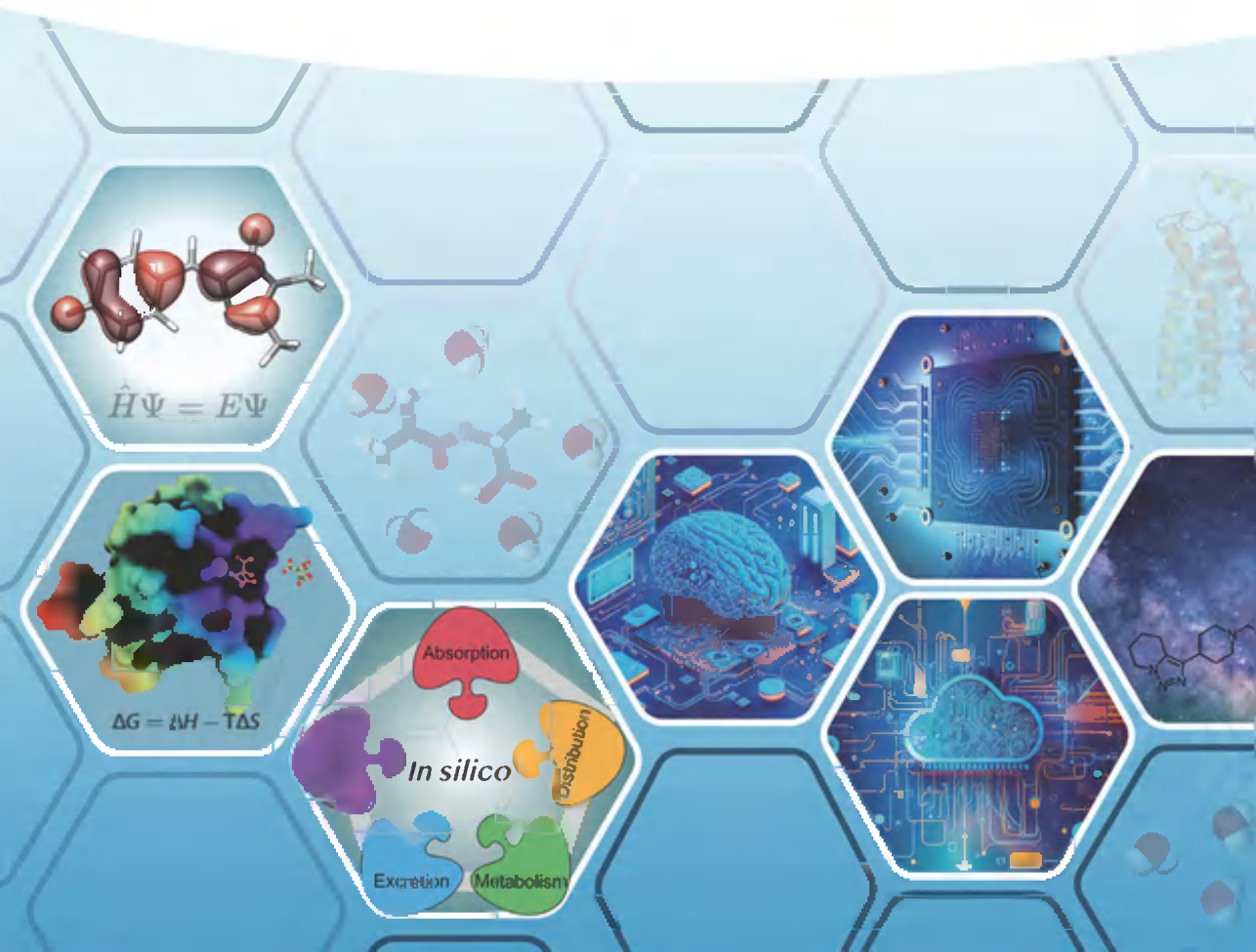


Edited by Vasanthanathan Poongavanam
and Vijayan Ramaswamy

Computational Drug Discovery

Methods and Applications

Volumes 1–2



Computational Drug Discovery

Computational Drug Discovery

Methods and Applications

*Edited by Vasanthanathan Poongavanam and
Vijayan Ramaswamy*

Volume 1

Editors

Dr. Vasanthanathan Poongavanam

Uppsala University
Department of Chemistry-BMC
751 05 Uppsala
Sweden

Dr. Vijayan Ramaswamy

University of Texas MD Anderson
Cancer Center
Institute for Applied Cancer Science
TX
United States

Cover: © Vasanthanathan Poongavanam

■ All books published by **WILEY-VCH** are carefully produced. Nevertheless, authors, editors, and publisher do not warrant the information contained in these books, including this book, to be free of errors. Readers are advised to keep in mind that statements, data, illustrations, procedural details or other items may inadvertently be inaccurate.

Library of Congress Card No.: applied for

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library.

Bibliographic information published by the Deutsche Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at <<http://dnb.d-nb.de>>.

© 2024 WILEY-VCH GmbH, Boschstraße 12, 69469 Weinheim, Germany

All rights reserved (including those of translation into other languages). No part of this book may be reproduced in any form – by photoprinting, microfilm, or any other means – nor transmitted or translated into a machine language without written permission from the publishers. Registered names, trademarks, etc. used in this book, even when not specifically marked as such, are not to be considered unprotected by law.

Print ISBN: 978-3-527-35374-3

ePDF ISBN: 978-3-527-84072-4

ePub ISBN: 978-3-527-84073-1

oBook ISBN: 978-3-527-84074-8

Typesetting Straive, Chennai, India

Contents

Volume 1

Preface xv

Acknowledgments xix

About the Editors xxi

Part I Molecular Dynamics and Related Methods in Drug Discovery 1

1 Binding Free Energy Calculations in Drug Discovery 3

Anita de Ruiter and Chris Oostenbrink

1.1 Introduction 3

1.1.1 Free Energy and Thermodynamic Cycles 4

1.2 Endpoint Methods 5

1.2.1 MM/PBSA and MM/GBSA 5

1.2.2 Linear Response Approximations 7

1.3 Alchemical Methods 9

1.3.1 Free Energy Perturbation 9

1.3.2 Thermodynamic Integration 10

1.3.3 Bennett's Acceptance Ratio 10

1.3.4 Nonequilibrium Methods 11

1.3.5 Multiple Compounds 11

1.3.6 One-Step Perturbation Approaches 12

1.3.7 Challenges in Alchemical Free Energy Calculations 13

1.4 Pathway Methods 15

1.5 Final Thoughts 17

References 17

2 Gaussian Accelerated Molecular Dynamics in Drug Discovery 21

Hung N. Do, Jinan Wang, Keya Joshi, Kushal Koirala, and Yinglong Miao

2.1 Introduction 21

2.2 Methods 22

2.2.1	Gaussian Accelerated Molecular Dynamics	22
2.2.2	Ligand Gaussian Accelerated Molecular Dynamics	24
2.2.3	Energetic Reweighting of GaMD for Free Energy Calculations	25
2.2.4	GLOW: A Workflow Integrating Gaussian Accelerated Molecular Dynamics and Deep Learning for Free Energy Profiling	26
2.2.5	Binding Kinetics Obtained from Reweighting of GaMD Simulations	26
2.2.6	Gaussian Accelerated Molecular Dynamics Implementations and Software	27
2.3	Applications	28
2.3.1	G-Protein-Coupled Receptors	28
2.3.1.1	Characterizing the Binding and Unbinding of Caffeine in Human Adenosine A _{2A} Receptor	28
2.3.1.2	Unraveling the Allosteric Modulation of Human A ₁ Adenosine Receptor	29
2.3.1.3	Ensemble Based Virtual Screening of Allosteric Modulators of Human A ₁ Adenosine Receptor	32
2.3.2	Nucleic Acids	33
2.3.2.1	Exploring the Binding of Risdiplam Splicing Drug Analog to Single-Stranded RNA	33
2.3.2.2	Uncovering the Binding of RNA to a Musashi RNA-Binding Protein	33
2.3.3	Human Angiotensin-Converting Enzyme 2 Receptor	35
2.3.4	Discovery of Novel Small-Molecule Calcium Sensitizers for Cardiac Troponin C	37
2.3.5	Binding Kinetics Prediction from GaMD Simulations	37
2.4	Conclusions	39
	References	40

3 MD Simulations for Drug-Target (Un)binding Kinetics 45

Steffen Wolf

3.1	Introduction	45
3.1.1	Preface	45
3.1.2	Motivation for Predicting (Un)binding Kinetics	45
3.1.3	The Time Scale Problem of MD Simulations	46
3.2	Theory of Molecular Kinetics Calculation	47
3.2.1	Nonequilibrium Statistical Mechanics in a Nutshell	47
3.2.2	Kramers Rate Theory	48
3.2.3	Biased MD Methods	49
3.2.3.1	Temperature- and Barrier-Scaling	49
3.2.3.2	Bias Potential-Based Methods	49
3.2.3.3	Bias Force-Based Methods	50
3.2.3.4	Knowledge-Biased Methods	50
3.2.3.5	Coarse-graining and Master Equation Approaches	51
3.3	Challenges and Caveats in Rate Prediction	51
3.3.1	Finding Reaction Coordinates and Pathways	51
3.3.2	Error Ranges of Estimates	52

3.3.3	A Need for Reliable Benchmarking Systems	53
3.3.4	Problems with Force Fields	53
3.4	Methods for Rate Prediction	53
3.4.1	Unbinding Rate Prediction	53
3.4.1.1	Empirical Predictions	53
3.4.1.2	Prediction of Absolute Unbinding Rates	54
3.4.2	Binding Rate Prediction	56
3.5	State-of-the-Art in Understanding Kinetics	57
3.6	Conclusion	57
	References	58
4	Solvation Thermodynamics and its Applications in Drug Discovery	65
	<i>Kuzhanthaivelan Saravanan and Ramesh K. Sista</i>	
4.1	Introduction	65
4.1.1	Protein Folding	65
4.1.2	Protein–Ligand Interactions	66
4.2	Tools to Assess the Solvation Thermodynamics	70
4.2.1	Watermap	71
4.2.2	GIST	72
4.2.3	3D-RISM	74
4.3	Case Studies	75
4.3.1	Watermap	75
4.3.1.1	Background and Approach	75
4.3.1.2	Results and Discussion	75
4.3.2	Grid Inhomogeneous Solvation Theory (GIST)	76
4.3.2.1	Objective and Approach	76
4.3.2.2	Results and Discussion	77
4.3.3	Three-Dimensional Reference Interaction-Site Model (3D-RISM)	78
4.3.3.1	Objective and Background	78
4.3.3.2	Results and Discussion	79
4.4	Conclusion	80
	References	80
5	Site-Identification by Ligand Competitive Saturation as a Paradigm of Co-solvent MD Methods	83
	<i>Asuka A. Orr and Alexander D. MacKerell Jr</i>	
5.1	Introduction	83
5.2	SILCS: Site Identification by Ligand Competitive Saturation	90
5.3	SILCS Case Studies: Bovine Serum Albumin and Pembrolizumab	97
5.3.1	SILCS Simulations	98
5.3.2	FragMap Construction	99
5.3.3	SILCS-MC	100
5.3.4	SILCS-Hotspots	102
5.3.5	SILCS-PPI	103

5.3.6	SILCS-Biologics	105
5.4	Conclusion	106
	Conflict of Interest	106
	Acknowledgments	107
	References	107

Part II Quantum Mechanics Application for Drug Discovery 119

6	QM/MM for Structure-Based Drug Design: Techniques and Applications	121
	<i>Marc W. van der Kamp and Jaida Begum</i>	
6.1	Introduction	121
6.2	QM/MM Approaches	122
6.2.1	Combined Quantum Mechanical/Molecular Mechanical Energy Calculations	122
6.2.2	QM/MM Methods for the Evaluation of Non-Covalent Inhibitor Binding	124
6.2.3	QM/MM Reaction Modeling	125
6.3	Applications of QM/MM for Covalent Drug Design and Evaluation	128
6.3.1	Covalent Tyrosine Kinase Inhibitors for Cancer Treatment	128
6.3.2	Evaluation of Antibiotic Resistance Conferred by β -Lactamases	133
6.3.3	Covalent SARS-CoV-2 Inhibitors: Mechanism and Insights for Design	138
6.4	Conclusions and Outlook	143
	References	144
7	Recent Advances in Practical Quantum Mechanics and Mixed-QM/MM-Driven X-Ray Crystallography and Cryogenic Electron Microscopy (Cryo-EM) and Their Impact on Structure-Based Drug Discovery	157
	<i>Oleg Borbulevych and Lance M. Westerhoff</i>	
7.1	Introduction	157
7.2	Feasibility of Routine and Fast QM-Driven X-Ray Refinement	159
7.3	Metrics to Measure Improvement	160
7.3.1	Ligand Strain Energy	160
7.3.2	ZDD of Difference Density	161
7.3.3	Overall Crystallographic Structure Quality Metrics: MolProbity Score and Clashscore	162
7.4	QM Region Refinement	162
7.5	ONIOM Refinement	165
7.6	XModeScore: Distinguish Protomers, Tautomers, Flip States, and Docked Ligand Poses	168

- 7.7 Impact of the QM-Driven Refinement on Protein–Ligand Affinity Prediction 169
- 7.7.1 Impact of Structure Inspection and Modification 172
- 7.7.2 Impact of Selecting Protomer States: Implications of XModeScore on SBDD 174
- 7.8 Conclusion 175
- Acknowledgments 177
- References 177

8 Quantum-Chemical Analyses of Interactions for Biochemical Applications 183

Dmitri G. Fedorov

- 8.1 Introduction 183
- 8.2 Introduction to FMO 184
- 8.3 Pair Energy Decomposition Analysis (PIEDA) 186
- 8.3.1 Formulation of PIEDA 186
- 8.3.2 Applications of PIEs and PIEDA 189
- 8.3.3 Example of PIEDA 189
- 8.4 Partition Analysis (PA) 190
- 8.4.1 Formulation of PA 193
- 8.4.2 Applications and an Example of PA 194
- 8.5 Partition Analysis of Vibrational Energy (PAVE) 195
- 8.5.1 Formulation of PAVE 196
- 8.5.2 Applications of PAVE 196
- 8.6 Subsystem Analysis (SA) 197
- 8.6.1 Formulation of SA 197
- 8.6.2 Examples of SA and PAVE 200
- 8.7 Fluctuation Analysis (FA) 201
- 8.8 Free Energy Decomposition Analysis (FEDA) 202
- 8.9 Other Analyses of Chemical Reactions 202
- 8.10 Conclusions 203
- References 203

Part III Artificial Intelligence in Pre-clinical Drug Discovery 211

9 The Role of Computer-Aided Drug Design in Drug Discovery 213

Storm van der Voort, Andreas Bender, and Bart A. Westerman

- 9.1 Introduction to Drug–Target Interactions, Hit Identification 213
- 9.2 Lead Identification and Optimization: QSAR and Docking-Based Approaches 215
- 9.3 DTI Machine Learning Methods 215
- 9.4 Supervised, Non-supervised and Semi-supervised Learning Methods 216

9.5	Graph-Based Methods to Label Data for DTI Prediction	217
9.6	The Importance of Explainable ML Methods: Linking Molecular Properties to Effects	218
9.7	Predicting Therapeutic Responses	219
9.8	ADMET-tox Prediction	220
9.9	Challenging Aspects of Using Computational Methods in Drug Discovery	220
9.9.1	What are Those Limitations?	221
	References	223

10 AI-Based Protein Structure Predictions and Their Implications in Drug Discovery 227

Tahsin F. Kellici, Dimitar Hristozov, and Inaki Morao

10.1	Introduction	227
10.2	Impact of AI-Based Protein Models in Structural Biology	229
10.2.1	Combination of AI-Based Predictions with Cryo-EM and X-Ray Crystallography	229
10.2.2	Combination of AI-Based Predictions with NMR Structures	232
10.2.3	Combination of AI-Based Predictions with Other Experimental Restraints	234
10.2.4	Impact of Deep Learning Models in Other Areas of Structural Biology	235
10.3	Combination of AI-Based Methods with Computational Approaches	236
10.3.1	Combination of Structure Prediction with Other Computational Approaches	242
10.4	Current Challenges and Opportunities	243
10.5	Conclusions	246
	References	246

11 Deep Learning for the Structure-Based Binding Free Energy Prediction of Small Molecule Ligands 255

Venkatesh Mysore, Nilkanth Patel, and Adegoke Ojewole

11.1	Introduction	255
11.2	Deep Learning Models for Reasoning About Protein–Ligand Complexes	257
11.2.1	Datasets	258
11.2.2	Convolutional Neural Networks	258
11.2.2.1	Background	258
11.2.2.2	Voxelized Grid Representation	258
11.2.2.3	Descriptors	259
11.2.2.4	Applications	259
11.2.3	Graph Neural Networks	260
11.2.3.1	Background	260
11.2.3.2	Graph Representation	260

11.2.3.3	Descriptors	260
11.2.3.4	Applications	260
11.2.3.5	Extension to Attention Based Models	261
11.2.3.6	Geometric Deep Learning and Other Approaches	261
11.3	Deep Learning Approaches Around Molecular Dynamics Simulations	261
11.3.1	Enhanced Sampling	262
11.3.2	Physics-inspired Neural Networks	262
11.3.3	Modeling Dynamics	263
11.3.3.1	Applications	263
11.4	Modifying AlphaFold2 for Binding Affinity Prediction	264
11.4.1	Modifying AlphaFold2 Input Protein Database for Accurate Free Energy Predictions	265
11.4.2	Modifying Multiple Sequence Alignment for AlphaFold2-Based Docking	265
11.5	Conclusion	266
11.5.1	New Models for Binding Affinity Prediction	266
11.5.2	Retrospective from the Compute Industry	266
11.5.2.1	Future DL-Based Binding Affinity Computation will Require Massive Scalability	267
11.5.2.2	Single GPU Optimizations for DL	267
11.5.2.3	Distributed DL Training and Inference	267
	References	268
12	Using Artificial Intelligence for <i>de novo</i> Drug Design and Retrosynthesis	275
	<i>Rohit Arora, Nicolas Brosse, Clarisse Descamps, Nicolas Devaux, Nicolas Do Huu, Philippe Gendreau, Yann Gaston-Mathé, Maud Parrot, Quentin Perron, and Hamza Tajmouati</i>	
12.1	Introduction	275
12.1.1	Traditional Drug Design and Discovery Process Is Slow and Expensive	275
12.1.2	Success and Limitations of Standard Computational Methods	276
12.1.3	AI-Based Methods can Accelerate Medicinal Chemistry	277
12.2	Quantitative Structure-Activity Relationship Models	278
12.2.1	Introduction to QSAR Models	278
12.2.2	QSAR Machine Learning Methods	278
12.2.3	QSAR Deep Neural Networks Methods	280
12.3	Modes of Generative AI in Chemistry	281
12.3.1	General Introduction	281
12.3.2	Generative AI in Lead Optimization	281
12.3.3	Fragment Growing	283
12.3.4	Novelty Generation	283
12.3.4.1	The Model	283
12.3.4.2	Optimization of the Novelty Generator	285

12.4	Importance of Synthetic Accessibility	285
12.4.1	Overview	285
12.4.2	Synthetic Scores	286
12.4.3	Integration of Synthetic Scores in Generative AI	288
12.4.3.1	An Example of a Lead Optimization Use Case	288
12.5	The Road Ahead	290
	References	290
13	Reliability and Applicability Assessment for Machine Learning Models	299
	<i>Fabio Urbina and Sean Ekins</i>	
13.1	Introduction	299
13.2	Challenges for Modeling	300
13.3	Example 1: BBB Applicability Domain Comparison	302
13.4	Example 2: Models for Uncertainty Estimation for Multitask Toxicity Predictions	303
13.5	Example 3: Class-Conditional Conformal Predictors	307
13.6	Conclusions	308
	Funding	309
	Competing Interests	309
	References	309

Volume 2

Preface	<i>xv</i>
Acknowledgments	<i>xix</i>
About the Editors	<i>xxi</i>

Part IV Chemical Space and Knowledge-Based Drug Discovery 315

14	Enumerable Libraries and Accessible Chemical Space in Drug Discovery	317
	<i>Tim Knehans, Nicholas A. Boyles, and Pieter H. Bos</i>	
15	Navigating Chemical Space	337
	<i>Ákos Tarscsay, András Volford, Jonathan Buttrick, Jan-Constantin Christopherson, Máte Erdős, and Zoltán B. Szabó</i>	
16	Visualization, Exploration, and Screening of Chemical Space in Drug Discovery	365
	<i>José J. Naveja, Fernanda I. Saldívar-González, Diana L. Prado-Romero, Angel J. Ruiz-Moreno, Marco Velasco-Velázquez, Ramón Alain Miranda-Quintana, and José L. Medina-Franco</i>	

- 17 SAR Knowledge Bases for Driving Drug Discovery 395**
Nishanth Kandepedu, Anil Kumar Manchala, and Norman Azoulay
- 18 Cambridge Structural Database (CSD) – Drug Discovery Through Data Mining & Knowledge-Based Tools 419**
Francesca Stanzione, Rupesh Chikhale, and Laura Friggeri
- Part V Structure-Based Virtual Screening Using Docking 441**
- 19 Structure-Based Ultra-Large Virtual Screenings 443**
Christoph Gorgulla
- 20 Community Benchmarking Exercises for Docking and Scoring 471**
Bharti Devi, Anurag TK Baidya, and Rajnish Kumar
- Part VI In Silico ADMET Modeling 495**
- 21 Advances in the Application of In Silico ADMET Models – An Industry Perspective 497**
Wenyi Wang, Fjodor Melnikov, Joe Napoli, and Prashant Desai
- Part VII Computational Approaches for New Therapeutic Modalities 537**
- 22 Modeling the Structures of Ternary Complexes Mediated by Molecular Glues 539**
Michael L. Drummond
- 23 Free Energy Calculations in Covalent Drug Design 561**
Levente M. Mihalovits, György G. Ferenczy, and György M. Keserű
- Part VIII Computing Technologies Driving Drug Discovery 579**
- 24 Orion® A Cloud-Native Molecular Design Platform 581**
Jesper Sørensen, Caitlin C. Bannan, Gaetano Calabrò, Varsha Jain, Grigory Ovanesyan, Addison Smith, She Zhang, Christopher I. Bayly, Tom A. Darden, Matthew T. Geballe, David N. LeBard, Mark McGann, Joseph B. Moon, Hari S. Muddana, Andrew Shewmaker, Jharrod LaFon, Robert W. Tolbert, A. Geoffrey Skillman, and Anthony Nicholls

- 25 Cloud-Native Rendering Platform and GPUs Aid
Drug Discovery 617**
*Mark Ross, Michael Drummond, Lance Westerhoff, Xavier Barbeau,
Essam Metwally, Sasha Banks-Louie, Kevin Jorissen, Anup Ojah, and
Ruzhu Chen*
- 26 The Quantum Computing Paradigm 627**
Thomas Ehmer, Gopal Karemore, and Hans Melo
- Index 679**

Preface

Computer-aided drug design (CADD) techniques are used in almost every stage of the drug discovery continuum, given the need to shorten discovery timelines, reduce costs, and improve the odds of clinical success. CADD integrates modeling, simulation, informatics, and artificial intelligence (AI) to design molecules with desired properties. Briefly, the application of CADD methodologies in drug discovery dates back to the 1960s, tracing its origin to the development of quantitative structure–activity relationship (QSAR) approaches. Between the 1970s and 1980s, computer graphics programs to visualize macromolecules began to take off together with advancements in computational power. This coincided with the emergence of more sophisticated techniques, including mapping energetically favorable binding sites on proteins, molecular docking, pharmacophore modeling, and modeling the dynamics of biomolecules. Since then, CADD has evolved as a powerful technique opening new possibilities, leading to increased adoption within the pharmaceutical industry and contributing to the discovery of several approved drugs.

Recent developments in CADD have been propelled by advancements in computing, breakthroughs in related fields such as structural biology, and the emergence of new therapeutic modalities. Notably, the advent of highly parallelizable GPUs and cloud computing have significantly increased computing power, while quantum computing holds promise to simulate complex systems at an unprecedented scale and speed. Advances in AI technologies, particularly generative AI for molecule design, are reducing cycle times during lead optimization. Meanwhile, the resolution revolution in cryo-electron microscopy (cryo-EM) and AI-powered structure biology are shedding light on the three-dimensional structure of many therapeutically relevant drug targets, thereby expanding our ability to carry out structure-based drug design against these targets. Other exciting breakthroughs that offer new opportunities include the explosion in the size of "make-on-demand" chemical libraries that enable ultra-large-scale virtual screening for hit identification, the big data phenomena in medicinal chemistry with the advent of bioactivity databases like ChEMBL and GOSTAR that provide access to millions of SAR data points useful for building predictive models and for knowledge-based compound, the emergence of new therapeutic modalities like targeted protein degradation like PROTACs and molecular glues, and viable approaches for targeting various reactive amino acid side chains beyond cysteine for developing covalent inhibitors. These

developments are also now enabling drug discovery scientists to tackle high-value drug targets previously considered undruggable.

The changing paradigm in drug discovery, complemented by technological advancements, has significantly expanded the toolbox available for computational chemists to enable drug discovery in recent years. Against this backdrop, we felt a need for a book that offers up-to-date information on the most important developments in the field of CADD. This book, titled “Computational Drug Discovery,” is meant to be a valuable resource for readers seeking a comprehensive account of the latest developments in CADD methods and technologies that are transforming small-molecule drug discovery. The intended target audience for this book is medicinal chemists, computational chemists, and drug discovery professionals from industry and academia.

The book is organized into eight thematic sections, each dedicated to a cutting-edge computational method, or a technology utilized in computational drug discovery. In total, it comprises 26 chapters authored by renowned experts from academia, pharma, and major drug discovery software providers, offering a broad overview of the latest advances in computational drug discovery.

Part I explores the role of molecular dynamics simulation and related approaches in drug discovery. It encompasses various topics such as the utilization of physics-based methods for binding free energy estimation, the theory and application of enhanced sampling methods like Gaussian Accelerated MD to facilitate efficient sampling of the conformational space, understanding binding and unbinding kinetics of compound binding through molecular dynamics simulation, the application of computational approaches like WaterMap and 3D-RISM framework to understand the location and thermodynamic properties of solvents that solvate the binding pocket which offers rich physical insights compound design, and the use of mixed solvent MD simulations for mapping binding hotspots on protein surfaces based on the SILCS technology.

Part II focuses on the role of quantum mechanical approaches in drug discovery, covering topics such as the use of hybrid QM/MM method for modeling reaction mechanisms and covalent inhibitor design, refinement of X-ray and cryo-EM structures integrating QM and QM/MM approaches for accurate assignment of tautomer, protomers, and amide flip rotamers for downstream structure-based design, and quantifying protein–ligand interaction energies using QM methods at a reduced computational cost like the fragment molecular orbital (FMO) framework

Part III focuses on the application of AI in preclinical drug discovery, highlighting its growing importance across different stages of the drug discovery process. Given the recent advancements in AI and related technologies, we have chapters that outline advancements in deep learning for protein structure prediction, in particular the significant breakthrough achieved by AlphaFold2, the use of deep learning architectures such as Convolutional Neural Networks (CNNs), Graph Neural Networks (GNNs), and physics-inspired neural networks for predicting protein–ligand binding affinity, the emergence of generative modeling techniques for *de novo* design of synthetically tractable drug-like molecules that satisfy a defined set of constraints. In order to offer readers guidance on effectively applying

machine learning (ML) models and ensuring their validity and usefulness, this section includes a chapter that discusses different approaches for evaluating the reliability and domain applicability of ML models.

Part IV of this book focuses on how the concept of chemical space and the big data phenomenon are driving drug discovery. It includes chapters describing innovative approaches in reaction-based enumerations that enable the generation of virtual libraries containing tangible compounds, followed by computational solutions for visualizing and navigating this vast chemical space. Additionally, this section also highlights the use of SAR *knowledge bases like GOSATR* for extracting valuable insights and generating robust design ideas based on medicinal chemistry precedence. Wrapping up the section is a chapter highlighting how the wealth of knowledge gained by mining the data in CSD is proving valuable in various stages of drug discovery.

The ever-expanding size of compound libraries and the advent of make-on-demand compound libraries have elevated virtual screening to a whole new level.

Part V focuses on ultra-large-scale virtual screening using approaches that scale virtual screening methods to match the size of these massively large compound libraries. Although virtual screening using docking is a well-established approach for hit finding in drug discovery, the ability of docking programs to generate the correct binding mode and accurately estimate binding affinity is still a challenge. Hence, we have a chapter that reviews collaborative efforts within the scientific community for evaluating and comparing the performance of docking methods, establishing standardized metrics for assessing the efficiency of virtual screening techniques through rigorous competitive evaluations.

Early profiling of absorption, distribution, metabolism, excretion, and toxicity (ADMET) endpoints in early drug discovery is essential for designing and selecting compounds with superior ADMET properties. Consequently, major pharmaceutical companies have developed and implemented predictive models within their organizations for predicting multiple endpoints to enhance compound design. **Part VI** of the book chapter offers an overview of in silico ADMET methods and their practical applications in facilitating compound design within an industrial context. **Part VII** explores the role of computational techniques in accelerating the design of cutting-edge therapeutic modalities. This section provides a comprehensive focus on two key areas: the design of molecular glues and the design of covalent inhibitors.

In addition to the aforementioned methods and approaches that revolutionize the drug discovery process, computing technologies are further accelerating drug discovery with enhanced speed and accuracy.

Part VIII is dedicated to exploring how cloud computing and quantum computing significantly expand the range of drug discovery opportunities. Particularly, there is great hope and excitement surrounding the potential applications of quantum computing in drug discovery. *“The Quantum Computing Paradigm”* provides a comprehensive review on quantum computing from the perspective of drug discovery. In addition to discussing several drug discovery applications, including peptide design, the chapter also addresses challenges associated with this emerging drug discovery technology.

In conclusion, we believe that this book provides a thorough overview of the recent advancements in computational drug discovery, making it an engaging and captivating read. We would like to express our deepest appreciation to all the authors for their invaluable contributions to this book. Their expertise, insights, and unwavering commitment have greatly enriched its content and overall significance.

14 September 2023

Vasanthanathan Poongavanam, Uppsala, Sweden
Vijayan Ramaswamy, Texas, USA

Acknowledgments

First and foremost, we would like to extend our sincerest gratitude and profound appreciation to all the contributing authors. Their unwavering commitment, tireless efforts, and remarkable enthusiasm have been instrumental in bringing this book to fruition. It is their willingness to share their knowledge and experience that has greatly enriched its content, resulting in a truly valuable and comprehensive book that provides an account of the latest advancements in the field of computer-aided drug design.

We also extend our sincere gratitude to the external reviewers for their timely feedback and insightful suggestions that helped improve the quality of the book and shape the final outcome. Our special thanks to the following individuals for their invaluable contributions in reviewing the book chapters: Dr. Andreas Tosstorff (F. Hoffmann-La Roche, Switzerland), Dr. Sagar Gore, Dr. Suneel Kumar BVS (Molecular Forecaster, Canada), Dr. Pandian Sokkar, Dr. Ono Satoshi (Mitsubishi Tanabe Pharma, Japan), Dr. Octav Caldararu (Zealand Pharma, Denmark), Dr. Sundarapandian Thangapandian (HotSpot Therapeutics, Inc, USA), Dr. Vigneshwaran Namasivayam (Dewpoint Therapeutics, Germany), Dr. Yinglong Miao (University of Kansas, USA), Nanjie Deng (Pace University, USA), and Dr. Ansuman Biswas (Ernst & Young, India).

In conclusion, we would like to express our gratitude to the publisher Wiley for entrusting us with an opportunity to edit this book and for the fruitful collaboration. Especially, we convey our appreciation to Katherine Wong (Senior Managing Editor) and Dr. Lifan Yang (Program Manager) at Wiley for their unwavering support, encouragement throughout the editing process, and their commitment to ensuring the quality and excellence of this book. The editors also extend their thanks to Prof. Jan Kihlberg and Dr. Jason B. Cross for their continuous support, which made this project possible.

About the Editors

Vasanthanathan Poongavanam is a senior scientist in the Department of Chemistry-BMC, Uppsala University, Sweden. Before starting at Uppsala University in 2016, he was a postdoctoral fellow at the University of Vienna, Austria, and at the University of Southern Denmark. He obtained his PhD degree in Computational Medicinal Chemistry as a Drug Research Academy (DRA) Fellow at the University of Copenhagen, Denmark, on computational modeling of cytochrome P450. He has published more than 65 scientific articles, including reviews and book chapters. His scientific interests focus on in silico ADMET modeling, including cell permeability and solubility, and he has worked extensively on understanding the molecular properties that govern the pharmacokinetic profile of molecules in bRo5 property space, including macrocycles and PROTACs.

Vijayan Ramaswamy (R.S.K. Vijayan) is a senior research scientist affiliated with the Structural Chemistry division at the Institute for Applied Cancer Science, The University of Texas MD Anderson Cancer Center, TX, USA. In 2016, he joined MD Anderson Cancer after a brief tenure as a scientist in computational chemistry at PMC Advanced Technologies, New Jersey, USA. He undertook postdoctoral training at Rutgers University in New Jersey, USA, and Temple University in Pennsylvania, USA. He received his PhD as a CSIR senior research fellow from the Indian Institute of Chemical Biology, Kolkata, India. He is a named co-inventor on seven issued US patents, including an ATR kinase inhibitor that has advanced to clinical trials. He has published more than 20 scientific articles and authored one book chapter. His research focuses on applying computational chemistry methods to drive small-molecule drug discovery programs, particularly in oncology and neurodegenerative diseases.

Part I

Molecular Dynamics and Related Methods in Drug Discovery

1

Binding Free Energy Calculations in Drug Discovery

Anita de Ruiter¹ and Chris Oostenbrink^{1,2}

¹Institute for Molecular Modeling and Simulation, Department of Material Sciences and Process Engineering, University of Natural Resources and Life Sciences, Vienna, Muthgasse 18, 1190 Vienna, Austria

²Christian Doppler Laboratory for Molecular Informatics in the Biosciences, University of Natural Resources and Life Sciences, Vienna, Muthgasse 18, 1190 Vienna, Austria

1.1 Introduction

This chapter attempts to provide an overview of the different approaches and methods that are available to compute binding free-energy in drug design and drug discovery. We do not provide an exhaustive list of available methods and do not rigorously derive all of the methods from first principles. Instead, we aim to give a overview of available methods and to point at the intrinsic limitations and challenges of these methods, such that researchers applying these methods can make a fair estimate of the most appropriate methods for their aims.

Numerous methods for the calculation of binding free energies have been developed over the years [1]. Which method is the best choice depends on how many free energies need to be determined, the available computational resources, the accuracy one wishes to obtain, and other specific properties of the system under study. Let us start by separating the available methods into three classes. Binding free energies can be calculated with endpoint, alchemical, or pathway methods. These methods are very different, not only in terms of their underlying theory but also in their accuracy and efficiency. The endpoint methods are very efficient in terms of computational requirements, but, unfortunately, not very accurate. Alchemical methods, on the other hand, are considered one of the most accurate but also slow methods. Pathway methods are also computationally demanding but can give important information about the binding pathways. Which method is the best choice will mostly depend on the stage at which the drug discovery/design is currently at. In the very early stages, where whole databases of compounds need to be screened, one can likely not afford the computational costs of alchemical approaches. However, since the range of binding free energies that are to be predicted may also be rather large, the faster methods will be sufficient to pick up some hit compounds. In the lead optimization stage, where rather similar compounds are studied, a more accurate method is required that can detect smaller differences

in the binding free energies. Because the optimization stage also focuses on fewer leads, the higher computational demand for the more accurate method can actually be afforded.

1.1.1 Free Energy and Thermodynamic Cycles

First of all, we should look into the definition of free energy to find out what kind of property we are trying to determine. The definition of free energy in statistical mechanics is

$$G = -k_B T \ln Q_{NPT} \quad (1.1)$$

where G is the Gibbs free energy, k_B is the Boltzmann constant, T is the temperature, and Q_{NPT} is the partition function for a system with a constant number of particles, pressure, and temperature. The partition function is defined as

$$Q_{NPT} = \int \int e^{-H(\mathbf{r}, \mathbf{p})/k_B T} d\mathbf{r} d\mathbf{p} \quad (1.2)$$

with $\mathbf{r}$ and $\mathbf{p}$ as the positions and momenta of all atoms in the system, respectively, and $H(\mathbf{r}, \mathbf{p})$ as the Hamiltonian of the system, giving the total energy. It is clear from the integration of all positions and momenta in Eq. (1.2) that the free energy is intrinsically a property of a statistical mechanical ensemble and not something that can be estimated from a single configuration. Free energy is a property of all (relevant) configurations of the system together. Any effort to estimate the free energy from a single configuration will likely miss out on some relevant aspects of the ensemble, such as conformational changes and their entropic effects.

In the field of drug discovery, one is not interested in the absolute free energy of a certain state, but rather in the binding free energy of, e.g. a small molecule to a protein;

$$\Delta G_{bind} = -k_B T \ln \frac{Q_{bound}}{Q_{free}} \quad (1.3)$$

where Q_{bound} and Q_{free} are the partition functions for the system where the small molecule is bound to the protein and when both partners are free in solution, respectively.

Furthermore, during hit-to-lead or lead optimization stages, one is mostly interested in the relative binding free energy, i.e. which of the two ligands binds stronger to the protein than the other. This, together with the fact that free energy is a state function, makes it possible to design thermodynamic cycles to make it easier to calculate the free energies. Consider a thermodynamic cycle like that in Figure 1.1.

There are four states, one with ligand A bound to the protein, one with ligand A unbound from the protein, one with ligand B bound to the protein, and one with ligand B unbound from the protein. Since free energy is a state function, following the full thermodynamic cycle will lead to a free energy difference of 0:

$$\Delta G_{BA}(prot) - \Delta G_{bind}(B) - \Delta G_{BA}(free) + \Delta G_{bind}(A) = 0 \quad (1.4)$$

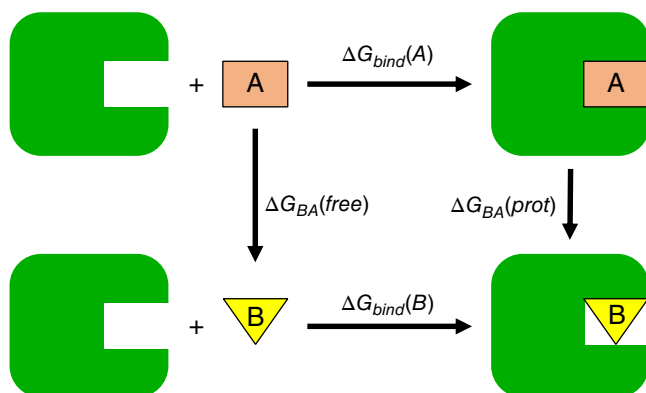


Figure 1.1 Thermodynamic cycle for the calculation of relative binding free energies of two small molecules A and B binding to a common receptor.

From this, it follows that $\Delta G_{bind}(B) - \Delta G_{bind}(A) = \Delta G_{BA}(prot) - \Delta G_{BA}(free)$. This means that we can determine the difference in binding free energy without performing a tedious simulation of the actual binding process. Although the modification of ligand A to ligand B is not something that is physically possible in the laboratory, it is possible with computer simulations and alchemical free-energy calculations. In fact, it is often easier to obtain converged results for these unphysical processes because modifying the ligands will most likely lead to much less reorganization of the protein than the binding process would. Modifying the ligand requires intermediate states, which will be discussed in more detail in the section on alchemical methods.

1.2 Endpoint Methods

As the name implies, endpoint methods only require the simulation of the endpoints of the system of interest. For binding free energy calculations, the endpoints would be the protein–ligand complex and the separate protein and ligand. That is, we explicitly simulate the states of the corners of the thermodynamic cycle of Figure 1.1. Their efficiency and reasonable accuracy make the endpoint free energy methods very popular in the early stages of drug discovery. Here, we will discuss two kinds of endstate methods: the molecular mechanics Poisson–Boltzmann surface area (MM/PBSA) methods and methods derived from linear response theory.

1.2.1 MM/PBSA and MM/GBSA

The most commonly used methods are MM/PBSA and the closely related molecular mechanics generalized Born surface area (MM/GBSA) [2, 3].

In MM/PBSA, the free energy of a state is composed of several contributions;

$$G = E_{MM} + G_{pol} + G_{np} - TS \quad (1.5)$$

$$E_{MM} = E_{bnd} + E_{el} + E_{vdW} \quad (1.6)$$

Here, E_{MM} is the molecular mechanics potential energy term, which consists of bonded interactions (E_{bnd}), electrostatic interactions (E_{el}), and van der Waals interactions (E_{vdW}). G_{pol} and G_{np} are the polar and nonpolar contributions to the solvation free energy, respectively. T represents the temperature of the system, and S is the entropy. Note that, although the free energy is a property of the ensemble and not an average over the ensemble, these methods assume that these terms together approximate the free energy of the state reasonably well and can be computed from individual configurations of the ensemble.

In order to calculate the absolute binding free energy of a system, the free energy of the free ligand (L), the unbound protein (P), as well as the complex (PL) needs to be computed;

$$\Delta G_{bind} = \langle G_{PL} \rangle_{PL} - \langle G_P \rangle_P - \langle G_L \rangle_L \quad (1.7)$$

Here, the angular brackets indicate an ensemble average from the simulation of the system indicated in the subscript. Equation (1.7) is the so-called three-average MM/PBSA (3A-MM/PBSA) since three different simulations need to be performed. The ensembles in Eq. (1.7) are generated from snapshots of molecular dynamics simulations with an explicit solvation model. Once these snapshots are generated, they are stripped from all solvent molecules and ions, and an implicit solvation model is used for further analysis.

G_{pol} is determined either by solving the Poisson–Boltzmann (PB) equation or by using the generalized Born equation (in which case the method would be called MM/GBSA). GB uses an analytical expression for the polar solvation energy and is thus much faster, but also likely to be less accurate, although this is system-dependent. G_{np} is estimated by using the solvent accessible surface area (SA). The assumption that G_{pol} and G_{np} can be approximated from an implicit solvation model means that solvent degrees of freedom are no longer treated explicitly in Eq. (1.2) and lead to the use of simple ensemble averages in Eq. (1.7). The calculation of G_{pol} furthermore depends strongly on the implicit solvation model that is used. Usually, the implicit solvation model requires a single dielectric constant to be chosen to describe the very complex electrostatic environment within the protein. This either makes the results unreliable, or the user can choose the constant such that the results are in agreement with known binding free energies for the system. In the latter case, MM/PBSA becomes more of an empirical method, where the parameters are optimized to reproduce experimental data. Finally, as a result of the implicit solvation model, MM/PBSA is not very well-suited when the binding site involves a highly charged environment or when critical water molecules are within the binding site.

The second reason that the ensemble property of Eqs. (1.1) and (1.2) may be approximated by a simple ensemble average in Eq. (1.7) is the explicit separation of the protein and ligand degrees of freedom into an energetic contribution (E_{MM}) and an entropic contribution (TS). The energetic term is computed from a force field, which is indeed well captured by an ensemble average. The entropy term is most

commonly estimated with normal mode analysis (NMA). However, this method, which estimates the curvature of the energy landscape and approximates the entropy based on the expected sampling on this surface, is rather time-consuming and therefore not suitable for larger systems. More efficient methods have been explored over the years, but especially when the interest lies with relative binding free energies (like in drug discovery), the entropy term is often simply ignored. The underlying assumption would be that similar ligands will have similar entropy terms. Effectively, however, this means that the free energy is approximated by an energy.

A further, very popular, approximation is to use the single-trajectory MM/PBSA (1A-MM/PBSA), where only the complex is simulated

$$\Delta G_{bind} = \langle G_{PL} - G_P - G_L \rangle_{PL} \quad (1.8)$$

G_P and G_L are determined from the ensemble of the complex by just removing the atoms that are not part of the state of interest, i.e. for G_P , the ligand atoms are removed from the complex simulations, and for G_L , the protein atoms are removed. There are two main advantages of 1A-MM/PBSA with respect to 3A-MM/PBSA. The most obvious one is that only a single simulation needs to be performed instead of three simulations, and therefore it is computationally more efficient. The second one comes from the fact that E_{bind} and all intramolecular contributions to E_{el} and E_{vdW} cancel in Eq. (1.8) because these energies for the apo protein and isolated ligand are calculated from exactly the same configuration as the complex. Also, the entropy estimate will seemingly become negligible since the ligand does not sample different conformations in the bound or in the free state. This significantly reduces the noise in the free energies, allowing for faster convergence of the results.

However, we need to consider what additional assumptions are being made with the 1A-MM/PBSA approach. It basically assumes that the protein and ligand visit the exact same conformations when they are in complex with each other as when they are each free in solution. This is not very likely to be the case. For example, the ligand can be forced to be more rigid (and/or bend) within a tight binding site, and a protein side chain or loop can be pushed aside upon binding of the ligand. The energetic and entropic effects of such conformational changes can be significant, but are entirely absent from the 1A-MM/PBSA approach. Unfortunately, the fact that the 1A-MM/PBSA often leads to less noise in the calculation does not make it more appropriate.

Recent advances try to address several of the approximations in the MM/PBSA and MM/GBSA methods with some promising results, but the optimal solutions remain rather system-dependent. For further reading, we suggest some recent reviews on the topic [3–5].

1.2.2 Linear Response Approximations

Other endpoint methods are based on the linear response theory. In the linear response approximation (LRA) framework [6, 7], two additional states need to be simulated, which are the neutralized states of the ligand when it is bound to the

protein and when it is free in solution. Any partial charges of the ligand are set to 0 in these simulations. The charging free energy difference is then calculated with

$$\begin{aligned}\Delta G_{N \rightarrow Q}^{LRA} &= G_Q - G_N = \frac{1}{2}[\langle H_Q - H_N \rangle_N + \langle H_Q - H_N \rangle_Q] \\ &= \frac{1}{2} \left(\langle V_{ls}^{el} \rangle_N + \langle V_{ls}^{el} \rangle_Q \right)\end{aligned}\quad (1.9)$$

where H_Q is the Hamiltonian of the charged state and H_N is the Hamiltonian of the neutralized state; subscripts after the angular brackets indicate which Hamiltonian was used to obtain the ensemble; and V_{ls}^{el} are the electrostatic interactions of the ligand with its surroundings.

In the Linear Interaction Energy (LIE) method [8], it is assumed that the electrostatic interactions of the charged ligand obtained from the ensemble of neutral states, $\langle V_{ls}^{el} \rangle_N$, will average to 0. Although this assumption is reasonable for the ligand in solution, it might not hold for the ligand bound to a protein. The protein is not likely to have a random electrostatic distribution around the neutral ligand and $\langle V_{ls}^{el} \rangle_N$ corresponds to the preorganization energy of the protein. The free energy difference of charging a ligand using LIE is then calculated with

$$\Delta G_{N \rightarrow Q}^{LIE} \approx \beta \langle V_{ls}^{el} \rangle_Q \quad (1.10)$$

where β is theoretically $1/2$. The nonpolar interactions are also assumed to have a linear relationship with the free energy difference, even though this is only based on observations that the free energy of solvation for nonpolar particles and the interaction energy both seem to be linearly correlated with the size of the molecule. The binding free-energy difference based on LIE can thus be calculated with

$$\Delta G_{bind}^{LIE} = \alpha \Delta \langle V_{ls}^{vdW} \rangle_Q + \beta \Delta \langle V_{ls}^{el} \rangle_Q + \gamma \quad (1.11)$$

where α and γ are empirical parameters, which can be used to fit to experimental data from a data set. Δ indicates the difference between the ensemble averages obtained from the simulation of the free ligand and when bound to the protein. Even though β has a theoretical value of $1/2$, it is also often used as an empirical parameter. Scaling the interactions with α , β , and adding γ helps compensate for the missing factors in the LIE approach, such as intramolecular energies, entropic confinement, and desolvation effects.

As an alternative to LIE, we have developed third power fitting (TPF), in which we do not assume linearity for the charging free energy [9]. Instead, the neutral and charged states are simulated, and the curvature of the charging curve is estimated by a third-order polynomial of a coupling parameter λ . Four constraints are used to find the best fit, which are based on the first ($dG/d\lambda$) and second ($d^2G/d\lambda^2$) derivatives of the free energy with respect to λ , from simulations in the N and Q states of LRA. It can be shown using the cumulant expansion that $d^2G/d\lambda^2$ is equal to the negative of the fluctuations of $dH/d\lambda$

$$\left. \frac{d^2G}{d\lambda^2} \right|_s = \frac{1}{k_B T} \left(\left\langle \left(\frac{\partial H}{\partial \lambda} \right)_s^2 \right\rangle - \left\langle \left(\frac{\partial H}{\partial \lambda} \right)_s \right\rangle^2 \right) = \frac{1}{k_B T} \left(\langle V_{ls}^{el} \rangle_s^2 - \langle (V_{ls}^{el})^2 \rangle_s \right) \quad (1.12)$$

With the subscript S corresponding to either N or Q . The advantage of TPF is that there is that the nonlinearity is captured without additional simulations or empirical parameters. One should keep in mind, though, that the fluctuations in Eq. (1.12) are slower to converge.

1.3 Alchemical Methods

Once one or more lead compounds have been discovered during the early stages of drug discovery, lead optimization can be performed with more rigorous alchemical methods. Especially relative binding free energies between compounds that do not differ too much can be calculated very accurately with these methods. As mentioned above, relative binding free energies can be determined by morphing ligand A into ligand B, when bound and when free in solution. This means that simulations are performed in the direction represented by the vertical arrows of the thermodynamic cycle in Figure 1.1. Molecular properties that are changed during this process can include atom type, (partial) charges, bond lengths, angles, and dihedrals. All these changes are usually performed with several intermediate steps since convergence of the simulations would otherwise not be reached. Since free energy is a state function, any intermediate state can be chosen to make the calculations more efficient, even if it is unphysical. The intermediate states are defined by a coupling parameter λ , where at $\lambda = 0$, ligand A is represented, and at $\lambda = 1$, ligand B is represented. This means that at intermediate values of λ , the ligand is a nonphysical representation of a mixture of both ligands.

1.3.1 Free Energy Perturbation

Thermodynamic integration (TI) [10] and Bennett acceptance ratio (BAR) [11] are the two most commonly used alchemical free energy methods. All of these methods are often referred to as free energy perturbation, even though we prefer to reserve that term for the perturbation equation of Zwanzig [12]. This equation shows that relative free energy can be expressed in terms of the ratio of the partition functions of the two states.

$$\Delta G_{BA}^{FEP} = G_B - G_A = -k_B T \ln \frac{\iint e^{-H_B/k_B T}}{\iint e^{-H_A/k_B T}} \quad (1.13)$$

Multiplying the first exponential by 1 written in the form of $e^{+H_A/k_B T} e^{-H_A/k_B T}$, we find an ensemble average

$$\Delta G_{BA}^{FEP} = -k_B T \ln \left\langle e^{-\frac{H_B - H_A}{k_B T}} \right\rangle_A \quad (1.14)$$

A simulation of only state A is thus predicting the free energy difference toward state B. Accurate results are only obtained if the simulation of state A also samples the relevant conformational states for state B. If this is not the case, additional intermediate states can be used to increase the phase space overlap.

1.3.2 Thermodynamic Integration

In TI, the free energy difference between two states A and B is calculated with

$$\Delta G_{BA}^{TI} = \int_0^1 \frac{dG(\lambda)}{d\lambda} = \int_0^1 \left\langle \frac{\partial H(\lambda)}{\partial \lambda} \right\rangle_{\lambda} d\lambda \quad (1.15)$$

where the term with the angular brackets indicates the ensemble average of the derivative of the Hamiltonian with respect to λ obtained from the simulation at that λ -value. After all simulations at a number (N) of intermediate states are finished, the free energy difference between the two end states is obtained by numerical integration. In most cases, the trapezoidal rule is used for integration, although other integration schemes have been suggested as well. The number and spacing of intermediate states that are required to get a reliable result are strongly system-dependent. Since the integration is done numerically, areas with large curvature need to have a denser spacing of λ . We ourselves have suggested the use of extended TI, in which the values of $\langle \partial H / \partial \lambda \rangle$ at many intermediate λ -values are predicted from simulations at a smaller number of λ -values, leading to smooth curves that explicitly capture the curvature [13].

1.3.3 Bennett's Acceptance Ratio

Another method to calculate the free energy difference between two states is BAR. This method uses ensemble averages obtained from both states to get a minimum statistical error of the free energy difference.

$$\Delta G_{ji}^{BAR} = k_B T \ln \left(\frac{\langle f(H_i - H_j + C) \rangle_j}{\langle f(H_j - H_i - C) \rangle_i} \right) + C \quad (1.16)$$

where $f(x) = 1/[1 + \exp(x/k_B T)]$ is the Fermi function and C is a constant. The optimal statistical estimate of the free energy difference is obtained when one solves for C such that the numerator and denominator of the logarithm are equal. It follows that

$$C = -k_B T \ln \left(\frac{N_j Q_i}{N_i Q_j} \right) \quad (1.17)$$

where N_j and N_i are the number of configurations in states j and i , respectively. Q_i and Q_j are the partition functions for states i and j , respectively. Since the partition functions are included in C , they need to be solved iteratively to get a self-consistent result. Convergence of the iterative process will only be reached if sufficient overlap between the forward and backward energy differences ($H_i - H_j$ and $H_j - H_i$ as obtained in the ensemble averages of j and i , respectively) is achieved. The final free energy difference between states A and B is determined with

$$\Delta G_{BA}^{BAR} = \sum_{i=1}^{n-1} \Delta G_{i+1,i}^{BAR} \quad (1.18)$$

BAR can be extended to using the configurations of all states to contribute to the free energy determination; this is referred to as Multistate Bennett Acceptance Ratio (MBAR) [14].

1.3.4 Nonequilibrium Methods

TI and BAR are both methods that rely on the simulation of states, which are at equilibrium. However, nonequilibrium simulations can also be used to determine free energy differences. According to the Jarzynski equation [15], the ensemble average over the exponential distribution of the work relates to the free energy;

$$\Delta G = -k_B T \ln \langle e^{-W/k_B T} \rangle \quad (1.19)$$

This means that alchemical perturbations can be performed in a single simulation, where λ is changing from 0 to 1, continuously. Since the system does not need to be in equilibrium, these simulations can be rather short. However, they must be performed many times, and initial configurations must be sampled from a proper ensemble, in order to get a reasonable estimate of the work distribution.

Another nonequilibrium method is the Crooks Gaussian intersection (CGI) [16]. Similar to the Jarzynski equation, it requires the work obtained from a number of nonequilibrium simulations. In addition, simulations are performed in the inverse direction (i.e. changing λ from 1 to 0). The forward and backward work distributions are fitted with Gaussian distributions, and the intersection of them leads to the free energy estimate;

$$\frac{P_f(W)}{P_b(W)} = e^{W - \Delta G/k_B T} \quad (1.20)$$

Here, $P_f(W)$ and $P_b(W)$ are the work distributions for the forward and backward simulations, respectively. CGI will give results with reasonable accuracy when the distributions are sufficiently sampled and there is an overlap between the forward and backward work distributions.

1.3.5 Multiple Compounds

The alchemical methods described above are used to determine a single relative binding free energy between two compounds. However, during the lead optimization stage, there are many more compounds of interest. Just performing all possible combinations of compounds will be extremely time-consuming and will likely lead to many non-converged simulations because the differences between the compounds are too large. So, how do we decide which combinations of the compounds are good to determine the relative binding free energies? For smaller sets of compounds, this can be done manually, but for larger sets, the lead optimization mapper (LOMAP) [17] has been introduced. LOMAP uses the maximal common substructure to group structurally similar compounds together. It makes sure that rings are preserved as much as possible and that the compounds have the same net charge when relative binding free energies are being calculated (see below for a short description of why such perturbations are particularly challenging). It then determines optimal thermodynamic cycles, like in Figure 1.2, which can be used as an internal validation. When the free energy along the thermodynamic cycle is close to 0, the free energy calculations are likely to have converged. LOMAP designs these cycles automatically and keeps them rather small in order to prevent cancellation of

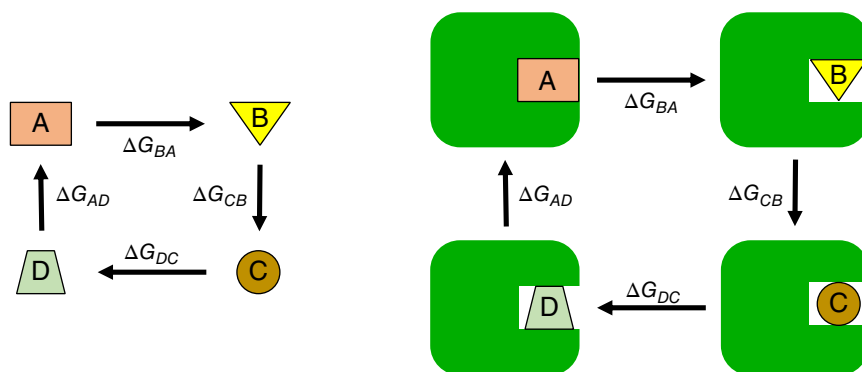


Figure 1.2 Thermodynamic cycles designed for internal validation of the calculated relative binding free energies.

errors. It also makes sure that all compounds (of the same net charge) are connected to each other.

1.3.6 One-Step Perturbation Approaches

Another approach to include multiple compounds is to generate simulation setups from which the binding free energies of more than two compounds can be estimated in one go. We have already discussed that intermediate states do not necessarily have to be physical. This means we can also use methods that rely on the simulation of only a single nonphysical (intermediate) state. An alternative thermodynamic cycle can be created, where one goes from state A to a reference state REF and then from REF to state B.

$$\Delta G_{A \rightarrow B} = \Delta G_{A \rightarrow \text{REF}} + \Delta G_{\text{REF} \rightarrow B} \quad (1.21)$$

The free energy difference can still be estimated with the Zwanzig relationship, based on the simulation of the REF state.

$$\Delta G_{A \rightarrow \text{REF}} = -k_B T \ln \left\langle e^{-(H_A - H_{\text{REF}})/k_B T} \right\rangle_{\text{REF}} \quad (1.22)$$

In one-step perturbation (OSP) approaches [18], REF can also be a nonphysical state that is designed to sample relevant conformational space for multiple real compounds. This works best if the compounds are closely related, so it is useful in the lead optimization stage. Note that in OSP, one is not restricted in the number of physical states the reference state represents, and the effect of small neutral substituents on a common scaffold can easily be screened for, e.g. a halogen scan. In practice, the method is more challenging if the differences between compounds involve larger changes in the charge distribution or the flexibility of the molecules.

The reference state in OSP needs to be selected carefully and may involve soft-core interactions, such that relevant conformations can be sampled for smaller and larger ligands in a single simulation. Alternative applications involve promiscuous stereocenters that sample multiple chemical configurations of a single scaffold. In practice,

it can be quite difficult to design the reference state such that it results in adequate phase space overlap for all ligands under investigation.

Another method that is based on the simulation of a single reference state is enveloping distribution sampling (EDS) [19]. The main difference with OSP is that EDS has an automated way of combining the Hamiltonians of the n ligands under investigation into the reference state. The n end-state Hamiltonians are combined into the reference state by Boltzmann weighting

$$H_{REF} = -k_B T \ln \left(\sum_{i=1}^n e^{-(H_i - \Delta F_i^R)/k_B T} \right) \quad (1.23)$$

where ΔF_i^R are free energy offset parameters. These ΔF_i^R correspond to the relative free energies of the end states if all states are sampled with equal probability. This complex energy surface is subsequently further smoothed by the use of a smoothing parameter or acceleration factors. The method typically requires two stages: a (set of) simulations in which the optimal parameters are derived, followed by a production simulation to obtain the free-energy differences.

1.3.7 Challenges in Alchemical Free Energy Calculations

The simplest functional form for connecting molecules A and B through the coupling parameter λ would be a linear combination of the two Hamiltonians

$$H(\lambda) = (1 - \lambda)H_A + \lambda H_B \quad (1.24)$$

However, in simulations where atoms are removed or added, this will cause major problems with the van der Waals interactions. When other atoms come close to the disappearing atom, E^{vdW} , which is an important nonbonded contribution of the Hamiltonians, becomes infinitely large, even at very small λ values. This restricts the sampling of configurational space. The same happens for the electrostatic interactions between atoms of equal charge. In order to address this problem, soft-core potentials are applied. The exact definition of the soft-core potential depends on the simulation software used. They all have in common that the singularity in the energy when the interatomic distance r_{ij} between two atoms approaches 0 is removed and the interaction function is smoothed for $\lambda \neq 0, 1$.

When alchemical changes involve a change in the net charge of the ligands, additional care needs to be taken. MD simulations are currently typically performed in explicit water under periodic boundary conditions, which means that the simulation systems are of the order of nanometers. At this scale, the simulation methodology will always treat the electrostatic interactions in an approximate way. One possibility is the use of cutoffs with reaction-field contributions to account for a homogeneous medium outside the cutoff sphere, and a second approximation is to use lattice summation to compute the electrostatic interaction for a strictly periodic molecular system. When a nonzero net charge is present during the simulation, these finite-size effects can lead to large charge-dependent artifacts that are propagated into the binding free energies [20]. To obtain results that do not depend on the size of the simulation box, one can either apply post-simulation

corrections [21, 22] or try to avoid the net charge changes. Under some conditions, the latter can, e.g. be done by simultaneously performing opposite charge changes on a counter ion [23, 24]. Addition of explicit ionic solution tends to further screen the remaining artifacts.

Generally, one will want to add or remove as few atoms as possible during an alchemical perturbation. However, this may not be the case if rings are involved. If one compound has a ring and the other is very similar, with just the difference that the ring is no longer closed, it may not always be a good idea to just break the bond. This perturbation will not converge since there will basically be no phase space overlap between the two compounds. Instead it is much better to remove all atoms involving the ring and let the other atoms appear, even if this means that many more atoms need to be added and removed. Several tools are available to determine the optimal perturbation pathway between two compounds [17, 25].

Water molecules can play an important part in binding as they are able to stabilize the interactions between the protein and ligand. When investigating several ligands in the same hydrated active site, it is possible that the optimal number of water molecules is different for different ligands [26]. This is important to keep in mind when relative binding free energy calculations are performed, especially when the active site is buried and the water molecules cannot easily move in or out of the active site. It might be necessary to alchemically make a water molecule disappear simultaneously with the changing ligand [27] or to perform the simulation in a grand-canonical ensemble [28], in which the number of water molecules in the active site may be adjusted during the simulations.

Alchemical methods can also be used to compute the full binding free energy of a ligand A, by defining the second molecule in the thermodynamic cycle as a noninteracting dummy molecule B. Effectively, the electrostatic and van der Waals interactions of the molecule of interest are scaled from fully interacting in state A to 0 in state B. For λ values close to 1, the lack of interactions will most likely lead to the molecule flying through the simulation box. In order to accelerate the convergence of the simulation of the (partly) decoupled ligand, restraints are applied to keep the ligand within the binding site. Several restraints can be used, i.e. a single distance restraint or additional angles and dihedrals can be used to restrain the ligand in its binding mode [29]. The obtained free energy difference subsequently needs to be corrected for the fact that a restraint was applied to the decoupled state, which can be done analytically.

Conformational changes of the protein (or ligand) during the perturbation can hamper the efficient convergence of the simulations. Longer simulation times might improve the situation if the free energy barrier between the conformations is not too high. Otherwise, enhanced sampling techniques can be applied to improve the convergence of the simulations. There are many techniques available; one of them is replica exchange molecular dynamics (REMD) [30]. Here, multiple noninteracting replicas are simulated simultaneously, each at a different temperature. The replicas at higher temperatures are more likely to overcome energy barriers, whereas replicas at a low temperature can get trapped in a local energy minimum. At certain intervals, an attempt is made to switch the configurations of two neighboring replicas.

The switch is then either accepted or rejected, according to the Metropolis criterion. This ensures that the copy with the lowest temperature (which is the original temperature of the system) also gets conformations for which the energetic barrier has been overcome. Instead of replicas with different temperatures, REMD can also be done with different Hamiltonians (HREMD), which can correspond to the end-states and the intermediate states along the free-energy calculation. When the source of the energetic barrier is known, precautions can be taken to ensure that the barrier is absent or easily crossed in at least one of the intermediate replicates.

Many different issues for alchemical methods remain, many of which are discussed in recent best-practices reviews [31, 32]. In addition, the performance of several free energy calculation methods has been evaluated in industrial drug design settings, offering an insight into real-life challenges [33–36].

1.4 Pathway Methods

The above methods give an insight into the (relative) binding free energy, the binding poses, and the interactions between the ligand and protein. However, there is no information on the binding process itself. Especially when the active site is buried, information on the binding path can also be very valuable, as it will give more insight into the binding kinetics, rather than just the binding thermodynamics. This is where pathway methods come into play [37, 38]. Probably the most intuitive way is to run an MD simulation of a solution with protein and ligand molecules present. Run this simulation long enough such that multiple binding and unbinding events are sampled and determine the binding free energy from the equilibrium binding constant

$$\Delta G_{bind} = -k_B T \ln K_{bind}^\circ = -k_B T \ln \left(\frac{[PL]C^\circ}{[P][L]} \right) \quad (1.25)$$

where K_{bind}° is the equilibrium constant of the reversible binding process, $[PL]$, $[P]$, and $[L]$ are the concentrations of the complex, the protein, and the ligand, respectively, and C° is the standard-state concentration. In practice, this is not as simple as it sounds. Because simulations typically only involve a single protein and a single ligand, this is intrinsically different from an equilibrium of many bound and unbound molecules, requiring additional adjustments to the sampled volume, and the fraction in the equation above should only contain the number of observed bound vs. unbound configurations $P_{bound}/P_{unbound}$ [39].

But there are more challenges to observe the actual binding equilibrium in a straightforward simulation. First of all, a large simulation box is required, in order to be able to sample ligand configurations that are not interacting with the protein at all. Second, the ligand can spend a lot of time moving through the (large) simulation box before it finds the binding site of the protein. Third, after the ligand is finally bound to the protein, the unbinding still needs to be sampled to really observe the binding equilibrium. In the case of a strong binder, this often takes prohibitively

long. All in all, the simulation time required to sample the reversible binding and unbinding events until equilibrium is reached is only very rarely feasible.

In order to speed up the binding and/or unbinding processes, additional restraining or pulling forces can be applied to the ligand. Starting from a bound configuration, one can define a reaction coordinate along which the ligand will move. This is usually the radial or linear distance between the centers of mass between the protein and the ligand, but much more elaborate coordinates can be designed. Unbinding can then be sampled by either a single simulation in which the ligand is gradually pulled toward a predefined free state (nonequilibrium simulation) or by multiple simulations with the ligand restrained to a slightly different part of the reaction coordinate (e.g. with umbrella sampling [US]). The nonequilibrium pulling simulations need a careful choice of pulling strength. It should be strong enough such that there is sufficient speed up in the simulation but not too strong, to avoid disruption of the protein structure. In order to obtain an equilibrium binding free energy estimate, the nonequilibrium simulations need to be repeated many times, and consequently, an exponential averaging according to the Jarzynski formalism needs to be performed.

In US, several intermediate states are generated in which the ligand is restrained to a different distance along the reaction coordinate. These biasing potentials make sure that unfavorable regions, as well as regions that require a conformational change, are properly sampled. When the individual umbrella simulations are converged and the phase space overlap between them is sufficient, the results are corrected for their biasing potentials, and the potential of mean force (PMF) can be constructed. Keeping the standard state correction in mind, the binding free energy can then be determined from the PMF. The efficiency of the US is very dependent on the system and the choice of the restraints. When neighboring umbrellas are too far apart, there is not enough phase space overlap, and the PMF is not properly converged. Similar problems occur when the restraining potential is too strong, such that only a very narrow range of distances is sampled during a simulation. Too weak restraints will cause the ligand to avoid regions with higher energy. Simulations at umbrellas that show conformational changes of the protein or the ligand can require very long simulations in order to reach convergence. This is especially the case with buried binding pockets where, i.e. amino acid side chains need to make space for the ligand to pass.

Pathway methods usually focus on the dissociation of the ligand from the protein. The advantage here is that it is not necessary to have knowledge about the binding path prior to the simulations. As mentioned above, the dissociation can be, e.g. enforced along a radial distance or a predefined path between the ligand and the (center of mass of the) active site. Enforcing the association process, however, is not so straightforward if the binding path is not known *a priori*. Gradually pulling the ligand to smaller radial distances is not very likely to result in binding. At a large radial distance, the ligand has a lot of space to move and can easily go to the other site of the protein. Pulling the ligand closer at this moment will just result in the ligand getting stuck at the surface of the protein, far away from the active site. Once a defined path is sampled and the simulations are converged, the binding free energy

can be calculated. Although this value will be correct, it is possible that alternative paths are available, which are not sampled here, and thus the free energy profile along the path might be wrong.

1.5 Final Thoughts

We have outlined some of the commonly used methods above to compute the binding free energy in the context of drug discovery and drug design. In the last decade, the methods have typically become much more user-friendly and are partially incorporated into large drug discovery pipelines and software packages. This is a development that is extremely satisfying from the point of view of academic method development and furthermore crucial to ensure a broad application of such methods in efficient drug design. Recent examples of computational workflows (which include free energy calculations) driving drug discovery forward include the identification of novel allosteric binders for KRAS^{G12C} [40], the discovery of potent noncovalent inhibitors of the main protease of SARS-CoV-2 [41], and lead optimization of an inhibitor of phosphodiesterase 2A (PDE2A) [42], and others [43].

However, we also emphasize that all of these methods come with their own set of approximations and limitations, which we have tried to highlight wherever appropriate. Even if everyday use of the methods is becoming easier, we feel it is crucial for the user to understand the background of the methods that are being applied and to be aware of the intrinsic limitations they come with [36]. This is the only way to ensure that a free energy is predicted that is appropriate for the problem at hand and that can be used to guide new experiments and new designs.

The list of methods is far from exhaustive. Many alternative methods, modifications to the ones described, and further improvements have been described. The aim of this work was not to give a full overview of all methods available, but rather to offer a starting point to understand the key principles in free-energy calculations. The interested reader is encouraged to check out the work in the further reading section below.

References

- 1 Chipot, C. and Pohorille, A. (eds.) (2007). *Free Energy Calculations. Theory and Applications in Chemistry and Biology*. Berlin, New York: Springer Verlag.
- 2 Srinivasan, J., Cheatham, T.E., Cieplak, P. et al. (1998). Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate–DNA helices. *J. Am. Chem. Soc.* 120 (37): 9401–9409. <https://doi.org/10.1021/ja981844+>.
- 3 Genheden, S. and Ryde, U. (2015). The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin. Drug Discov.* 10 (5): 449–461. <https://doi.org/10.1517/17460441.2015.1032936>.

- 4 Wang, E., Sun, H., Wang, J. et al. (2019). End-point binding free energy calculation with MM/PBSA and MM/GBSA: strategies and applications in drug design. *Chem. Rev.* <https://doi.org/10.1021/acs.chemrev.9b00055>.
- 5 King, E., Aitchison, E., Li, H., and Luo, R. (2021). Recent developments in free energy calculations for drug discovery. *Front. Mol. Biosci.* 8.
- 6 Lee, F.S., Chu, Z.-T., Bolger, M.B., and Warshel, A. (1992). Calculations of antibody-antigen interactions: microscopic and semi-microscopic evaluation of the free energies of binding of phosphorylcholine analogs to McPC603. *Protein Eng.* 5 (3): 215–228. <https://doi.org/10.1093/protein/5.3.215>.
- 7 Sham, Y.Y., Chu, Z.T., Tao, H., and Warshel, A. (2000). Examining methods for calculations of binding free energies: LRA, LIE, PDLD-LRA, and PDLD/S-LRA calculations of ligands binding to an HIV protease. *Proteins Struct. Funct. Bioinforma.* 39 (4): 393–407. [https://doi.org/10.1002/\(SICI\)1097-0134\(20000601\)39:4<393::AID-PROT120>3.0.CO;2-H](https://doi.org/10.1002/(SICI)1097-0134(20000601)39:4<393::AID-PROT120>3.0.CO;2-H).
- 8 Åqvist, J., Medina, C., and Samuelsson, J.-E. (1994). A new method for predicting binding affinity in computer-aided drug design. *Protein Eng.* 7 (3): 385–391. <https://doi.org/10.1093/protein/7.3.385>.
- 9 de Ruiter, A. and Oostenbrink, C. (2012). Efficient and accurate free energy calculations on trypsin inhibitors. *J. Chem. Theory Comput.* 3686–3695. <https://doi.org/10.1021/ct200750p>.
- 10 Kirkwood, J.G. (1935). Statistical mechanics of fluid mixtures. *J. Chem. Phys.* 3 (5): 300. <https://doi.org/10.1063/1.1749657>.
- 11 Bennett, C.H. (1976). Efficient estimation of free energy differences from Monte Carlo data. *J. Comput. Phys.* 22 (2): 245–268.
- 12 Zwanzig, R.W. (1954). High temperature equation of state by a perturbation method. I. Nonpolar gases. *J. Chem. Phys.* 22: 1420.
- 13 de Ruiter, A. and Oostenbrink, C. (2016). Extended thermodynamic integration: efficient prediction of lambda derivatives at nonsimulated points. *J. Chem. Theory Comput.* 12 (9): 4476–4486. <https://doi.org/10.1021/acs.jctc.6b00458>.
- 14 Shirts, M.R. and Chodera, J.D. (2008). Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.* 129 (12): 124105. <https://doi.org/10.1063/1.2978177>.
- 15 Jarzynski, C. (1997). Equilibrium free-energy differences from nonequilibrium measurements: a master-equation approach. *Phys. Rev. E* 56 (5): 5018. <https://doi.org/10.1103/PhysRevE.56.5018>.
- 16 Goette, M. and Grubmüller, H. (2009). Accuracy and convergence of free energy differences calculated from nonequilibrium switching processes. *J. Comput. Chem.* 30 (3): 447–456.
- 17 Liu, S., Wu, Y., Lin, T. et al. (2013). Lead optimization mapper: automating free energy calculations for lead optimization. *J. Comput. Aided Mol. Des.* 27 (9): <https://doi.org/10.1007/s10822-013-9678-y>.
- 18 Mark, A.E., Xu, Y., Liu, H., and van Gunsteren, W.F. (1995). Rapid non-empirical approaches for estimating relative binding free energies. *Acta Biochim. Pol.* 42 (4): 525–535.

- 19 Christ, C.D. and van Gunsteren, W.F. (2007). Enveloping distribution sampling: a method to calculate free energy differences from a single simulation. *J. Chem. Phys.* 126 (18): 184110. <https://doi.org/10.1063/1.2730508>.
- 20 Hunenberger, P.; Reif, M. Single-Ion Solvation; 2011. <https://doi.org/10.1039/9781849732222>.
- 21 Rocklin, G.J., Mobley, D.L., Dill, K.A., and Hünenberger, P.H. (2013). Calculating the binding free energies of charged species based on explicit-solvent simulations employing lattice-sum methods: an accurate correction scheme for electrostatic finite-size effects. *J. Chem. Phys.* 139 (18): 184103. <https://doi.org/10.1063/1.4826261>.
- 22 Reif, M.M. and Oostenbrink, C. (2014). Net charge changes in the calculation of relative ligand-binding free energies via classical atomistic molecular dynamics simulation. *J. Comput. Chem.* 35 (3): 227–243. <https://doi.org/10.1002/jcc.23490>.
- 23 Chen, W., Deng, Y., Russell, E. et al. (2018). Accurate calculation of relative binding free energies between ligands with different net charges. *J. Chem. Theory Comput.* 14 (12): 6346–6358. <https://doi.org/10.1021/acs.jctc.8b00825>.
- 24 Clark, A.J., Negron, C., Hauser, K. et al. (2019). Relative binding affinity prediction of charge-changing sequence mutations with FEP in protein–protein interfaces. *J. Mol. Biol.* 431 (7): 1481–1493. <https://doi.org/10.1016/j.jmb.2019.02.003>.
- 25 Petrov, D. (2021). Perturbation free-energy toolkit: an automated alchemical topology builder. *J. Chem. Inf. Model.* 61 (9): 4382–4390. <https://doi.org/10.1021/acs.jcim.1c00428>.
- 26 Bodnarchuk, M.S. (2016). Water, water, everywhere ... it's time to stop and think. *Drug Discov. Today* 21 (7): 1139–1146. <https://doi.org/10.1016/j.drudis.2016.05.009>.
- 27 Maurer, M., Hansen, N., and Oostenbrink, C. (2018). Comparison of free-energy methods using a tripeptide-water model system. *J. Comput. Chem.* 39 (26): 2226–2242. <https://doi.org/10.1002/jcc.25537>.
- 28 Bruce Macdonald, H.E., Cave-Ayland, C., Ross, G.A., and Essex, J.W. (2018). Ligand binding free energies with adaptive water networks: two-dimensional grand canonical alchemical perturbations. *J. Chem. Theory Comput.* 14 (12): 6586–6597. <https://doi.org/10.1021/acs.jctc.8b00614>.
- 29 Boresch, S., Tettinger, F., Leitgeb, M., and Karplus, M. (2003). Absolute binding free energies: a quantitative approach for their calculation. *J. Phys. Chem. B* 107 (35): 9535–9551. <https://doi.org/10.1021/jp0217839>.
- 30 Sugita, Y. and Okamoto, Y. (1999). Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* 314 (1–2): 141–151. [https://doi.org/10.1016/S0009-2614\(99\)01123-9](https://doi.org/10.1016/S0009-2614(99)01123-9).
- 31 Lee, T.-S., Allen, B.K., Giese, T.J. et al. (2020). Alchemical binding free energy calculations in AMBER20: advances and best practices for drug discovery. *J. Chem. Inf. Model.* 60 (11): 5595–5623. <https://doi.org/10.1021/acs.jcim.0c00613>.
- 32 Mey, A.S.J.S., Allen, B.K., McDonald, H.E.B. et al. (2020). Best practices for alchemical free energy calculations [Article v1.0]. *Living J. Comput. Mol. Sci.* 2 (1): 18378–18378. <https://doi.org/10.33011/livecoms.2.1.18378>.

- 33 Homeyer, N., Stoll, F., Hillisch, A., and Gohlke, H. (2014). Binding free energy calculations for lead optimization: assessment of their accuracy in an industrial drug design context. *J. Chem. Theory Comput.* 10 (8): 3331–3344. <https://doi.org/10.1021/ct5000296>.
- 34 Breznik, M., Ge, Y., Bluck, J.P. et al. (2022). Prioritizing small sets of molecules for synthesis through in-silico tools: a comparison of common ranking methods. *ChemMedChem* e202200425. <https://doi.org/10.1002/cmdc.202200425>.
- 35 Schindler, C.E.M., Baumann, H., Blum, A. et al. (2020). Large-scale assessment of binding free energy calculations in active drug discovery projects. *J. Chem. Inf. Model.* 60 (11): 5457–5474. <https://doi.org/10.1021/acs.jcim.0c00900>.
- 36 Meier, K., Bluck, J.P., and Christ, C.D. (2021). Use of free energy methods in the drug discovery industry. In: *Free Energy Methods in Drug Discovery: Current State and Future Directions; ACS Symposium Series*, vol. 1397, 39–66. American Chemical Society <https://doi.org/10.1021/bk-2021-1397.ch002>.
- 37 Dickson, A., Tiwary, P., and Vashisth, H. (2017). Kinetics of ligand binding through advanced computational approaches: a review. *Curr. Top. Med. Chem.* 17 (23): 2626–2641.
- 38 Bruce, N.J., Ganotra, G.K., Kokh, D.B. et al. (2018). New approaches for computing ligand–receptor binding kinetics. *Curr. Opin. Struct. Biol.* 49: 1–10. <https://doi.org/10.1016/j.sbi.2017.10.001>.
- 39 De Jong, D.H., Schäfer, L.V., De Vries, A.H. et al. (2011). Determining equilibrium constants for dimerization reactions from molecular dynamics simulations. *J. Comput. Chem.* 32 (9): 1919–1928. <https://doi.org/10.1002/jcc.21776>.
- 40 Mortier, J., Friberg, A., Badock, V. et al. (2020). Computationally empowered workflow identifies novel covalent allosteric binders for KRASG12C. *ChemMedChem* 15 (10): 827–832. <https://doi.org/10.1002/cmdc.201900727>.
- 41 Zhang, C.-H., Stone, E.A., Deshmukh, M. et al. (2021). Potent noncovalent inhibitors of the main protease of SARS-CoV-2 from molecular sculpting of the drug perampanel guided by free energy perturbation calculations. *ACS Cent. Sci.* 7 (3): 467–475. <https://doi.org/10.1021/acscentsci.1c00039>.
- 42 Tresadern, G., Velter, I., Trabanco, A.A. et al. (2020). [1,2,4]Triazolo[1,5-*a*]pyrimidine phosphodiesterase 2A inhibitors: structure and free-energy perturbation-guided exploration. *J. Med. Chem.* 63 (21): 12887–12910. <https://doi.org/10.1021/acs.jmedchem.0c01272>.
- 43 Abel, R. (2022). Advanced computational modeling accelerating small-molecule drug discovery. In: *Contemporary Accounts in Drug Discovery and Development*, 9–25. Wiley. <https://doi.org/10.1002/9781119627784.ch2>.

2

Gaussian Accelerated Molecular Dynamics in Drug Discovery

Hung N. Do, Jinan Wang, Keya Joshi, Kushal Koirala, and Yinglong Miao

Center for Computational Biology and Department of Molecular Biosciences, University of Kansas, Lawrence, KS 66047, USA

2.1 Introduction

Molecular dynamics (MD) is a powerful computational technique for simulating biomolecular dynamics at an atomistic level [1]. Due to advancements in computing hardware and software, timescales accessible to MD simulations have increased while costs have decreased [2, 3]. However, conventional MD (cMD), which makes no use of any enhanced sampling schemes, is often limited to tens to hundreds of microseconds [3–10] for simulations of biomolecular systems and cannot reach the timescales required to observe many biological processes of interest, which typically occur over milliseconds or longer, due to high energy barriers (e.g. 8–12 kcal/mol) [3–10].

Many enhanced sampling techniques have been developed during the last several decades to overcome the challenges mentioned above [11–15]. One class of enhanced sampling techniques uses predefined collective variables (CVs) or reaction coordinates (RCs), including umbrella sampling (US) [16, 17], metadynamics [18, 19], adaptive biasing force [20, 21], and steered MD [22]. However, it can be challenging to define proper CVs prior to simulation [3], and predefined CVs might significantly limit the sampling of conformational space during simulations [3]. Another class of enhanced sampling techniques, including replica exchange MD (REMD) [23, 24] or parallel tempering [25], self-guided Langevin MD [26–28], and accelerated MD (aMD) [29, 30], do not require predefined CVs. The latter class of unconstrained enhanced sampling techniques remains attractive to improve the sampling of biomolecular dynamics and obtain sufficient accuracy in free energy calculations.

Gaussian accelerated molecular dynamics (GaMD) is an unconstrained enhanced sampling technique that works by applying a harmonic boost potential to smooth the biomolecular potential energy surface [31]. Since this boost potential usually exhibits a near Gaussian distribution, cumulant expansion to the second order

(“Gaussian approximation”) can be applied to achieve proper energy reweighting [32]. GaMD allows for simultaneous, unconstrained enhanced sampling and free energy calculations of large biomolecules [31]. GaMD has been successfully demonstrated on enhanced sampling of ligand binding [31, 33–36], protein folding [31, 35], protein conformational changes [34, 37–40], protein-membrane [41], protein-protein [42–44], and protein-nucleic acid [45, 46] interactions.

Furthermore, GaMD has been combined with other enhanced sampling methods, including the REMD [47, 48] and weighted ensemble [49], to further improve conformational sampling and free energy calculations [3]. Notably, a novel multi-level enhanced sampling strategy was developed by combining GaMD, US, and a new adaptive sampling technique (AS) with the “dual-water” mode and AMOEBA [50–52] polarizable force field in TINKER-HP [53, 54] to accelerate the convergence of GaMD simulations [49]. In addition, “selective GaMD” algorithms, including ligand GaMD (LiGaMD) [55], peptide GaMD (Pep-GaMD) [56], and protein-protein interaction GaMD (PPI-GaMD) [44], have been developed to enable repetitive binding and dissociation of small-molecule ligands, highly flexible peptides, and proteins within microsecond simulations, which allow for highly efficient and accurate calculations of ligand/peptide/protein binding free energy and kinetic rate constants [3]. Furthermore, GaMD has been combined with deep learning (DL) and free energy profiling in a workflow (GLOW) to predict molecular determinants and map free energy landscapes of biomolecules [37].

Here, we review the principles and selected applications of GaMD in drug discovery. In particular, GaMD has been applied for advanced simulations of various biomolecular systems, including G-protein-coupled receptors (GPCRs), nucleic acids, and human angiotensin-converting enzyme 2 (ACE2) receptors.

2.2 Methods

2.2.1 Gaussian Accelerated Molecular Dynamics

Consider a system comprised of N atoms with their coordinates $r \equiv \{\vec{r}_1, \dots, \vec{r}_N\}$ and momenta $p \equiv \{\vec{p}_1, \dots, \vec{p}_N\}$. The system Hamiltonian can be expressed as:

$$H(r, p) = K(p) + V(r), \quad (2.1)$$

where $K(p)$ and $V(r)$ are the system’s kinetic and total potential energies, respectively. Next, we decompose the potential energy into the following terms:

$$V(r) = V_b(r) + V_{nb}(r) \quad (2.2)$$

where V_b and V_{nb} are the bonded and nonbonded potential energies, respectively. According to classic force fields, the bonded and nonbonded potential energies can be expressed as the following:

$$V_b = V_{bonds} + V_{angles} + V_{dih} + V_{improbers} + V_{UB} + V_{CMAP}, \quad (2.3)$$

$$V_{nb} = V_{elec} + V_{vdW}. \quad (2.4)$$

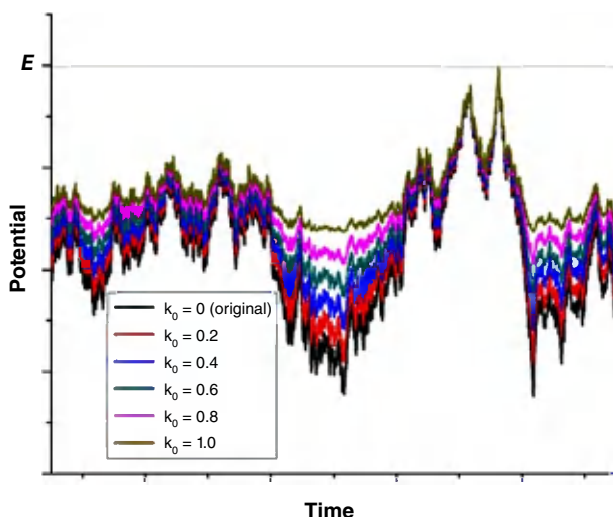


Figure 2.1 Schematic illustration of GaMD. When the threshold energy is set to the maximum potential ($E = V_{max}$), the system potential energy surface is smoothened by adding a harmonic boost potential that follows Gaussian distribution. The coefficient k_0 in the range of 0–1 determines the magnitude of the applied boost potential. With greater k_0 , higher boost potential is added to the original energy surface in cMD, which provides enhanced sampling of biomolecules across decreased energy barriers. Source: Reproduced with permission of Miao et al. [31]/American Chemical Society / Public Domain CC BY 3.0.

To enhance biomolecular conformational sampling, we will add boost potential based on these potential energetic terms (*bonds, angles, dih, impropers, UB, CMAP, elec, VDW*).

GaMD works by adding a harmonic boost potential to smooth the potential energy surface when the system potential drops below a reference energy E [31] (Figure 2.1):

$$\Delta V(r) = \begin{cases} \frac{1}{2}k(E - V(\vec{r}))^2, & V(\vec{r}) < E \\ 0, & V(\vec{r}) \geq E, \end{cases} \quad (2.5)$$

where k is the harmonic force constant. The two adjustable parameters E and k can be determined based on three enhanced sampling principles. First, for any two arbitrary potential values $V_1(\vec{r})$ and $V_2(\vec{r})$ found on the original energy surface, if $V_1(\vec{r}) < V_2(\vec{r})$, ΔV should be a monotonic function that does not change the relative order of the biased potential values; i.e. $V_1^*(\vec{r}) < V_2^*(\vec{r})$. Second, if $V_1(\vec{r}) < V_2(\vec{r})$, the potential difference observed on the smoothed energy surface is smaller than that of the original, i.e. $V_2^*(\vec{r}) - V_1^*(\vec{r}) < V_2(\vec{r}) - V_1(\vec{r})$. The reference energy needs to be set in the following range:

$$V_{max} \leq E \leq V_{min} + \frac{1}{k}, \quad (2.6)$$

where V_{max} and V_{min} are the system's minimum and maximum potential energies. To ensure that Eq. (2.6) is valid, k must satisfy: $k \leq \frac{1}{V_{max} - V_{min}}$. Let us define $k \equiv k_0 \frac{1}{V_{max} - V_{min}}$, then $0 \leq k_0 \leq 1$. Third, the standard deviation of ΔV needs to be

small enough (i.e. narrow distribution) to ensure proper energetic reweighting [32]: $\sigma_{\Delta V} = k(E - V_{\text{avg}})\sigma_V \leq \sigma_0$, where V_{avg} and σ_V are the average and standard deviation of the system's potential energies, $\sigma_{\Delta V}$ is the standard deviation of ΔV with σ_0 as a user-specified upper limit (e.g. $10k_B T$) for proper reweighting. When E is set to the lower bound $E = V_{\text{max}}$, k_0 can be calculated as:

$$k_0 = \min(1.0, k'_0) = \min\left(1.0, \frac{\sigma_0}{\sigma_V} \frac{V_{\text{max}} - V_{\text{min}}}{V_{\text{max}} - V_{\text{avg}}}\right), \quad (2.7)$$

Alternatively, when the threshold energy E is set to its upper bound $E \leq V_{\text{min}} + \frac{1}{k}$, k_0 is set to:

$$k_0 = k''_0 \equiv \left(1.0 - \frac{\sigma_0}{\sigma_V}\right) \frac{V_{\text{max}} - V_{\text{min}}}{V_{\text{avg}} - V_{\text{min}}}, \quad (2.8)$$

if k''_0 is found to be between 0 and 1. Otherwise, k_0 is calculated using Eq. (2.7).

2.2.2 Ligand Gaussian Accelerated Molecular Dynamics

On the basis of GaMD, selective GaMD algorithms, including LiGaMD [55], have been developed for more efficient simulations and calculations of both free energy and kinetics of biological processes [3]. LiGaMD, in particular, allows for simulations of repetitive binding and dissociation of small-molecule ligands for calculating ligand binding free energies and kinetics [3, 55]. Given a simulation system consisting of ligand L , protein P and biological environment E , the system potential energy $V(r)$ from Eq. (2.1) could be decomposed into the following terms:

$$\begin{aligned} V(r) = & V_{P,b}(r_P) + V_{L,b}(r_L) + V_{E,b}(r_E) + V_{PP,nb}(r_P) + V_{LL,nb}(r_L) + V_{EE,nb}(r_E) \\ & + V_{PL,nb}(r_{PL}) + V_{PE,nb}(r_{PE}) + V_{LE,nb}(r_{LE}) \end{aligned} \quad (2.9)$$

where $V_{P,b}$, $V_{L,b}$, and $V_{E,b}$ are the bonded potential energies in protein P , ligand L , and environment E , respectively. $V_{PP,nb}$, $V_{LL,nb}$, and $V_{EE,nb}$ are the self-nonbonded potential energies in protein P , ligand L , and environment E , respectively. $V_{PL,nb}$, $V_{PE,nb}$, and $V_{LE,nb}$ are the nonbonded interaction energies between $P-L$, $P-E$, and $L-E$, respectively. On the other hand, the nonbonded potential energies are usually calculated as:

$$V_{nb} = V_{\text{elec}} + V_{\text{vdW}} \quad (2.10)$$

where V_{elec} and V_{vdW} are the system's electrostatic and van der Waals potential energies. In general, ligand binding involves mostly the nonbonded interaction energies of the ligand ($V_{LL,nb}$, $V_{PL,nb}$, and $V_{LE,nb}$). Therefore, a selective boost potential is added to the ligand nonbonded potential energy based on the GaMD algorithm [31]:

$$\Delta V_{L,nb}(r) = \begin{cases} \frac{1}{2}k_{L,nb}(E_{L,nb} - V_{L,nb}(r))^2, & V_{L,nb}(r) < E_{L,nb} \\ 0, & V_{L,nb}(r) \geq E_{L,nb} \end{cases} \quad (2.11)$$

where $E_{L,nb}$ is the threshold energy for applying boost potential and $k_{L,nb}$ is the harmonic constant.

In addition, multiple ligand molecules can be added to the solvent to facilitate ligand binding to proteins in MD simulations [3]. The higher the ligand concentration, the faster the ligand binds, as long as the ligand concentration is still within its solubility concentration [3]. Therefore, besides the selective boost added to the bound ligand, another boost potential could be applied to the unbound ligand molecules, protein, and environment to facilitate ligand dissociation and rebinding [3, 55]. The second boost potential is calculated using the total system potential energy excluding the nonbonded potential energy of the bound ligand as:

$$\Delta V_D(r) = \begin{cases} \frac{1}{2}k_D(E_D - V_D(r))^2, & V_D(r) < E_D \\ 0, & V_D(r) \geq E_D \end{cases} \quad (2.12)$$

where V_D is the total system potential energy excluding the nonbonded potential energy of the bound ligand, E_D is the threshold energy for applying the second boost potential, and k_D is the harmonic constant. Therefore, dual-boost LiGaMD has a total boost potential of $\Delta V(r) = \Delta V_{L,nb}(r) + \Delta V_D(r)$ [3, 55].

2.2.3 Energetic Reweighting of GaMD for Free Energy Calculations

For energetic reweighting of GaMD simulations, the probability distribution along a selected RC can be calculated from simulations as $p^*(A)$. Given the boost potential $\Delta V(r)$ of each frame in GaMD simulations, $p^*(A)$ can be reweighted to recover the canonical ensemble distribution, $p(A)$, as:

$$p(A_j) = p^*(A_j) \frac{\langle e^{\beta \Delta V(\bar{r})} \rangle_j}{\sum_{i=1}^M \langle p^*(A_i) e^{\beta \Delta V(\bar{r})} \rangle_i}, j = 1, \dots, M \quad (2.13)$$

where M is the number of bins, $\beta = k_B T$ and $\langle e^{\beta \Delta V(\bar{r})} \rangle_j$ is the ensemble-averaged Boltzmann factor of $\Delta V(\bar{r})$ for simulation frames found in the j^{th} bin. The ensemble-averaged reweighting factor can be approximated using cumulant expansion [31, 32]:

$$\langle e^{\beta \Delta V(\bar{r})} \rangle = \exp \left\{ \sum_{k=1}^{\infty} \frac{\beta^k}{k!} C_k \right\}, \quad (2.14)$$

where the first three cumulants are given by:

$$\begin{aligned} C_1 &= \langle \Delta V \rangle, \\ C_2 &= \langle \Delta V^2 \rangle - \langle \Delta V \rangle^2 = \sigma_{\Delta V}^2, \\ C_3 &= \langle \Delta V^3 \rangle - 3\langle \Delta V \rangle^2 \langle \Delta V \rangle + 2\langle \Delta V \rangle^3, \end{aligned} \quad (2.15)$$

The boost potential obtained from GaMD simulations usually shows near-Gaussian distribution [57]. Cumulant expansion to the second order thus provides a good approximation for computing the reweighting factor [31, 32]. The reweighted free energy $F(A) = -k_B T \ln p(A)$ is calculated as:

$$F(A) = F^*(A) - \sum_{k=1}^2 \frac{\beta^k}{k!} C_k + F_c, \quad (2.16)$$

where $F^*(A) = -k_B T \ln p^*(A)$ is the modified free energy obtained from GaMD simulation and F_c is a constant.

To characterize the extent to which ΔV follows a Gaussian distribution, its distribution anharmonicity γ is calculated as [32]:

$$\gamma = S_{max} - S_{\Delta V} = \frac{1}{2} \ln(2\pi e \sigma_{\Delta V}^2) + \int_0^\infty p(\Delta V) \ln(p(\Delta V)) d\Delta V \quad (2.17)$$

where ΔV is dimensionless as divided by $k_B T$ with k_B and T being the Boltzmann constant and system temperature, respectively, and $S_{max} = \frac{1}{2} \ln(2\pi e \sigma_{\Delta V}^2)$ is the maximum entropy of ΔV [32]. When γ is zero, ΔV follows the exact Gaussian distribution with sufficient sampling. Reweighting by approximating the exponential average term with cumulant expansion to the second order is able to accurately recover the original free energy landscape. As γ increases, the ΔV distribution becomes less harmonic, and the reweighted free energy profile obtained from cumulant expansion to the second order would deviate from the original. The anharmonicity of ΔV distribution serves as an indicator of the enhanced sampling convergence and accuracy of the reweighted free energy.

2.2.4 GLOW: A Workflow Integrating Gaussian Accelerated Molecular Dynamics and Deep Learning for Free Energy Profiling

Recently, GaMD and DL were integrated to develop GLOW—a workflow to identify important RCs and map free energy profiles of biomolecules [37]. First, dual-boost GaMD simulations are performed on the biomolecules of interest. Since simulation trajectories are collections of static PDB snapshots, the residue contact map of each simulation frame can be calculated and transformed into images. The specialized type of neural network for image classification, the two-dimensional (2D) convolutional neural network (CNN), is employed to classify the residue contact maps of target biomolecules. The default DL architecture of GLOW consists of four convolutional layers of 3×3 kernel size, with 32, 32, 64, and 64 filters, respectively, followed by three fully connected (dense) layers, the first two of which include 512 and 128 filters with a dropout rate of 0.5 each. The final fully connected layer is the classification layer. “ReLU” activation is used for all layers in the 2D CNN, except the classification layer, in which “softmax” activation is used. A maximum pooling layer of 2×2 kernel size is added after each convolutional layer [37]. The saliency (attention) map of residue contact gradients is calculated through backpropagation by vanilla gradient-based attribution [58] using the residue contact map of the most populated structural cluster of each system, from which important RCs can be identified. Finally, the free energy profiles of these RCs are calculated through reweighting of GaMD simulations to characterize the biomolecular systems of interest.

2.2.5 Binding Kinetics Obtained from Reweighting of GaMD Simulations

Provided sufficient sampling of repetitive protein dissociation and binding in the simulations, we recorded the time periods and calculated their averages for the

ligand sampled in the bound (τ_B) and unbound (τ_U) states from the simulation trajectories. The τ_B corresponds to the protein residence time. Then, the ligand dissociation and binding rate constants (k_{off} and k_{on}) were calculated as:

$$k_{\text{off}} = \frac{1}{\tau_B} \quad (2.18)$$

$$k_{\text{on}} = \frac{1}{\tau_U \cdot [L]} \quad (2.19)$$

where $[L]$ is the ligand concentration in the simulation system.

According to Kramers' rate theory, the rate of a chemical reaction in the large viscosity limit is calculated as [59]:

$$k_R \cong \frac{w_m w_b}{2\pi\xi} e^{-\Delta F/k_B T} \quad (2.20)$$

where w_m and w_b are frequencies of the approximated harmonic oscillators (also referred to as curvatures of free energy surface [60, 61]) near the energy minimum and barrier, respectively, ξ is the frictional rate constant and ΔF is the free energy barrier of transition. The friction constant ξ is related to the diffusion coefficient D with $\xi = k_B T/D$. The apparent diffusion coefficient D can be obtained by dividing the kinetic rate calculated directly using the transition time series collected directly from simulations by that using the probability density solution of the Smoluchowski equation [62]. In order to reweight protein kinetics from the GaMD simulations using the Kramers' rate theory, the free energy barriers of protein binding and dissociation are calculated from the original (reweighted, ΔF) and modified (no reweighting, ΔF^*) PMF profiles, similarly for curvatures of the reweighed (w) and modified (w^* , no reweighting) PMF profiles near the protein bound ("B") and unbound ("U") low-energy wells and the energy barrier ("Br"), and the ratio of apparent diffusion coefficients from simulations without reweighting (modified, D^*) and with reweighting (D). The resulting numbers are then plugged into Eq. (2.20) to estimate accelerations of the ligand binding and dissociation rates during GaMD simulations [59], which allows us to recover the original kinetic rate constants.

2.2.6 Gaussian Accelerated Molecular Dynamics Implementations and Software

GaMD has been implemented in widely used simulation packages including AMBER [31], NAMD [35], GENESIS [48], TINKER-HP [63], and OpenMM [64]. Overall, GaMD simulations consist of three main stages [55], including short cMD, GaMD equilibration, and GaMD production. During short cMD, a number of preparatory steps are first performed to equilibrate the system, and then the potential statistics (including V_{max} , V_{min} , V_{avg} , and σ_V) are collected [55]. During the GaMD pre-equilibration stage, boost potential is applied, but no boost parameters are updated [55]. In the GaMD equilibration stage, boost potential continues to be applied, and boost parameters are updated [55]. In the final stage of GaMD production, boost potential is applied, and the boost parameters are held fixed [55]. In a typical GaMD simulation, users need to specify the number of simulation

steps for each stage as well as the number of simulation steps used to calculate the average and standard deviation of potential energies, a flag to apply boost potential, a flag to set the threshold energy E for applying boost potentials, a flag to restart GMD simulation, and the upper limits of the standard deviations of the first and second boost potentials. Additional resources related to GaMD can be found here: <https://www.med.unc.edu/pharm/miaolab/resources/GaMD>

2.3 Applications

2.3.1 G-Protein-Coupled Receptors

GPCRs are the largest family of human membrane proteins and the primary targets of ~34% of currently marketed drugs [65]. On the basis of sequence homology and functional similarity, GPCRs are classified into six different classes, including class A (Rhodopsin-like), class B (secretin receptors), which is further divided into subclasses of B1 (classical hormone receptors), B2 (adhesion GPCRs), B3 (methuselah-type receptors), class C (metabotropic glutamate receptors), class D (fungal mating pheromone receptors), class E (cyclic AMP receptors), and class F (frizzled/TAS2 receptors) [66, 67]. GPCRs share a characteristic structural fold of seven transmembrane (TM) α -helices (TM1-TM7), connected by three extracellular loops (ECL1-ECL3) and three intracellular loops (ICL1-ICL3). Next, we will review recent applications of GaMD in the studies of GPCRs (including adenosine (ADO) and muscarinic acetylcholine receptors).

2.3.1.1 Characterizing the Binding and Unbinding of Caffeine in Human Adenosine A_{2A} Receptor

Adenosine receptors (ARs) are a subfamily of class A GPCRs with ADO as the endogenous ligand [68], with four known subtypes: A_1 AR, A_{2A} AR, A_{2B} AR, and A_3 AR [69]. Despite their broad distribution in human tissues and functional differences, ARs share common antagonists of caffeine (CFF) and theophylline, both of which antagonize the receptors upon binding. Sequence alignment shows that the seven TM helix bundles of the A_1 AR share high similarity with A_{2A} AR by 71%, A_{2B} AR by 70%, and A_3 AR by 77%. Their sequence similarity is significantly reduced in the ECLs, being 43% for A_{2A} AR, 45% for A_{2B} AR, and 35% for A_3 AR when compared with A_1 AR [33].

All-atom GaMD simulations were performed to determine the binding and dissociation pathways of CFF to the human A_{2A} AR [33], starting from the X-ray structure of A_{2A} AR in complex with CFF (PDB: 5MZZ) [70]. The T4-lysozyme, lipid molecules, CFF, water, and heteroatom molecules were removed. A total of 10 CFF ligand molecules were placed randomly at a distance $>15\text{\AA}$ from the extracellular surface of the A_{2A} AR. Spontaneous binding and dissociation of CFF in the receptor were successfully captured in the GaMD simulations [33]. A dominant binding pathway of CFF to the A_{2A} AR was identified from the 63-ns GaMD equilibration (Figure 2.2a). CFF approached the A_{2A} AR through interactions with ECL2, the

extracellular mouth between ECL2, ECL3, and TM7, and finally, the receptor orthosteric site located deeply within the receptor TM bundle (Figure 2.2d). A slightly different binding pathway was observed when the orthosteric pocket of the A_{2A}AR was already occupied by one CFF molecule in Sim2 (Figure 2.2b). In this pathway, the second CFF first explored a region between ECL3 and TM7 during the binding process (Figure 2.2e). The dissociation pathway of CFF was mostly the reverse of the dominant binding pathway (Figure 2.2c,f).

2.3.1.2 Unraveling the Allosteric Modulation of Human A₁ Adenosine Receptor

Recently, we investigated the allosteric effect of a newly discovered PAM MIPS521 in the A₁AR [71]. MIPS521 was identified to bind a novel lipid-facing pocket of the A1R and hardly changed the cryo-EM structure of the ADO-A1R-G_{i2} protein complex. Therefore, using the latest cryo-EM structures of A₁AR, we performed further GaMD simulations on four systems: ADO-A₁AR-G_{i2}-MIPS521, A₁AR-G_{i2}, ADO-A₁AR-MIPS521 and ADO-A₁AR (Figure 2.3) [71]. Three replicas of GaMD simulations lasting 500 ns or 1000 ns were performed. MIPS521 underwent high fluctuations in the ADO-A₁AR-MIPS521 system and could even dissociate in the GaMD simulations, while it remained bound to the ADO-A₁AR-G_{i2} complex during both the GaMD and cMD simulations. We then focused on the dynamics of the agonist, ADO. In the absence of G protein, ADO sampled a large conformational space in the orthosteric pocket and exhibited higher flexibilities in simulations with and without MIPS521 (Figure 2.3b,c). The presence of G protein decreased the conformational dynamics of ADO, consistent with a ternary complex model where the G protein allosterically stabilizes agonist binding in the orthosteric pocket (Figure 2.3d) [71]. In the presence of MIPS521, ADO was stabilized even further when G_{i2} was present in the A1R complex (Figure 2.3e). Next, we examined whether the effect of MIPS521 on ADO stability in the orthosteric pocket resulted from changes in receptor and G protein dynamics in the simulations. In the absence of G protein, the active receptor in the ADO-A₁AR GaMD simulations relaxed toward the inactive structure, for which the R105^{3,50}-E229^{6,30} distance could decrease to ~8.3 Å (Figure 2.3f) [71]. Simulations in the presence of G protein, MIPS521, or both revealed TM3-6 distances consistent with the A₁AR in the active conformation, suggesting a direct stabilization of the A₁AR in a “G protein-bound-like” conformation by MIPS521 post-removal of G_{i2} (Figure 2.3g-i).

Finally, GLOW was applied to characterize GPCR activation and allosteric modulation, using the A₁AR as a model system [37]. GLOW characterization of GPCR activation was obtained through the classification of the A₁AR bound by “Antagonist” (PSB36), “Agonist” (adenosine), and “Agonist – Gi.” GLOW achieved an overall accuracy of 99.34% and a loss of 1.85% on the validation data set after 15 epochs. GLOW revealed that the removal of the G_{i2} protein from the A₁AR bound by “Agonist-Gi” led to the deactivation of the A₁AR. The intracellular halves of TM3, TM5, TM6, and TM7 underwent significant conformational changes. In particular, TM6 drew closer to TM3, while TM7 moved away from TM3 and TM5 [37]. The receptor TM5 and TM6 intracellular domains and helix 8 also became more flexible in the absence of the G_{i2} protein. Replacement of “Agonist” with

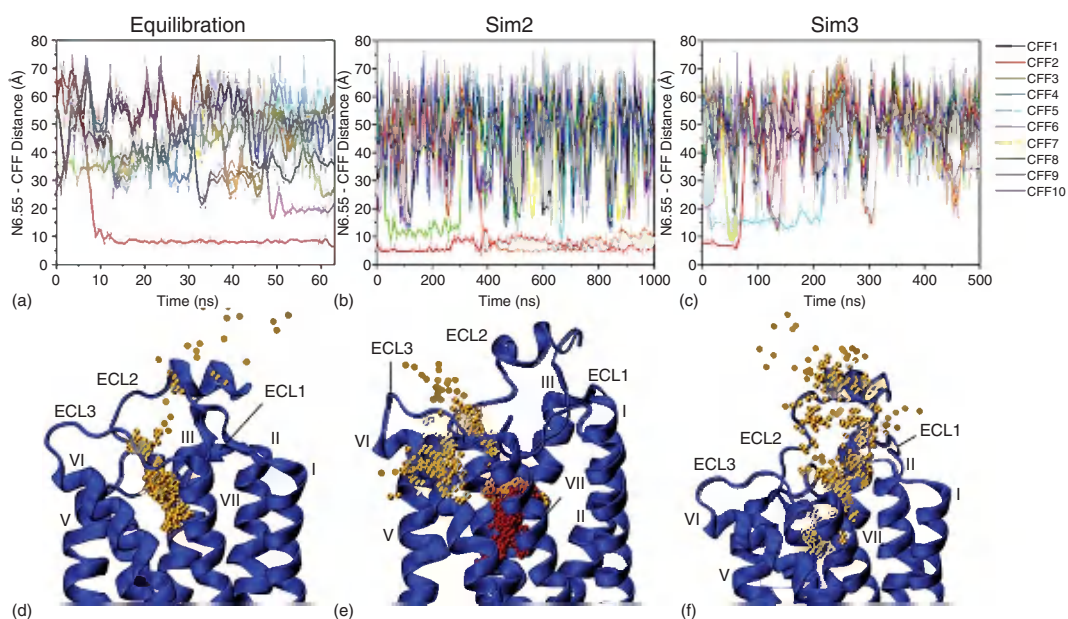


Figure 2.2 Binding and dissociation pathways of caffeine (CFF) from the A_{2A}AR revealed from GaMD simulations. (a–c) Time courses of the distance between receptor residue N6.55 atom ND2 and CFF atom N1 calculated from GaMD equilibration, Sim2, and Sim3 GaMD production simulations. (d–f) Trace of CFF (orange and red) in the A_{2A}AR observed in the GaMD equilibration, Sim2, and Sim3 GaMD production simulations. The seven transmembrane helices are labeled I–VII, and extracellular loops 1–3 are labeled ECL1–ECL3. Source: Adapted with permission from Do et al. [33]. Copyright 2021 Do, Akhter, and Miao. <https://www.frontiersin.org/articles/10.3389/fmolb.2021.673170/full>. Further permissions related to the material excerpted should be directed to Frontiers.

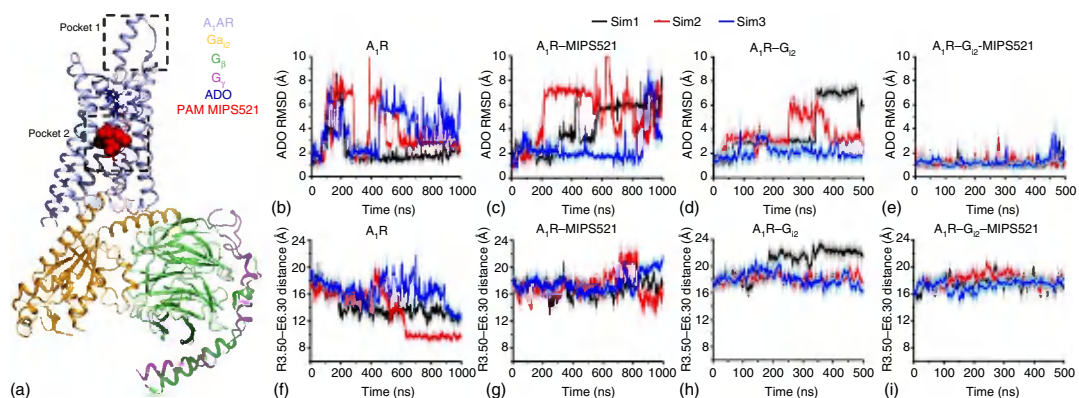


Figure 2.3 Effects of allosteric drug leads on the human adenosine receptor A₁AR. (a) Allosteric binding sites (pocket 1 and pocket 2) in the A₁AR. (B-E) RMSD (Å) of adenosine (ADO) orthosteric ligand calculated from GaMD simulations in the presence (b) and absence (c) of the allosteric drug MIP5521, G₁₂ (d), or both (e). (f–i) Distance between the intracellular ends of TM3 and TM6 (measured as the distance between charge centers of residues R3.50 and E6.30) in the absence (f) or presence (g) of MIP5521, G₁₂ (h) or both (i). Source: Reproduced with permission of Draper-Joyce et al. [71]/Springer Nature.

“Antagonist” in A₁AR led to the complete closure of the intracellular pocket, where TM3 formed intracellular residue contacts with TM6 and the NPxxY motif in TM7 moved away from TM3 and TM5. In addition, the ligand-binding extracellular domains and intracellular G-protein binding domains were found to be loosely coupled in GPCR activation [37, 72]. A₁AR sampled a more “open” conformational state of the extracellular mouth in the “Agonist-Gi” bound system due to mostly increased flexibility of the ECL1 and ECL2. Removal of the G_{i2} protein from the “Agonist-Gi”-A₁AR complex reduced the conformational space of the receptor ECL2, while replacement of “Agonist” by “Antagonist” confined the ECL2 significantly [37]. Characterization of GPCR allosteric modulation was done through the classification of the A₁AR bound by “Agonist-Gi” and “Agonist-Gi-PAM,” with the PAM being MIPS521. GLOW achieved an overall accuracy of 99.27% and a loss of 1.78% on the validation data set after 15 epochs. Furthermore, GLOW showed that ECL2 played a critical role in the allosteric modulation of A₁AR [37], which is consistent with previous mutagenesis, structure, and molecule modeling studies [71, 73–76]. GLOW revealed that binding of a PAM (MIPS521) to the agonist-Gi-A₁AR complex biased the receptor conformational ensemble, especially in the ECL1 and ECL2 regions. PAM binding stabilized agonist binding within the orthosteric pocket of A₁AR, which confined the extracellular mouth of the receptor. Furthermore, PAM binding disrupted the N148^{ECL2}-V152^{ECL2} α -helical hydrogen bond and distorted this portion of the ECL2 helix [37]. A small pocket was formed in the ECL2-TM5-ECL3 region because of this ECL2 helix distortion. Taken together, GaMD simulations provided mechanistic insight into the residue level of the positive cooperativity of MIPS521 in the A₁AR-G_{i2} complex.

2.3.1.3 Ensemble Based Virtual Screening of Allosteric Modulators of Human A₁ Adenosine Receptor

Virtual screening has been widely used for agonist/antagonist design targeting GPCRs [77]. However, it is rather challenging to apply virtual screening to identify allosteric modulators due to their low affinity compared with the agonist/antagonist. Recently, retrospective ensemble docking calculations of PAMs to the A₁AR combining GaMD simulations and Autodock [78] were performed [79]. Receptor structural ensembles obtained from GaMD simulations were used to increase the docking performance of known PAMs using the A₁AR as a model GPCR.

The GaMD simulations implemented in AMBER [31] and NAMD [35] were applied to generate receptor ensembles. The flexible docking and rigid body docking at different levels (i.e. short, medium, and long) were all evaluated. Docking scores corrected by the GaMD-reweighted free energy of the receptor structural cluster further improved the docking performances. The calculated docking enrichment factors and the area under the receiver operating characteristic curves are increased using ranking by the average binding energy in comparison with the minimum binding energy. Ensembles obtained from AMBER dual-boost GaMD simulations of the VCP171-bound ADO-A₁AR-Gi complex outperformed other ensembles for docking. Interactions between the PAM and receptor ECL2 in the VCP171-bound ADO-A₁AR-Gi complex might induce more suitable conformations

for PAM binding, which were difficult to be sampled in the simulations of PAM-free (i.e. apo) A₁AR. Dual-boost GaMD with higher boost potential was observed to perform better than the dihedral-boost GaMD for ensemble docking. Overall, flexible docking performed significantly better than rigid-body docking at different levels with AutoDock, suggesting that the flexibility of protein side chains is also important in ensemble docking. In summary, docking performance has been highly improved by combining GaMD simulations with flexible docking, which effectively accounts for the flexibility of the backbone and side chains in receptors. Such an ensemble docking protocol will greatly facilitate future PAM design of the A₁AR and other GPCRs [3].

2.3.2 Nucleic Acids

2.3.2.1 Exploring the Binding of Risdiplam Splicing Drug Analog to Single-Stranded RNA

Risdiplam is the first approved small-molecule splicing drug for the treatment of spinal muscular atrophy (SMA). Five 500 ns GaMD simulations were performed to explore the binding interaction of a known risdiplam analog, SMN-C2, which binds to the target nucleic acids with GA-rich sequence [36]. The simulation structures of (Seq6) of DNA and RNA were built using NAB in the AmberTools [80] package. The ligand molecule was placed randomly at >15 Å away from the nucleic acid. The AMBER force field BSC1 [81] parameter sets were used for DNA, OL3 [82] for RNA, GAFF2 [83] for ligand, and TIP3P [84] for water in the system. Spontaneous binding of the SMN-C2 to both RNA and DNA Seq6 was observed in the GaMD simulations [36]. For the binding of SMN-C2 to the RNA Seq6, three low-energy conformational states were identified, including the “Unbound/Unfolded,” “Bound/Intermediate” and “Bound/Folded” states (Figure 2.4a). SMN-C2 was able to interact with RNA Seq6 in two bound states (Figure 2.4b). When bound to RNA Seq6, the COM distance between RNA and ligand is reduced to ~6 Å and the RNA radius of gyration (*R_g*) reduced to ~8.0 Å. While for the binding of SMN-C2 to the DNA Seq6, the location of the bound small molecule was slightly different from that observed in RNA, as demonstrated by the COM distance between DNA and ligand being reduced to ~4 Å in the bound states [36] (Figure 2.4b). Three low-energy conformational states of the DNA-ligand system were also identified for DNA Seq6, including the “Bound/Unfolded,” “Intermediate,” and “Bound/Folded” states (Figure 2.4c). In the Bound/Folded state, a similar binding mode of SMN-C2 was observed in DNA as in RNA, with subtle differences [36] (Figure 2.4d). In both RNA and DNA, it appeared that the AAG trinucleotide in the GAAG motif is thus important for the binding of SMN-C2 through π -stacking interactions. Furthermore, the unfolded RNA Seq6 appeared to be more flexible than the unfolded DNA, and SMN-C2 did not spontaneously bind to the unfolded RNA [36].

2.3.2.2 Uncovering the Binding of RNA to a Musashi RNA-Binding Protein

The Musashi 1 (MSI1) protein serves as a therapeutic drug target for treating several cancers, such as colorectal, ovarian, bladder, and myeloid leukemia. It is known

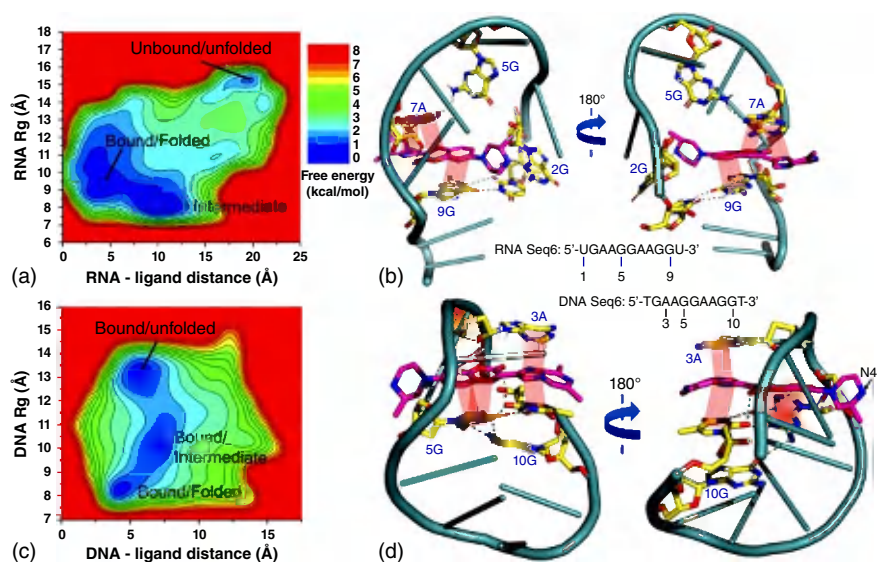


Figure 2.4 GaMD simulations revealed spontaneous binding of risdiplam to RNA and DNA Seq6. (a) 2D free energy profile of the center-of-mass (COM) distance between RNA-ligand and RNA radius of gyration (Rg). Three low-energy states were identified, namely the Unbound/Unfolded, Intermediate, and Bound/Folded. (b) Representative conformation of risdiplam-bound RNA Seq6 in the folded state. (c) 2D free energy profile of the COM distance between DNA-ligand and DNA radius of gyration (Rg). Three low-energy states were identified, namely the Unbound/Unfolded, Intermediate, and Bound/Folded. (d) Representative conformation of risdiplam-bound DNA Seq6 in the folded state. The color scheme is as follows: magenta = risdiplam, yellow = interacting nucleotides, cyan = other nucleotides, green dashed line = polar interaction, light red shade = π - π or lone pair- π stacking. Source: Reproduced with permission of Tang et al. [36]/Oxford University Press/Public Domain CC BY 4.0.

to bind and suppress translation of the 3'-UTR of Numb mRNA [86], however, the molecular mechanism of this interaction remains elusive, which is important for effective drug design. GaMD simulations were performed to study the binding mechanism between Numb RNA and MSI1 [85]. For system setup, the Numb RNA was placed ~ 30 Å away from the MSI1. The AMBER force fields were used with ff14SBonlysc for protein, RNA.LJbb [82] for RNA, and TIP3P [84] model for water molecules. Spontaneous binding of Numb RNA to the MSI1 protein was successfully captured in 6 out of the 19 independent 1200 ns of GaMD simulations. In Sim1, RNA binding was observed at ~ 100 ns, where the RMSD of RNA relative to the NMR structure reduced to ~ 2.50 Å. In Sim2, RNA binding was observed at ~ 1010 – 1130 ns followed by RNA dissociation into the bulk solvent. The RNA bound to MSI1 after ~ 800 ns in Sim3, Sim4, and Sim5. In Sim6, spontaneous binding of RNA was observed after ~ 1000 ns. Five low-energy minima were characterized from GaMD simulations, including the “Bound,” “Intermediate I1,” “Intermediate I2,” “Intermediate I3,” and “Unbound” (Figure 2.5). These states were identified at the backbone RMSD of the RNA core and N_{contacts} as (2.0 Å, 1500), (5.2 Å, 480), (9.5 Å, 200), (25.0 Å, 10), and (40 Å, 0), respectively. The “Unbound,” “Intermediate I1,”

“Intermediate I2,” and “Intermediate I3” states were identified at the backbone RMSD of the RNA core and R_g of Numb being (40.0 Å, 6.2 Å), (5.0 Å, 7.2 Å), and (6.9 Å, 6.2 Å), respectively (Figure 2.5a–d). From the GaMD simulations, the R_g of the Numb RNA in the “Bound” state was observed to have a wider range as compared to the “Unbound” and “Intermediate” conformations, which suggested an induced fit mechanism of Numb RNA binding to MSI1. The I1 state showed interactions of Numb RNA with the β 2- β 3 loop and C terminus of MSI1. Three hydrogen bonds were formed between MS1 C terminus residue R99 and the RNA nucleotide A106 (Figure 2.5e). The I2 state showed flipping of the MS1 residue R61 sidechain toward the solvent, leading to the formation of hydrogen bonds and salt-bridge interactions with the phosphate oxygen of the sidechain and backbone of RNA nucleotide A106, respectively (Figure 2.5f). The I3 state showed large conformational changes in the MS1 C terminus, where hydrogen bond and salt bridge interactions were observed between the C terminus residue R99 and the sidechain and backbone of RNA nucleotide A106, respectively (Figure 2.5g). These important understandings of the RNA binding mechanism to MSI1 provided by GaMD simulations can aid rational structure-based drug design against MSI1 and other related diseases.

2.3.3 Human Angiotensin-Converting Enzyme 2 Receptor

Angiotensin-converting enzyme 2 (ACE2), which helps to convert angiotensin II to angiotensin 1–7 and angiotensin to angiotensin 1–9 [87], has been identified as the function receptor for severe acute respiratory syndrome coronaviruses, including SARS-CoV and SARS-CoV-2 [88]. SARS-CoV-2 is responsible for the 2019 coronavirus pandemic (COVID-19). The entry of SARS-CoV-2 is mediated by the interaction of the receptor binding domain (RBD) in the virus spike protein S1 subunit with the host ACE2 receptor. Therefore, inhibiting the interaction between the viral RBD and host ACE2 presents a promising strategy for blocking SARS-CoV-2 entry into human cells [34].

LiGaMD simulations were performed to explore the binding mechanism of MLN-4760, a highly selective and potent inhibitor, in the human ACE2 receptor [34], starting from the 1R4L [89] PDB structure. Besides the bound ligand, nine additional ligands were placed randomly in the solvent. During the equilibration phase, the bound MLN-4760 was observed to dissociate from ACE2 active site to the bulk solvent after the conformational changes in the subdomain I of ACE2 [34]. LiGaMD simulations were further extended to ten 700–2000 ns independent production simulations with randomized initial atomic velocities. Repetitive binding and unbinding of a ligand molecule to ACE2 was observed in three out of ten simulations [34]. The receptor was found to undergo various large-scale conformational changes, with four primary low-energy conformational states observed, including “Open,” “Partially Open,” “Closed,” and “Fully Closed” conformations (Figure 2.6a,b). Among the four low-energy conformations, three of them were already observed in the experimental structures, including “Open,” “Partially Open,” and “Closed” conformations, whereas the “Fully Closed” conformation

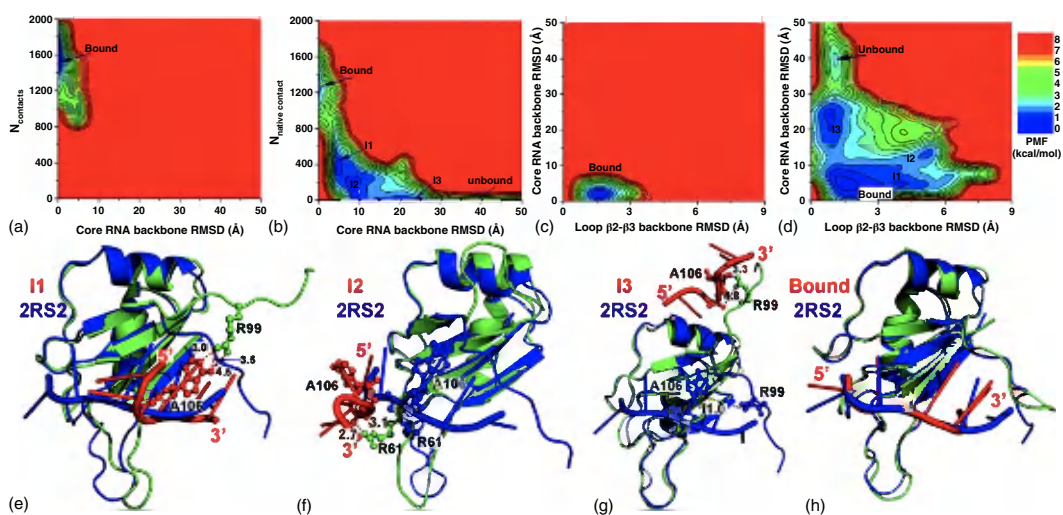


Figure 2.5 Binding of Musashi 1 (MSI1) RNA-binding protein to Numb mRNA. (a, b) 2D free energy profiles of the core RNA backbone RMSD relative to the first NMR conformation (PDB: 2RS2) and number of native contacts between MSI1 and Numb mRNA calculated from GaMD simulations starting from the (a) Bound and (b) Unbound states of the MSI1-Numb system. (c, d) 2D free energy profiles of the MSI1 β 2- β 3 loop backbone RMSD and core RNA backbone RMSD relative to the first NMR conformation (PDB: 2RS2) calculated from GaMD simulations starting from the (c) Bound and (d) Unbound states of the MSI1-Numb system. (E, H) Low-energy conformational states (I1, I2, and I3) and “Bound” state as identified from the 2D free energy profiles of the MSI1-RNA simulation system started from the Unbound state. The MSI1 protein and Numb RNA are shown in green and red, respectively. The NMR structure of the MSI1-Numb complex is shown in blue for comparison. Source: Reproduced with permission of Wang et al. [85]/Elsevier.

was uncovered from the LiGaMD simulations (Figure 2.6a,b) [34]. The presence of the polar and charged groups in a different part of the receptor was found to make favorable interactions with polar chloride and nitrogen atoms and the charged carboxylate group of the ligand molecules. Similarly, subdomain II of protein was relatively stable during LiGaMD simulations, adopting $\sim 2\text{--}4$ Å RMSD in relative to 1R4L PDB structure, whereas subdomain I showed higher flexibility with conformational changes ranging $\sim 3\text{--}10$ Å. Notably, two primary binding and dissociation pathways were observed in these production simulations (Figure 2.6c,d). The binding pathway involved the opening between $\alpha 2$ and $\alpha 4$ helices during the transition of subdomain I from the “Closed” to “Open” conformation [34] (Figure 2.6c), while the dissociation required interactions between MLN-4760 molecules and an interface formed between $\alpha 5$ helices and ACE2 3_{10}H_4 [34] (Figure 2.6d).

2.3.4 Discovery of Novel Small-Molecule Calcium Sensitizers for Cardiac Troponin C

Cardiac troponin C (TnC) is a calcium-dependent protein in the troponin complex responsible for the activation of muscle contraction. Disorders of TnC may trigger heart diseases and then cause death. One of the current therapies [90] requires the design of small molecules that can stabilize the open structure of the TnC and facilitate binding of the TnC switch peptide. To identify the potent small molecules, Coldren *et al.* [91] combined GaMD with high-throughput virtual screening to predict binding conformations and affinities of small molecules in TnC. GaMD simulations were compared with experiments for the TnC protein structures in complex with calcium sensitivity modulators. Independent 300ns GaMD simulations were performed on each system to obtain protein-ligand binding poses. GaMD trajectories were clustered to obtain the top 10 most populated conformations using the agglomerative hierarchical clustering algorithm, which were then used for virtual screening and docking. Their work identified a number of novel compounds that reduced the calcium dissociation rate and showed an overall calcium sensitization effect. One of the compounds exhibited high binding affinity in TnC and was further verified by the stopped-flow kinetic experiment.

2.3.5 Binding Kinetics Prediction from GaMD Simulations

Binding kinetics have recently been recognized to be more relevant for drug design. In particular, the dissociation rate constant that determines the drug residence time appears to better correlate with drug efficacy [92, 93]. However, binding kinetic rates are even more challenging to predict than binding free energies, largely due to slow dissociation processes [93]. In order to efficiently capture both the binding and unbinding processes of ligand/peptide/protein, we have recently developed selective GaMD methods, including LiGaMD, Pep-GaMD, and PPI-GaMD. The LiGaMD, Pep-GaMD, and Pep-GaMD simulations were able to capture multiple events of ligand/peptide/protein binding and unbinding within microsecond of simulation time. These highly efficient simulations thus allowed us to accurately

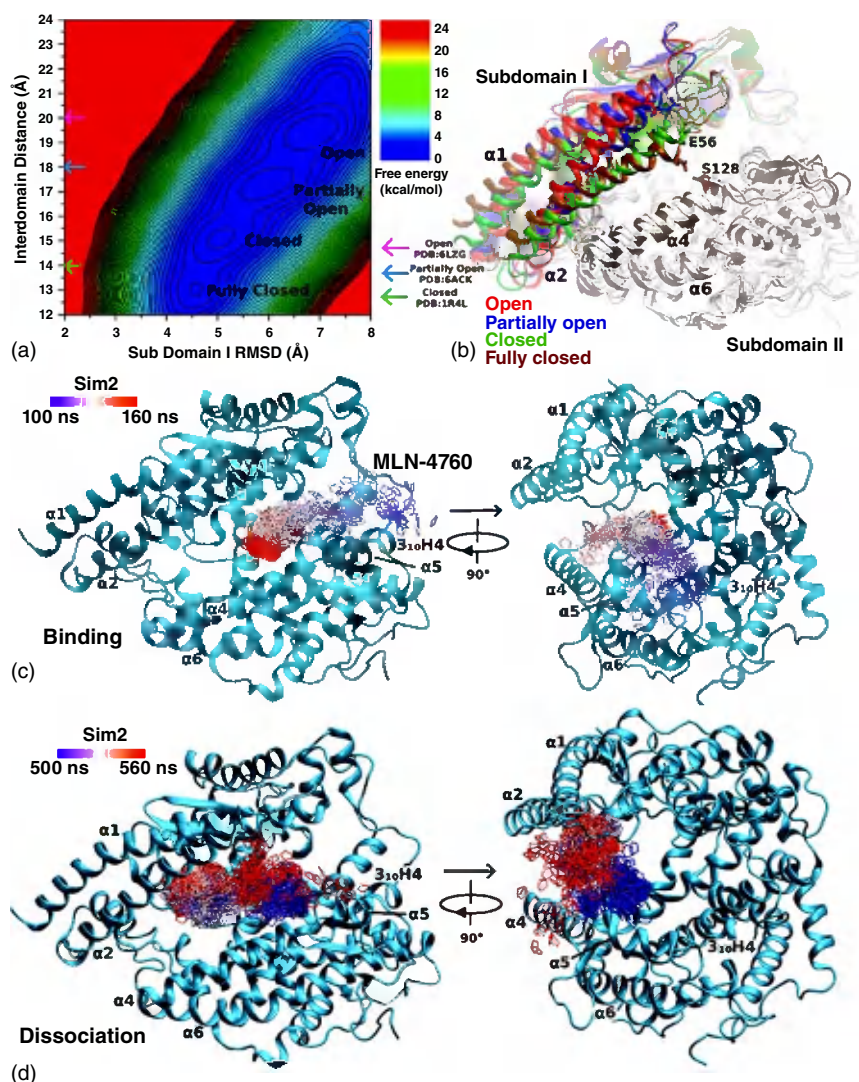


Figure 2.6 Binding and dissociation of the MLN-4760 inhibitor in the human ACE2 receptor. (a) 2D potential of mean force (PMF) of the subdomain I RMSD and interdomain distance calculated by combining the ten LiGaMD simulations. (b) Low-energy conformations of the ACE2 receptor with subdomain I found in the "Open" (red), "Partially Open" (blue), "Closed" (green), and "Fully Closed" (brown) states in the LiGaMD simulations. Subdomain II is stable and colored in white. (c) Two different views of the ligand binding pathways were observed in "Sim2," for which the center ring of MLN-4760 is represented by lines and colored by simulation time in a blue-white-red (BWR) color scale. (d) Two different views of the ligand dissociation pathway were observed in "Sim2," for which the center ring of MLN-4760 is represented by lines and colored by simulation time in a blue-white-red (BWR) color scale. Source: Reproduced with permission of Bhattarai et al. [34]/American Chemical Society.

characterize the ligand/peptide/protein binding thermodynamics and kinetics [55, 56].

LiGaMD has been proposed to quantitatively characterize ligand binding thermodynamics and kinetics [55]. Hundreds-of-nanosecond LiGaMD simulations captured repetitive guest binding and unbinding in the β -cyclodextrin host. The calculated guest binding free energies were in good agreement with experimental data with errors <1.0 kcal/mol in comparison with converged μ s-timescale cMD simulations. Particularly, ligand kinetic rate constants estimated using Kramers' rate theory were accurately predicted. Furthermore, repetitive binding and unbinding of the benzamidine inhibitor in trypsin was observed in 1 μ s LiGaMD simulations, allowing us to accurately calculate ligand binding free energy and kinetic rate constants. The predicted values were in excellent agreement with the experimental data [55].

Pep-GaMD [56] has been demonstrated on binding of three model peptides to the SH3 domains [94, 95], which include "PPPVP_{PRR}" (PDB: 1CKB), "PPPALPP_{KK}" (PDB: 1CKA) and "PAMPAR" (PDB: 1SSH). Repetitive peptide binding and unbinding processes were captured in independent 1 μ s Pep-GaMD simulations, allowing us to calculate peptide binding thermodynamics and kinetics. The predicted values from Pep-GaMD were in good agreement with available experimental data. PPI-GaMD [56] has been demonstrated on a model system of the ribonuclease barnase binding to barstar. Six independent 2 μ s PPI-GaMD simulations have successfully captured repetitive barstar dissociation and rebinding events. The barstar binding free energy predicted from PPI-GaMD was highly consistent with experimental data (-17.79 vs -18.90 kcal/mol) [96]. In addition, the predicted protein binding kinetic rates k_{on} and k_{off} were $21.7 \pm 13.8 \times 10^8 \text{ M}^{-1} \text{ s}^{-1}$ and $7.32 \pm 4.95 \times 10^{-6} \text{ s}^{-1}$, being highly consistent with the corresponding experimental values of $6.0 \times 10^8 \text{ M}^{-1} \text{ s}^{-1}$ and $8.0 \times 10^{-6} \text{ s}^{-1}$, respectively. Furthermore, PPI-GaMD simulations have provided mechanistic insights into barstar binding to barnase, which involve long-range electrostatic interactions and multiple binding pathways, being consistent with previous experimental and computational findings of this model system.

2.4 Conclusions

In this work, we have reviewed the important developments and applications of GaMD in the field of drug discovery. GaMD is an unconstrained enhanced sampling technique that allows for the exploration of large biomolecular conformational spaces and complex biological interactions. Furthermore, the boost potential in GaMD exhibits a Gaussian distribution, enabling accurate reweighting of the simulations using cumulant expansion to the second order. Given its strengths, GaMD was applied to reveal the binding mechanisms of various ligands to GPCRs, nucleic acids, and human ACE2 receptors, as well as the effects of allosteric drug leads in GPCRs at the residue level. Additional applications of GaMD uncovered the mechanisms of protein-membrane [41] interactions, identified cryptic pockets

within the SARS-CoV-2 main protease [97], explored drug binding to protease [98], and revealed the conformational landscape of drug binding to GPCRs [99]. Nevertheless, more efficient GaMD algorithms and enhanced sampling methods are still needed to characterize the thermodynamics and kinetics of important protein–protein/nucleic acid interactions and explore the structural dynamics in systems of increasing sizes, such as viruses and cells [3]. GaMD can be potentially applied to predict ADMET properties (e.g. membrane permeation), especially when combined with compatible enhanced sampling techniques such as replica exchange US [48]. Further developments in both supercomputing hardware and enhanced sampling methods should help tackle these challenges in the future.

References

- 1 Karplus, M. and McCammon, J.A. (2002). *Nat. Struct. Biol.* 9: 646–652, <https://doi.org/10.1038/Nsb0902-646>.
- 2 Hollingsworth, S. and Dror, R. (2018). *Neuron* 99: 1129–1143.
- 3 Wang, J. et al. (2021). *WIREs Comput. Mol. Sci.* e1521, <https://doi.org/10.1002/wcms.1521>.
- 4 Henzler-Wildman, K. and Kern, D. (2007). *Nature* 450: 964–972.
- 5 Harvey, M.J., Giupponi, G., and Fabritiis, G.D. (2009). *J. Chem. Theory Comput.* 5: 1632–1639.
- 6 Johnston, J.M. and Filizola, M. (2011). *Curr. Opin. Struct. Biol.* 21: 552–558.
- 7 Shaw, D.E. et al. (2010). *Science* 330: 341–346.
- 8 Lane, T.J., Shukla, D., Beauchamp, K.A., and Pande, V.S. (2013). *Curr. Opin. Struct. Biol.* 23: 58–65.
- 9 Vilardaga, J.-P., Bünemann, M., Krasel, C. et al. (2008). *Nat. Biotechnol.* 21: 807–812.
- 10 Miao, Y. and Ortoleva, P.J. (2006). *J. Chem. Phys.* 125: 214901.
- 11 Spiwok, V., Sucur, Z., and Hosek, P. (2015). *Biotechnol. Adv.* 33: 1130–1140.
- 12 Gao, Y.Q., Yang, L.J., Fan, Y.B., and Shao, Q. (2008). *Int. Rev. Phys. Chem.* 27: 201–227.
- 13 Liwo, A., Czaplewski, C., Oldziej, S., and Scheraga, H.A. (2008). *Curr. Opin. Struct. Biol.* 18: 134–139.
- 14 Christen, M. and van Gunstere, W. (2008). *J. Comput. Chem.* 29: 157–166.
- 15 Miao, Y. and McCammon, J.A. (2016). *Mol. Simul.* 42: 1046–1055.
- 16 Torrie, G. and Valleau, J. (1977). *J. Comput. Phys.* 23: 187–199.
- 17 Kumar, S., Rosenberg, J., Bouzida, D. et al. (1992). *J. Comput. Chem.* 13: 1011–1021.
- 18 Laio, A. and Gervasio, F. (2008). *Rep. Prog. Phys.* 71: 126601.
- 19 Besker, N. and Gervasio, F. (2012). *Computational Drug Discovery and Design*, 501–513. Berlin: Springer.
- 20 Darve, E., Rodriguez-Gomez, D., and Pohorille, A. (2008). *J. Chem. Phys.* 128: 144120.
- 21 Darve, E., Wilson, M., and Pohorille, A. (2002). *Mol. Simul.* 28: 113–144.

- 22 Isralewitz, B., Baudry, J., Gullingsrud, J. et al. (2001). *J. Mol. Graph. Model.* 19: 13–25.
- 23 Sugita, Y. and Okamoto, Y. (1999). *Chem. Phys. Lett.* 314: 141–151.
- 24 Okamoto, Y. (2004). *J. Mol. Graph. Model.* 22: 425–439.
- 25 Hansmann, U. (1997). *Chem. Phys. Lett.* 281: 140–150.
- 26 Wu, X. and Brooks, B. (2003). *Chem. Phys. Lett.* 381: 512–518.
- 27 Wu, X., Brooks, B., and Vanden-Eijnden, E. (2016). *J. Comput. Chem.* 37: 595–601.
- 28 Wu, X. and Wang, S. (1998). *J. Phys. Chem. B.* 102: 7238–7250.
- 29 Hamelberg, D., Mongan, J., and McCammon, J.A. (2004). *J. Chem. Phys.* 120: 11919–11929.
- 30 Voter, A. and Hyperdynamics, F. (1997). *Phys. Rev. Lett.* 78: 3908.
- 31 Miao, Y., Feher, V.A., and McCammon, J.A. (2015). *J. Chem. Theory Comput.* 11: 3584–3595.
- 32 Miao, Y. et al. (2014). *J. Chem. Theory Comput.* 10: 2677–2689.
- 33 Do, H., Akhter, S., and Miao, Y. (2021). *Front. Mol. Biosci.* 8: 242.
- 34 Bhattarai, A., Pawnikar, S., and Miao, Y. (2021). *J. Phys. Chem. Lett.* 12: 4814–4822.
- 35 Pang, Y., Miao, Y., and McCammon, J.A. (2017). *J. Chem. Theory Comput.* 13: 9–19.
- 36 Tang, Z. et al. (2021). *Nucleic Acids Res.* 49: 7870–7883.
- 37 Do, H., Wang, J., Bhattarai, A., and Miao, Y. (2022). *J. Chem. Theory Comput.* 18: 1423–1436.
- 38 Bhattarai, A., Devkota, S., Bhattarai, S. et al. (2020). *ACS Central Sci.* 6: 969–983.
- 39 Bhattarai, A. et al. (2022). *J. Am. Chem. Soc.* 144: 6215–6226.
- 40 Miao, Y. and McCammon, J.A. (2016). *Proc. Natl. Acad. Sci. U. S. A.* 113: 12162–12167.
- 41 Bhattarai, A., Wang, J., and Miao, Y. (2020). *J. Comput. Chem.* 41: 460–471.
- 42 Miao, Y. and McCammon, J.A. (2018). *Proc. Natl. Acad. Sci. U. S. A.* 115: 3036–3041.
- 43 Wang, J. and Miao, Y. (2019). *J. Phys. Chem. B.* 123: 6462–6473.
- 44 Wang, J. and Miao, Y. (2022). *J. Chem. Theory Comput.* 18: 1275–1285.
- 45 East, K.W. et al. (2020). *J. Am. Chem. Soc.* 142: 1348–1358.
- 46 Ricci, C.G. et al. (2019). *ACS Central Sci.* 5: 651–662.
- 47 Huang, Y.-M., McCammon, J.A., and Miao, Y. (2018). *J. Chem. Theory Comput.* 14: 1853–1864.
- 48 Oshima, H., Re, S., and Sugita, Y. (2019). *J. Chem. Theory Comput.* 15: 5199–5208.
- 49 Ahn, S.H., Ojha, A.A., Amaro, R.E., and McCammon, J.A. (2021). *J. Chem. Theory Comput.* 17: 7938–7951.
- 50 Ponder, J.W. et al. (2010). *J. Phys. Chem. B.* 8: 2549–2564.
- 51 Shi, Y. et al. (2013). *J. Chem. Theory Comput.* 9: 4046–4063.
- 52 Zhang, C. et al. (2018). *J. Chem. Theory Comput.* 14: 2084–2108.
- 53 Lagardere, L. et al. (2018). *Chem. Sci.* 9: 956–972.
- 54 Adjoua, O. et al. (2021). *J. Chem. Theory Comput.* 17: 2034–2053.

- 55 Miao, Y., Bhattarai, A., and Wang, J. (2020). *J. Chem. Theory Comput.* 16: 5526–5547.
- 56 Wang, J. and Miao, Y. (2020). *J. Chem. Phys.* 153: 154109.
- 57 Miao, Y. and McCammon, J.A. (2017). *Annu. Rep. Comp. Chem.* 13: 231–278, <https://doi.org/10.1016/bs.arcc.2017.06.005>.
- 58 Keras-Vis (GitHub, 2017).
- 59 Miao, Y. (2018). *J. Chem. Phys.* 149: 072308, <https://doi.org/10.1063/1.5024217>.
- 60 Doshi, U. and Hamelberg, D. (2011). *J. Chem. Theory Comput.* 7: 575–581, <https://doi.org/10.1021/ct1005399>.
- 61 Frank, A.T. and Andricioaei, I. (2016). *J. Phys. Chem. B* 120: 8600–8605, <https://doi.org/10.1021/acs.jpcc.6b02654>.
- 62 Hamelberg, D., Shen, T., and Andrew McCammon, J. (2005). *J. Chem. Phys.* 122: 241103, <https://doi.org/10.1063/1.1942487>.
- 63 Celerse, F. et al. (2022). *J. Chem. Theory Comput.* 18: 968–977.
- 64 Copeland, M.C. et al. (2022). *J. Phys. Chem. B.* 126: 5810–5820.
- 65 Hauser, A.S. et al. (2018). *Cell* 172: 41–54.
- 66 Stevens, R.C. et al. (2013). *Nat. Rev. Drug Discov.* 12: 25–34.
- 67 Isberg, V. et al. (2016). *Nucleic Acids Res.* 44: D365–D364.
- 68 Fredholm, B. et al. (1997). *Trends Pharmacol. Sci.* 18: 79–82.
- 69 Jacobson, K.A. and Gao, Z.-G. (2006). *Nat. Rev. Drug Discov.* 5: 247–264.
- 70 Cheng, R. et al. (2017). *Structure* 25: 1275–1285.
- 71 Draper-Joyce, C.J. et al. (2021). *Nature* 597: 571–576, <https://doi.org/10.1038/s41586-021-03897-2>.
- 72 Dror, R.O. et al. (2011). *Proc. Natl. Acad. Sci. U. S. A.* 108: 18684–18689.
- 73 Avlani, V. et al. (2007). *J. Biol. Chem.* 282: 25677–25686.
- 74 Peeters, M. et al. (2012). *Biochem. Pharmacol.* 84: 76–87.
- 75 Nguyen, A. et al. (2016). *Mol. Pharmacol.* 90: 715–725.
- 76 Miao, Y., Bhattarai, A., Nguyen, A. et al. (2018). *Sci. Rep.* 8: 16836.
- 77 Wang, J. et al. (2020). *GPCRs* (ed. B. Jastrzebska and P.S.H. Park), 283–293. Academic Press.
- 78 Morris, G.M. et al. (2009). *J. Comput. Chem.* 30: 2785–2791.
- 79 Bhattarai, A., Wang, J., and Miao, Y. (2020). *Biochim. Biophys. Acta Gen. Subj.* 1864: 129615.
- 80 Amber 2021 (University of California, San Francisco, 2021).
- 81 Ivani, I. et al. (2016). *Nat. Methods* 13: 55–58.
- 82 Zgarbova, M., Otyepka, M., Šponer, J. et al. (2011). *J. Chem. Theory Comput.* 7: 2866–2902.
- 83 Wang, J., Wolf, R., Caldwell, J. et al. (2004). *J. Comput. Chem.* 25: 1157–1174.
- 84 Mark, P. and Nilsson, L. (2001). *Chem. A Eur. J.* 105: 9954–9960.
- 85 Wang, J., Lan, L., Wu, X., Xu, L. & Miao, Y. bioRxiv, 2020.2010.2030.362756, <https://doi.org/10.1101/2020.10.30.362756> (2021).
- 86 Kudinov, A.E., Karanickolas, J., Golemis, E.A., and Boumber, Y. (2017). *Clin. Cancer Res.* 23: 2143–2153, <https://doi.org/10.1158/1078-0432.Ccr-16-2728>.
- 87 Gross, L.Z.F. et al. (2020). *ChemMedChem* 15: 1682–1690, <https://doi.org/10.1002/cmdc.202000368>.

- 88 Hoffmann, M. et al. (2020). *Cell* 181: 271.
- 89 Towler, P. et al. (2004). *J. Biol. Chem.* 279: 17996–18007.
- 90 Remme, W.J. and Swedberg, K. (2001). *Eur. Heart J.* 22: 1527–1560, <https://doi.org/10.1053/euhj.2001.2783>.
- 91 Coldren, W.H., Tikunova, S.B., Davis, J.P., and Lindert, S. (2020). *J. Chem. Inf. Model.* 60: 3648–3661, <https://doi.org/10.1021/acs.jcim.0c00452>.
- 92 Schuetz, D.A. et al. (2017). *Drug Discov. Today* 22: 896–911, <https://doi.org/10.1016/j.drudis.2017.02.002>.
- 93 Tonge, P. and Drug-Target, J. (2018). *ACS Chem. Neurosci.* 9: 29–39, <https://doi.org/10.1021/acscchemneuro.7b00185>.
- 94 Ahmad, M. and Helms, V. (2009). *Chem. Cent. J.* 3: O22.
- 95 Ball, L.J., Kuhne, R., Schneider-Mergener, J., and Oschkinat, H. (2005). *Angew. Chem. Int. Ed. Engl.* 44: 2852–2869, <https://doi.org/10.1002/anie.200400618>.
- 96 Schreiber, G. and Fersht, A.R. (1993). *Biochemistry* 32: 5145–5150.
- 97 Sztain, T., Amaro, R.E., and McCammon, J.A. (2021). *J. Chem. Inf. Model.* 61: 3495–3501.
- 98 Wang, Y.-T. et al. (2022). *Phys. Chem. Chem. Phys.* 24: 22898–22904.
- 99 Zhang, H. et al. (2021). *Nat. Commun.* 12: 4151.

3

MD Simulations for Drug-Target (Un)binding Kinetics

Steffen Wolf

Biomolecular Dynamics, Institute of Physics, University of Freiburg, Hermann-Herder-Strasse 3, Freiburg 79104, Germany

3.1 Introduction

3.1.1 Preface

Compared to the free energy calculations employed for drug design presented in the preceding chapters, determining protein-drug binding and unbinding kinetics from molecular dynamics (MD) simulations is a comparatively young field. Though a large body of methods has evolved within the last decade, these developments have yet to result in commonly accepted best practices as well as a gold standard of test systems at the time of writing this chapter.

Because of the still-fast-paced rate of development of the field, the author here can only attempt to give a general overview of the current state-of-the-art. Additionally, the versatility of approaches employed prohibits an exhaustive explanation of the underlying theory. Instead, this chapter aims at providing the interested computer-aided drug design (CADD) researcher with a basic overview of the problem of predicting kinetics, its formal basis, and a practical guideline on the strengths and shortcomings of selected methods. For the interested reader, links to the core literature and helpful reviews will be provided.

3.1.2 Motivation for Predicting (Un)binding Kinetics

Experimental approaches in drug design focus on determining a diverse set of constants, e.g. inhibition constants K_i , half maximal effective concentrations EC_{50} , or dissociation constants K_D . Especially the latter K_D can be reformulated in the form of a standard binding free energy

$$\Delta F^0 = -k_B T \ln K_D. \quad (3.1)$$

with an implicit reference concentration of 1 M. Accordingly, finding a compound with large ΔF^0 directly translates into finding a high-affinity compound, which is the basic principle behind free energy-based computational as well as early-stage experimental drug development.

Computational Drug Discovery: Methods and Applications, First Edition.

Edited by Vasanthanathan Poongavanam and Vijayan Ramaswamy.

© 2024 WILEY-VCH GmbH. Published 2024 by WILEY-VCH GmbH.

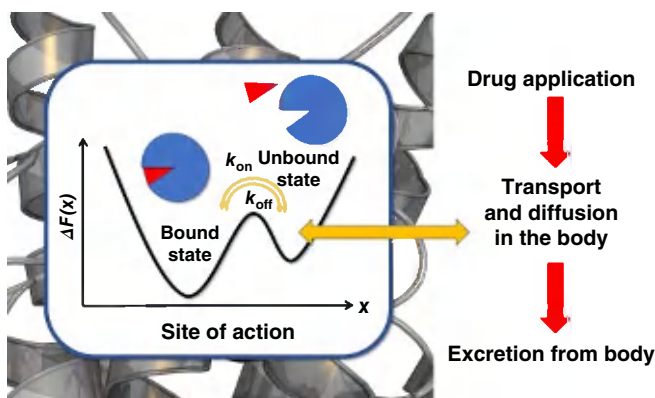


Figure 3.1 Schematic of nonequilibrium aspects of drug binding and pharmacokinetics. Red arrows represent irreversible steps and orange arrows reversible steps. The coordinate x represents the distance between the ligand and its binding site in a target protein.

Starting in the midst of the 2000s, this focus on high-affinity compounds was challenged by several works proposing that the efficacy of a compound, which is the finally relevant criterion for the successful applicability of a compound as a drug, may not result from high affinity but from long residence times of a drug at its target protein [1–4]. A prime example of this hypothesis is the tyrosine kinase inhibitor Gleevec (Imatinib) [2, 5, 6], which gains its selectivity for Abl over other kinases by such slow unbinding kinetics.

Where does this difference between affinity and residence time come from? In short, from the difference between equilibrium and nonequilibrium statistical mechanics. Free energies are thermodynamic variables describing closed systems in equilibrium in which no exchange of particles with the surrounding occurs. This is markedly different from the situation in the human body, which is a strictly nonequilibrium environment with particle uptake and release. Figure 3.1 depicts a simplified yet instructive representation of the process: if a dose of a drug is administered, one initially observes transport and diffusion to its active sites, followed by a local equilibration between the drug-bound and the unbound states. Concurrently, the liver and kidneys start to remove the drug from the body and thus deplete the population of unbound drugs. In case that the bound and unbound states exchange quickly, a drug that binds tightly to its target will nevertheless be quickly eliminated and thus be only shortly active. Consequently, the compound needs to be applied more often, raising the possibility for off-target binding and thus undesired side effects.

3.1.3 The Time Scale Problem of MD Simulations

Given the large interest in predicting residence times, it is natural to consider performing MD simulations for their calculation. Here, however, one faces a major challenge based on the fundamental physics underlying MD simulations: this

method is based on numerical solutions of Newton's equations of motion, in which in a new position of an atom $q(t)$ can be calculated in its simplest form as

$$q(t + \Delta t) = q(t) + v(t)\Delta t + \frac{f(t)}{2m}\Delta t^2 \quad (3.2)$$

based on the atom's mass m , velocity $v(t)$, and the force $f(t)$ the atom experiences from its environment. Such solutions require the definition of a time increment Δt , which needs to be small enough to sample the fastest motion in a simulation system. In biomolecular systems, these motions are O–H stretch vibrations with a period of about ten femtoseconds. Accordingly, Δt needs to be on the order of 1–2 fs. This time range stands in stark contrast to pharmacologically relevant timescales, as especially unbinding events occur on the order of seconds to hours (see, e.g. [7, 8]). As the solution of Eq. (3.2) needs to be performed iteratively, parallelization only remedies the computational costs of calculating forces per time step but cannot solve this discrepancy. At the time of writing of chapter, the best-performing purpose-built simulation computer, Anton3 of D. E. Shaw research [9], was said to be able to calculate 100 μ s per day. While this is an achievement of its own, even simulating a single second of real-time would still require 33 years. Therefore, there is no computational hardware in sight that may solve this conundrum. Instead, smart simulation approaches from the group of enhanced sampling techniques are required, which help to overcome this timescale problem. While this type of technique has been introduced in Chapter 2, we have to briefly review the theory underlying such calculations to better understand the way these approaches can help in calculating molecular process rates.

3.2 Theory of Molecular Kinetics Calculation

3.2.1 Nonequilibrium Statistical Mechanics in a Nutshell

To better understand the challenges encountered in the calculation of kinetics, we briefly need to review their formal basis and the differences to free energy calculations.¹ Let us assume that for investigating the (un)binding process of a ligand from its target protein, it is sufficient to regard the process only along the distance $x(t)$ of the center of mass (COM) between ligand and protein. This distance is formally a *collective variable* (CV), i.e. a linear combination of atomic mass-weighted atom positions $q(t)$. If a CV is capable of sufficiently describing a molecular process of interest, it qualifies as a *reaction coordinate*. Given an MD simulation system in an *NVT* ensemble,² one can calculate a free energy profile along x

$$\Delta F(x) = -k_B T \ln P(x) \quad (3.3)$$

with the Boltzmann constant k_B and the probability distribution $P(x)$. This distribution can simply be obtained, e.g. by histogramming $x(t)$ from an unbiased

1 For readers who want to dive deeper into the formal basis of nonequilibrium statistical mechanics, the author recommends the respective books by Zwanzig [10] and Pottier [11].

2 We use the ensemble here for formal reasons. The solutions in a *NPT* ensemble contain an additional volume work term, but are analogous in their form.

MD simulation along x . One can directly see that $P(x)$ does not contain the time dependence of $x(t)$ anymore. Instead, $\Delta F(x)$ constitutes the time-independent potential of mean force with

$$-\frac{dF(x)}{dx} = \langle f(x) \rangle. \quad (3.4)$$

where $\langle \cdot \rangle$ denotes a mean over time (specifically here overall points in t that correspond to a specific x). To recover the formal time dependence of $x(t)$, it is possible to rationalize dynamics along the CV of choice, e.g. as a Markovian Langevin equation [10]

$$m\ddot{x} = -\frac{dF(x)}{dx} - \dot{x}\Gamma + \sqrt{2k_B T \Gamma} \xi(t), \quad (3.5)$$

with a friction coefficient Γ . On the right side of Eq. (3.5), the first term is the mean force as given in Eq. (3.4) and the second term is a friction force. The third term, representing a fluctuating force, consists of a zero-mean, normally distributed stochastic term $\xi(t)$ with variance unity (i.e. a random number drawn from a standard Gaussian probability distribution) and an amplitude $\sqrt{2k_B T \Gamma}$ that is coupled to the friction due to the fluctuation-dissipation theorem [11]. Friction factors can be calculated, e.g. based on the force autocorrelation function (ACF) [12]

$$C(t') = \langle (f(t) - \langle f \rangle)(f(t+t') - \langle f \rangle) \rangle \quad (3.6)$$

as integral over time

$$\Gamma = \frac{1}{k_B T} \int_{t_0}^{\infty} C(t') dt' \quad (3.7)$$

If we assume that $C(t')$ follows an exponential decay $C(t') \approx \langle f_0^2 \rangle \exp(-t'/\tau)$, Eq. (3.7) becomes $\Gamma = \frac{\tau}{k_B T} \langle f_0^2 \rangle$. To be able to recover the dynamics along x , we need not only to know the time-independent mean force given by the free energy but also the variance of the force. On a microscopic level, the mean force represents the average interaction of a ligand with its surroundings, while the variance contains the effect of chaotic dynamics such as collisions of the ligand with water molecules or the motion of side chains. The latter two effects can both decelerate and accelerate the ligand, hence the appearance of friction factors in both friction and fluctuation forces.

3.2.2 Kramers Rate Theory

Based on Eq. (3.5), we require both $\Delta F(x)$ and Γ for being able to calculate rates. Indeed, Kramers derived an analytical expression for such rates based on these two values [13, 14] for the high friction regime: given a free energy surface as displayed in Figure 3.1, i.e. with a single free energy minimum and a single free energy barrier with a height $\Delta F^\ddagger$, a ligand will cross the barrier at a rate

$$k = \frac{\sqrt{|F''(x_{\min})| |F''(x_{\text{barrier}})|}}{2\pi\Gamma} \exp\left(-\frac{\Delta F^\ddagger}{k_B T}\right) \quad (3.8)$$

Here, the absolute value of the second derivative at the minimum $|F''(x_{min})|$ and the barrier $|F''(x_{barrier})|$, and the friction Γ is hard to calculate. As a consequence, the fraction in Eq. (3.8) usually cannot be solved analytically. However, the structure of the equation helps us to understand the principles of calculating molecular rates: the fraction in Eq. (3.8) has units of inverse time and represents the *attempt frequency*, i.e. how often a ligand attempts to leave the bound state. The exponential term represents the Boltzmann probability to reach the least likely position along x from the bound state, which is the *transition state* barrier with a height $\Delta F^\ddagger$. All biased techniques employed for rate prediction modulate these two parts of Eq. (3.8) such that a sufficient number of transitions occurs in a reasonable time frame of simulations (preferably within a few nanoseconds), and that the underlying unbiased rate can be extracted from these calculations with a suitable reweighing scheme.

3.2.3 Biased MD Methods

We here briefly review the theory of biased MD simulations, which are of most importance for the calculation of molecular kinetics. For further information, the author would like to refer the reader to Chapter 2 of this book as well as to a recent review [15].

3.2.3.1 Temperature- and Barrier-Scaling

As can be seen directly in Eq. (3.8), the simplest approach to accelerate sampling is to include a tuning factor a into the Boltzmann probability, leading to accelerated kinetics

$$k_a = k \exp \left(-\frac{\Delta F^\ddagger}{k_B T} [a - 1] \right) \quad (3.9)$$

There are two options available on how to implement factor a into simulations: on one hand, it can be applied as a scaling factor of temperature, leading to *temperature-accelerated MD* [16, 17]. While this approach is well applicable in simulations of protein folding [18], it is less suited for the problems discussed here, as protein structures are sensitive to high temperatures that would be necessary for sufficient acceleration. Therefore, given that $\Delta F^\ddagger = \Delta U^\ddagger - T\Delta S^\ddagger$ with potential (inner) energy U and entropy S , the factor is usually included in the free energy as a modification of the potential energies U , leading to the approach of *scaled* or *smoothed potential MD* [19]. An alternative development here is Gaussian accelerated MD [20], which uses a quadratic scaling of U below threshold energy that dampens the influence of deep minima in ΔF .

3.2.3.2 Bias Potential-Based Methods

Instead of accelerating dynamics per se, biased potential methods employ a potential $V_{bias}(x)$ that is added to the system Hamiltonian, leading to an adjusted probability $P_{bias}(x)$ in Eq. (3.3). The free energy of the unbiased system $\Delta F(x)$ can then be calculated from the biased simulation as [21]

$$\Delta F(x) = -k_B T \ln P_{bias}(x) - V_{bias}(x) \quad (3.10)$$

If $V_{bias}(x)$ is well-chosen, then $P_{bias}(x)$ is close to uniform enough to sample well the relevant range of x . Torrie and Valleau originally suggested the usage of harmonic potentials $V_{bias}(x) = \frac{1}{2}(x - x_0)^2$ [21]. Alternatively, in the approach of *metadynamics* [22, 23], $V_{bias}(x)$ is constructed from a sum of Gaussian functions

$$V_{bias}(x) = \sum_i A_i \exp(-[x - x_i]^2 / 2\sigma^2), \quad (3.11)$$

which are sequentially placed in i steps during simulation until $P_{bias}(x)$ is constant, i.e. a ligand diffuses freely in and out of its binding pocket. Under these conditions, $-k_B T \ln P_{bias}(x)$ becomes a constant that can be ignored, and $\Delta F(x) = -V_{bias}(x)$.

3.2.3.3 Bias Force-Based Methods

Instead of employing bias potentials to drive a ligand out from its binding site, it is feasible to pull it out actively, i.e. to apply an external force. This approach is reminiscent of atomic force microscopy, and indeed, the first such computational approaches were carried out on the streptavidin-biotin complex in combination with such experiments [24]. Employing bias forces

$$f_{bias}(x) = -\frac{d}{dx} V_{bias} \quad (3.12)$$

from a harmonic potential with a time-dependent center point

$$V_{bias}(x) = \frac{1}{2}(x - [x_0 + vt])^2 \quad (3.13)$$

gives rise to *steered MD* simulations [25, 26].

A theory-intrinsic alternative to these harmonic restraints is *targeted MD* [27], which employs a constant velocity constraint

$$\Phi(x) = x - [x_0 + vt] = 0, \quad (3.14)$$

which then gives rise to a constraint force

$$f_{bias}(x) = \lambda \frac{d}{dx} \Phi(x). \quad (3.15)$$

with a Lagrange multiplier λ . The constraint force is calculated in each time step to ensure that the ligand moves out of the binding site with a velocity v .

For both approaches listed above, $f_{bias}(x)$ is applied within MD simulations along a pre-defined vector between protein and ligand, e.g. the distance between the ligand's COM and the COM of some reference atoms within the protein. This setup requires a general idea in which direction an unbinding pathway can be found. If this knowledge is not given, it is possible to apply a small f_{bias} with a pre-set amplitude, but in a random direction, giving rise to *random acceleration MD* [28]. This approach allows for employing simpler reaction coordinates, e.g. root mean square distances of a ligand from its initial position, and therefore allows to explore possible unbinding pathways without prior knowledge about them.

3.2.3.4 Knowledge-Biased Methods

All three biasing approaches described above have the disadvantage that the introduced bias interferes with the dynamics of the system. Another option is to apply

a knowledge-based bias: similar to the action of a “Maxwell’s daemon,” trajectories can be sorted based on whether they naturally evolve along a reaction coordinate of choice, resulting in such methods as *Weighted Ensembles* (WE) [29] or *Milestoning* [30]. Giving an appropriate short simulation time, the final structures from trajectories that have successfully crossed a pre-defined threshold along the reaction coordinate are used as seeds for new simulations. The cycle is repeated until the final state sought after has been reached. To find the correct $\Delta F^\ddagger$, *Transition Path Sampling* (TPS) [31, 32] can be carried out. Here, after an initial (enforced) transition from bound to unbound state, a redistribution of velocities is applied to a structure at or around the transition state, and the resulting state is numerically propagated forward and backward in time to check if the unbound and bound states are reached, respectively. If a successful transition is found, the procedure is repeated to generate an ensemble of transition trajectories.

3.2.3.5 Coarse-graining and Master Equation Approaches

All methods listed before focus on accelerating dynamics in biasing fully-atomistic MD simulations. However, alternative approaches exist that coarse-grain the overall dynamics, which reduces the necessary computational power by several orders of magnitude. One such approach is the integration of the Langevin Equation (3.5) [33]. Alternatively, it is possible to reduce the dynamics to the waiting times in the bound and unbound states and any intermediate states in between. If the transitions between all these states are Markovian, i.e. the transition between two states at time t does not depend on any information from a prior time, a *Markov State Model* (MSM) [34, 35] can be constructed. From the MSM, binding and unbinding times can then be determined. The advantage of this approach is that short trajectories, e.g. from knowledge-based simulation methods, can be used for the construction of an MSM rate matrix that encodes the transition times between the states.

3.3 Challenges and Caveats in Rate Prediction

3.3.1 Finding Reaction Coordinates and Pathways

At this position in the chapter, it is necessary to highlight the challenges a researcher faces when attempting to calculate (un)binding rates. The major and crucial difference to standard free energy calculations is the coordinate dependence of $\Delta F(x)$. While free energy calculations in principle are possible based on the single-step difference between bound and unbound states using Zwanzig’s equation [36], determining $\Delta F(x)$ requires finding both the correct reaction coordinate x and the right transition state $\Delta F^\ddagger$ along it. In time-dependent processes, x describes the slowest process in the molecular system of interest. If this process is, say, a conformational change of the protein, such as in the Abl kinase mentioned above, then a simulation of unforced ligand unbinding ignoring this change will inevitably yield wrong results. Furthermore, a good reaction coordinate such as dihedral angles describing this conformational change may not be a suitable biasing coordinate,

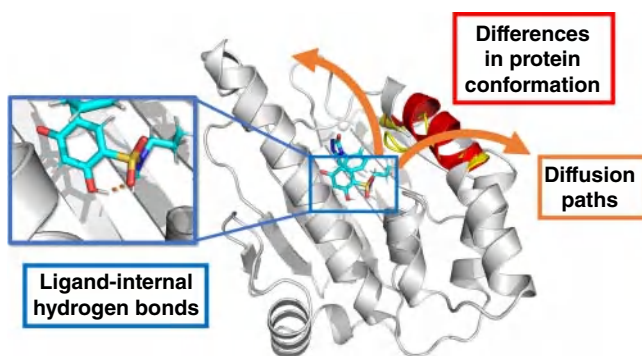


Figure 3.2 Possible reaction coordinates that need to be taken into account in the search for pathways for the example of the N-terminal domain of Hsp90. Protein as cartoon, ligand as sticks.

as a biasing coordinate requires the possibility to define forces that can serve as input for the integrator Eq. (3.2). Lastly, it may be that x is not one-dimensional but a multi-dimensional $\mathbf{x}$, e.g. if conformational change and ligand diffusion to and from the binding site are coupled with each other, if a ligand can take different routes out of a binding site (see Figure 3.2), or if ligand translation is coupled to a specific bond rotation. In such cases, ligands may take paths through $\mathbf{x}$ that need to be found to understand the unbinding mechanism and to determine the most relevant unbinding path. Accordingly, methods have been developed for finding or learning unbinding reaction coordinates on the fly [32, 37, 38], detecting pathways in trajectory ensembles a posteriori [33, 39–43], or performing outright brute-force pathway exploration [44–46]. The reader needs to note that pathways can change drastically between ligands despite only small chemical differences. For example, we found for two ligands bound to the N-terminal domain of Hsp90 that swapping an amide with a sulfonamide moiety results in a completely different unbinding behavior [43]. The reason here is the formation of a ligand-internal hydrogen bond, which is not formed with a carbonyl oxygen but exists with a sulfonyl oxygen atom. The presence of the hydrogen bond led to an additional rotation barrier within the ligand, causing the sulfonamide ligand to be less flexible than its amide counterpart. The corresponding hydrogen bond donor/acceptor distance indeed turned out to be a good additional reaction coordinate to discriminate unbinding pathways.

3.3.2 Error Ranges of Estimates

Another general issue is the error range in absolute (un)binding rate predictions. We see in Eq. (3.8) that this error is exponentially dependent on the accuracy of transition barrier height estimation. If we consider the mean error range currently achievable by free energy calculations [47] $\Delta\Delta F^\ddagger \approx 1.7 \text{ kcal/mol} \approx 3 k_B T$, then the corresponding mean error of a rate is $\Delta k \approx k \cdot \exp(3)$, so about a factor of 17. Predicting a binding rate correctly within a magnitude is already considered a successful prediction and within the best range achievable.

3.3.3 A Need for Reliable Benchmarking Systems

Lastly, a significant challenge is the current absence of a gold standard system to benchmark theoretical prediction methods. Commonly employed systems that allow a computational comparison between rates predicted by biased methods and the true unbiased rates are a NaCl ion pair in water [33], the F506 binding protein (FBKP) with bound DMSO, DSS of 4-hydroxy-2-butanone as ligands [48], or the L99A mutant of T4 lysozyme with bound benzene or xylene [49]. It is easy to note that these compounds lack the chemical complexity that constitutes a real drug. On the side of experimentally measured kinetics, the trypsin-benzamidine complex [50] has emerged as a generally investigated benchmarking system. The problem here is that the kinetics of only a single ligand is known and used for predictions, which does not allow filtering of true from false positives. A system that can serve as a “gold standard” for benchmarking drug kinetics has not been agreed on yet. The author would like to highlight the N-terminal domain of Hsp90 [8] as well as the A2 adenosine receptor [7] as a possible benchmarking system, for which diverse sets of protein–ligand complexes have been characterized both in their atomic structure and their binding kinetics within the framework of the kinetics for drug discovery consortium [51]. Especially for Hsp90, a library of chemically diverse ligands is available that cover a range in rates of about four orders of magnitude in time [8, 52, 53]. Furthermore, long ligand residence is believed to depend on a conformational change of a protein helix covering the binding site [8], which poses a challenge for reaction coordinate prediction methods.

3.3.4 Problems with Force Fields

Besides all these issues, a further problem might be found in the force fields used in MD simulations themselves: transition state barrier heights $\Delta F^\ddagger$ may be overestimated because of missing polarization effects. While this problem is well known from simulations of ion transfer rates over membrane channels [54–56], it has recently been proposed to exist in the case of the M2 muscarinic receptor, as well [57]. A possible remedy could be the utilization of polarizable force fields [58, 59] introduced in Chapter 6, for which reliable ligand force fields still have to be developed.

3.4 Methods for Rate Prediction

In the following, several high-accuracy (within the expected error range, see above) applications of theories from Section 3.2 will be briefly introduced, and their advantages and limitations will be discussed. Comprehensive and recent reviews of existing methods and their detailed performance can be found in Refs. [60–67].

3.4.1 Unbinding Rate Prediction

3.4.1.1 Empirical Predictions

The first class to consider are methods that fulfill a similar role like Docking in the prediction of high-affinity ligands. These approaches do not attempt to predict

absolute rate constants, but relative parameters with a proposed proportionality to rate constants such as mean biased unbinding times or the mean nonequilibrium work to enforce unbinding, which are then used to score ligands. Such simulations are comparatively inexpensive and require a few to several hundred nanoseconds of simulated time per ligand. The accuracy of predictions suffers from similar issues as Docking [68], e.g. a dependence on prediction accuracy on protein target and ligand chemical scaffold, but at least allows a qualitative to semi-quantitative classification of ligands that is fast.

Scaled MD and smoothed potential MD have been successfully employed for scoring ligands bound to the N-terminal domain of Hsp90 [40, 42] as well as glucokinases [69]. The advantage of these approaches is that no reaction coordinate needs to be employed to accelerate such simulations. The first limitation of scaled and smoothed potential MD is the comparatively high simulation time for slowly unbinding compounds of several microseconds. Furthermore, scaling potential energy barriers following Eq. (3.9) causes a destabilization of the protein's structure, requiring the application of soft restraints onto protein atoms, which causes problems when a protein conformational change is involved in the unbinding process.

From the side of biased simulations, fast adiabatic metadynamics [70] and steered MD [71] have been employed for several G protein-coupled receptor-ligand combinations. The currently fastest scoring method [63] is nonequilibrium targeted MD [72], which only requires a few nanoseconds of cumulative simulation time to score a ligand and is independent of the unbinding rate of a ligand. The major limitation of this approach is that only compounds with similar chemical scaffolds can be compared to each other. Furthermore, all three approaches require a relatively good intuition along which pathway unbinding occurs for a definition of the biasing coordinate.

The best compromise between computational cost, prediction accuracy, and versatility for empirical predictions is found in tauRAMD [53], based on random acceleration MD. This method has been successfully applied to a range of different protein-ligand complex classes, including G protein-coupled receptors (GPCRs) and kinases [49, 73, 74]. Computationally, investigating a single slowly unbinding compound may still require hundreds of nanoseconds of simulation time. However, it is particularly well suited in cases where unbinding pathways are not known and, therefore, can serve for pathway exploration as well.

3.4.1.2 Prediction of Absolute Unbinding Rates

In comparison to scoring ligands according to unbinding rates, the prediction of absolute unbinding rates is still a daunting task and requires significant computational effort. As transitions along x need to be sufficiently sampled, the calculation of explicit kinetical rates even when using biased simulations usually requires accumulated trajectory data on the order of several μ s for a single ligand. Computation of explicit kinetics, therefore, is a precision tool that should only be used in the later stages of a drug development program.

The so far best-established method for unbinding rate prediction is infrequent metadynamics [39, 75]: filling up the initial free energy minimum of the bound state results in an increased transition probability in Eq. (3.8). The unbiased rate k can then be extracted from the biased results via

$$k = k_M \left\langle \exp \left(\frac{V_{bias,M}}{k_B T} \right) \right\rangle_M^{-1}, \quad (3.16)$$

with the mean rate observed in the M bias-accelerated simulations k_M and the bias potential $V_{bias,M}$ that needed to be added in each simulation until an unbinding event happened. The method has already been successfully applied to a range of proteins, from test systems [39, 48] to GPCRs [37, 57] and kinases [6, 76]. Due to the flexible implementation of the MD simulation interface PLUMED [77], a range of different reaction coordinates can be used for biasing. Judged by the 5 μ s of accumulated simulation data for the trypsin-benzamidine complex, the computational requirements are relatively low. As a downside, metadynamics requires a prior definition of a range of parameters in Eq. (3.11), especially amplitudes, width, and placement frequency, as well as the position of the added Gaussian functions, for which a suitable choice depends on the underlying free energy landscape and therefore is not known beforehand.

Concerning methods that do not require the introduction of a bias on dynamics, Milestoning approaches have been used in the form of the SEEKR algorithm [78, 79] on the example of the trypsin–benzamidine complex. This approach further reduces the computational cost for the generation of single trajectories via the implementation of Brownian Dynamics at distances larger than a suitable threshold from the binding site, where the details of near-ordering and dynamics of water molecules are not required to be taken into account anymore. Here, the unbinding and binding rates are calculated from the flux, i.e. the number of trajectories, crossing thresholds along x , and the respective time they require for doing so. The computational requirements are moderate: ca 20 μ s of accumulated MD simulation time were needed for the trypsin-benzamidine complex [78]. A similar approach is *adaptive multisplitting* [80], which resulted in a comparable prediction accuracy, albeit with only requiring ca 2.5 μ s simulation time for said complex.

A completely different approach is taken in *dissipation-corrected targeted MD* (dcTMD) [81]: calculating a bias work $W(x) = \int_{x_0}^x dx' f_{bias}(x')$ from the constant velocity constraint bias work of in Eq. (3.15). Based on a second-order cumulant expansion of Jarzynski's equality [82], free energies and friction profiles are then calculated as

$$\Delta F(x) = \langle W \rangle_N - \frac{1}{2k_B T} \langle \Delta W^2 \rangle_N \quad (3.17)$$

$$\Gamma(x) = \frac{1}{2k_B T v} \frac{d}{dx} \langle \Delta W^2 \rangle_N \quad (3.18)$$

with the mean and variance over a set of N -independent pulling simulations. The two profiles then serve as input for numerical integration of the Langevin equation (3.5), which already allows for the calculation of unbinding kinetics on the order of microseconds within a few minutes of wall clock time. To reach

biomedically relevant time scales, a temperature acceleration similar to that in Eq. (3.9) can be employed: determining $\Delta F(x)$ and $\Gamma(x)$ at a temperature of interest T and setting them to be independent of temperature, Langevin equation simulations at higher temperatures T' yield rate constants k' . These constants then can be scaled back to predict k at the temperature T via the *temperature-boosting* equation [33]

$$k = k' \exp \left(-\frac{\Delta F^\ddagger}{k_B} \left[\frac{1}{T} - \frac{1}{T'} \right] \right), \quad (3.19)$$

which in turn is based on Kramers rate theory Eq. (3.8). dcTMD and temperature-boosted Langevin simulations have been applied to several proteins such as the trypsin-benzamidine complex and the N-terminal domain of Hsp90 [33] as well as to a library of fragments bound to SARS-CoV-2 amin protease [83]. The advantage of this approach is a small computational requirement for pulling data acquisition, which was only 0.4 μ s for the case of the trypsin-benzamidine complex, while the cost of the Langevin equation simulations is negligible. In addition, $\Delta F(x)$ and $\Gamma(x)$ yield valuable information on the molecular effects determining the dynamics of protein-ligand unbinding. Recently, an implementation of dcTMD into the Galaxy platform [83] has become available, allowing for a close connection with chemoinformatic tools. The major limit of this approach is the second-order cumulant expansion mentioned above, which requires W to follow a normal distribution and only appears to hold for individual unbinding pathways [43, 56]. Before dcTMD analysis, pathways, therefore, have to be found based on inter-trajectory similarity.

3.4.2 Binding Rate Prediction

Contrary to the prediction of unbinding rates, few options exist to simulate binding events. A ligand finding the correct conformation and near-ordering of contacts with surrounding amino acids within the binding site simply takes time and cannot be biased. Conversely, enforcing unbinding simply requires breaking these contacts [72]. A few examples of binding simulations exist in the form of long-time simulations on the β_2 -adrenergic receptor with artificially high ligand concentrations [84]. Alternatively, structural coarse-graining of a protein and ligand can be employed to speed up binding simulations and assess details of this process [85], but this approach comes at the cost of losing the connection between coarse-grained simulation time and real-world (fine-grained) binding time. MSMs are able to capture the binding kinetics of molecular fragments [86] and of benzamidine to trypsin [87, 88], and it should be noted that they have been successfully applied for the prediction of protein-protein association rates [89], but these methods require a significant computational effort up to several hundred microseconds of total simulated time.

An exception here is the combination of dcTMD and temperature-boosted Langevin equation simulations [33]. As the free energy profiles generated with this approach cover the full range of x , both binding and unbinding events are sampled at the same time, i.e. binding and unbinding rates are obtained at the same time. Recently, a similar capability of capturing both dissociation and association

kinetics has been reported for *ligand Gaussian accelerated molecular dynamics* (LiGaMD) for the trypsin-benzamidine complex [90] and SARS-CoV2 protease [91]. Furthermore, SEEKR and adaptive multi-splitting provide information on both binding and unbinding rates, as well, as the respective calculated trajectory flux.

3.5 State-of-the-Art in Understanding Kinetics

With all these methods having been established, we now take a look at what knowledge has been gained on the molecular discriminants determining such kinetics and how this knowledge may be exploited for the design of drugs with tailored binding and unbinding rates. From Eq. (3.8), we see that both $\Delta F^\ddagger$ and Γ have an impact on kinetics, with the transition state being the major contributor. On a general basis, the ligand mass, hydrophobic effect, and protein conformational changes have been implied to impact residence times [7, 92]. A special role has been found for electrostatic interactions in the N-terminal domain of Hsp90 [72]: while a salt bridge between a ligand and the protein in the bound state may prolong the ligand's residence time, transient charge interactions between a ligand and an amino acid along the unbinding path can accelerate unbinding of bulkier ligands by facilitating ligand-internal conformational changes as well.

Major contributors to (un)binding dynamics seem to be protein- and ligand-bound water molecules: in trypsin [93] as well as in Hsp90 [52], binding site and ligand desolvation appear to steer the ligand binding rate, which is in agreement with experimental investigations [94]. An intermediate ligand desolvation site has been observed in the before-mentioned simulations of ligand binding to the β_2 -adrenergic receptor [84]. Conversely, ligand and binding site hydration shell formation or rearrangements cause a rise in friction and the relaxation of the free energy directly after the transition state in trypsin [33]. Therefore, both changing ligand charge and hydration characteristics appear as feasible start points for an optimization of small organic molecules kinetics.

3.6 Conclusion

While the field of calculating binding and unbinding kinetics is a comparatively new field, much progress has been made over the last decade. It is therefore feasible to assume that the fine-tuning of a drug's kinetic profile will become a standard option in the CADD toolbox in the coming decade. Currently, the largest barrier for this to happen is the high computational cost of such an undertaking. As an experimentalist once pointed out to the author of this chapter, computational predictions of kinetics are only helpful if they are either faster than synthesizing a library of small compounds or if they provide a significant (monetary or informational) benefit over such a brute-force experimental approach. The coming years will show if MD simulation-based predictions of kinetics, possibly with the help of machine learning models trained by MD input data [37, 38, 43, 95–97], can fulfill the promise of providing such benefits.

References

- 1 Copeland, R.A., Pompliano, D.L., and Meek, T.D. (2006). Drug–target residence time and its implications for lead optimization. *Nat. Rev. Drug Discov.* 5 (9): 730–739.
- 2 Swinney, D.C. (2006). Biochemical mechanisms of new molecular entities (NMEs) approved by United States FDA during 2001–2004: Mechanisms leading to optimal efficacy and safety. *Curr. Top. Med. Chem.* 6 (5): 461–478.
- 3 Swinney, D.C. (2012). Applications of Binding Kinetics to Drug Discovery. *Pharm. Med.* 22 (1): 23–34.
- 4 Copeland, R.A. (2016). The drug-target residence time model: a 10-year retrospective. *Nat. Rev. Drug Discov.* 15 (2): 87–95.
- 5 Agafonov, R.V., Wilson, C., Otten, R. et al. (2014). Energetic dissection of Gleevec’s selectivity toward human tyrosine kinases. *Nat. Struct. Mol. Biol.* 21 (10): 848–853.
- 6 Shekhar, M., Smith, Z., Seeliger, M.A., and Tiwary, P. (2022). Protein flexibility and dissociation pathway differentiation can explain onset of resistance mutations in kinases. *Angew. Chem. Int. Ed. Engl.* 61 (28): e202200983.
- 7 Segala, E., Guo, D., Cheng, R.K.Y. et al. (2016). Controlling the dissociation of ligands from the adenosine A2A receptor through modulation of salt bridge strength. *J. Med. Chem.* 59 (13): 6470–6479.
- 8 Amaral, M., Kokh, D.B., Bomke, J. et al. (2017). Protein conformational flexibility modulates kinetics and thermodynamics of drug binding. *Nat. Commun.* 8 (1): 2276.
- 9 Shaw, D.E., Adams, P.J., Azaria, A. et al. (2021). Anton 3: twenty microseconds of molecular dynamics simulation before lunch. In: SC, <https://doi.org/10.1145/3458817.3487397>.
- 10 Zwanzig, R.W. (2001). *Nonequilibrium Statistical Mechanics*. New York, NY: Oxford Univ. Press.
- 11 Pottier, N. (2010). *Nonequilibrium Statistical Physics: Linear Irreversible Processes*. New York, NY: Oxford Univ. Press.
- 12 Vogelsang, R. and Hoheisel, C. (1987). Determination of the friction coefficient via the force autocorrelation function. A molecular dynamics investigation for a dense Lennard-Jones fluid. *J. Stat. Phys.* 47 (1–2): 193–207.
- 13 Kramers, H.A. (1940). Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica* 7 (4): 284–304.
- 14 Hänggi, P., Talkner, P., and Borkovec, M. (1990). Reaction-rate theory: fifty years after Kramers. *Rev. Mod. Phys.* 62 (2): 251–341.
- 15 Hénin, J., Lelievre, T., Shirts, M.R. et al. (2022). Enhanced sampling methods for molecular dynamics simulations [Article v1.0]. *Living J. Comp. Mol. Sci.* 4 (1): 1583–1583. <https://doi.org/10.33011/livecoms.4.1.1583>.
- 16 Rensen, M.R.S. and Voter, A.F. (2000). Temperature-accelerated dynamics for simulation of infrequent events. *J. Chem. Phys.* 112 (21): 9599–9606.

- 17 Maragliano, L. and Vanden-Eijnden, E. (2006). A temperature accelerated method for sampling free energy and determining reaction pathways in rare events simulations. *Chem. Phys. Lett.* 426 (1–3): 168–175.
- 18 Abrams, C.F. and Vanden-Eijnden, E. (2010). Large-scale conformational sampling of proteins using temperature-accelerated molecular dynamics. *Proc. Natl. Acad. Sci. U. S. A.* 107 (11): 4961–4966.
- 19 Mollica, L., Decherchi, S., Zia, S.R. et al. (2015). Kinetics of protein-ligand unbinding via smoothed potential molecular dynamics simulations. *Sci. Rep.* 5: 11539.
- 20 Miao, Y., Feher, V.A., and McCammon, J.A. (2015). Gaussian accelerated molecular dynamics: unconstrained enhanced sampling and free energy calculation. *J. Chem. Theory Comput.* 11: 3584–3595.
- 21 Torrie, G.M. and Valleau, J.P. (1977). Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* 23 (2): 187–199.
- 22 Laio, A. and Parrinello, M. (2002). Escaping free-energy minima. *Proc. Natl. Acad. Sci. U. S. A.* 99 (20): 12562–12566.
- 23 Bussi, G. and Laio, A. (2020). Using metadynamics to explore complex free-energy landscapes. *Nat. Rev. Phys.* 23 (4): 1–13.
- 24 Grubmüller, H., Heymann, B., and Tavan, P. (1996). Ligand binding: molecular mechanics calculation of the streptavidin-biotin rupture force. *Science* 271 (5251): 997–999.
- 25 Izrailev, S., Stepaniants, S., Isralewitz, B. et al. (1999). Steered molecular dynamics, Computational Molecular Dynamics: Challenges, Methods, Ideas, Heidelberg. In: 39–65.
- 26 Isralewitz, B., Gao, M., and Schulten, K. (2001). Steered molecular dynamics and mechanical functions of proteins. *Curr. Opin. Struct. Biol.* 11 (2): 224–230.
- 27 Schlitter, J., Engels, M., and Krüger, P. (1994). Targeted molecular dynamics: a new approach for searching pathways of conformational transitions. *J. Mol. Graph.* 12 (2): 84–89.
- 28 Lüdemann, S.K., Lounnas, V., and Wade, R.C. (2000). How do substrates enter and products exit the buried active site of cytochrome P450cam? 1. Random expulsion molecular dynamics investigation of ligand access channels and mechanisms. *J. Mol. Biol.* 303 (5): 797–811.
- 29 Zuckerman, D.M. and Chong, L.T. (2017). Weighted ensemble simulation: review of methodology, applications, and software. *Annu. Rev. Biophys.* 46 (1): 43–57.
- 30 Faradjian, A.K. and Elber, R. (2004). Computing time scales from reaction coordinates by milestoning. *J. Chem. Phys.* 120 (23): 10880.
- 31 Dellago, C., Bolhuis, P.G., Csajka, F.S., and Chandler, D. (1998). Transition path sampling and the calculation of rate constants. *J. Chem. Phys.* 108 (5): 1964–1977.
- 32 Bolhuis, P.G., Chandler, D., Dellago, C., and Geissler, P.L. (2002). Transition path sampling: throwing ropes over rough mountain passes, in the dark. *Annu. Rev. Phys. Chem.* 53: 291–318.

- 33 Wolf, S., Lickert, B., Bray, S., and Stock, G. (2020). Multisecond ligand dissociation dynamics from atomistic simulations. *Nat. Commun.* 11 (1): 2918.
- 34 Bowman, G.R., Pande, V.S., and Noé, F. (2013). *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*. Springer Science & Business Media.
- 35 Thayer, K.M., Lakhani, B., and Beveridge, D.L. (2017). Molecular dynamics-Markov state model of protein ligand binding and allostery in CRIB-PDZ: conformational selection and induced fit. *J. Phys. Chem. B* 121 (22): 5509–5514.
- 36 Zwanzig, R.W. (1954). High-temperature equation of state by a perturbation method. I. Nonpolar gases. *J. Chem. Phys.* 22: 1420–1426.
- 37 Ribeiro, J.M.L., Provasi, D., and Filizola, M. (2020). A combination of machine learning and infrequent metadynamics to efficiently predict kinetic rates, transition states, and molecular determinants of drug dissociation from G protein-coupled receptors. *J. Chem. Phys.* 153 (12): 124105.
- 38 Badaoui, M., Buigues, P.J., Berta, D. et al. (2022). Combined free-energy calculation and machine learning methods for understanding ligand unbinding kinetics. *J. Chem. Theory Comput.* 18 (4): 2543–2555.
- 39 Tiwary, P., Limongelli, V., Salvalaglio, M., and Parrinello, M. (2015). Kinetics of protein–ligand unbinding: predicting pathways, rates, and rate-limiting steps. *Proc. Natl. Acad. Sci. U. S. A.* 112 (5): E386–E391.
- 40 Schuetz, D.A., Bernetti, M., Bertazzo, M. et al. (2019). Predicting residence time and drug unbinding pathway through scaled molecular dynamics. *J. Chem. Inf. Model.* 59 (1): 535–549.
- 41 Kokh, D.B., Doser, B., Richter, S. et al. (2020). A workflow for exploring ligand dissociation from a macromolecule: efficient random acceleration molecular dynamics simulation and interaction fingerprint analysis of ligand trajectories. *J. Chem. Phys.* 153 (12): 125102.
- 42 Bianciotto, M., Gkeka, P., Kokh, D.B. et al. (2021). Contact map fingerprints of protein-ligand unbinding trajectories reveal mechanisms determining residence times computed from scaled molecular dynamics. *J. Chem. Theory Comput.* 17 (10): 6522–6535.
- 43 Bray, S., Tänzle, V., and Wolf, S. (2022). Ligand unbinding pathway and mechanism analysis assisted by machine learning and graph methods. *J. Chem. Inf. Model.* 62 (19): 4591–4604.
- 44 Capelli, R., Carloni, P., and Parrinello, M. (2019). Exhaustive search of ligand and binding pathways via volume-based metadynamics. *J. Phys. Chem. Lett.* 3495–3499.
- 45 Capelli, R., Bochicchio, A., Piccini, G.M. et al. (2019). Chasing the full free energy landscape of neuroreceptor/ligand unbinding by metadynamics simulations. *J. Chem. Theory Comput.* 15 (5): 3354–3361.
- 46 Rydzewski, J. and Valsson, O. (2019). Finding multiple reaction pathways of ligand unbinding. *J. Chem. Phys.* 150 (22): 221101.

- 47 Fu, H., Zhou, Y., Jing, X. et al. (2022). Meta-analysis reveals that absolute binding free-energy calculations approach chemical accuracy. *J. Med. Chem.* 65 (19): 12970–12978.
- 48 Pramanik, D., Smith, Z., Kells, A., and Tiwary, P. (2019). Can one trust kinetic and thermodynamic observables from biased metadynamics simulations?: Detailed quantitative benchmarks on millimolar drug fragment dissociation. *J. Phys. Chem. B* 123 (17): 3672–3678.
- 49 Nunes-Alves, A., Kokh, D.B., and Wade, R.C. (2021). Ligand unbinding mechanisms and kinetics for T4 lysozyme mutants from tauRAMD simulations. *Curr. Res. Struct. Biol.* 3: 106–111.
- 50 Guillain, F. and Thusius, D. (1970). Use of proflavine as an indicator in temperature-jump studies of the binding of a competitive inhibitor to trypsin. *J. Am. Chem. Soc.* 92 (18): 5534–5536.
- 51 Schuetz, D.A., de Witte, W., Arnout, E. et al. (2017). Kinetics for drug discovery: an industry-driven effort to target drug residence time. *Drug Discov. Today* 22 (6): 896–911.
- 52 Schuetz, D.A., Richter, L., Amaral, M. et al. (2018). Ligand desolvation steers on-rate and impacts drug residence time of heat shock protein 90 (Hsp90) inhibitors. *J. Med. Chem.* 61 (10): 4397–4411.
- 53 Kokh, D.B., Amaral, M., Bomke, J. et al. (2018). Estimation of drug-target residence times by τ -random acceleration molecular dynamics simulations. *J. Chem. Theory Comput.* 14 (7): 3859–3869.
- 54 Peng, X., Zhang, Y., Chu, H. et al. (2016). Accurate evaluation of ion conductivity of the Gramicidin A channel using a polarizable force field without any corrections. *J. Chem. Theory Comput.* 12 (6): 2973–2982.
- 55 Ngo, V., Li, H., Mackerell, A.D. et al. (2021). Polarization effects in water-mediated selective cation transport across a narrow transmembrane channel. *J. Chem. Theory Comput.* 17 (3): 1726–1741.
- 56 Jäger, M., Koslowski, T., and Wolf, S. (2022). Predicting ion channel conductance via dissipation-corrected targeted molecular dynamics and Langevin equation simulations. *J. Chem. Theory Comput.* 18 (1): 494–502.
- 57 Capelli, R., Lyu, W., Bolnykh, V. et al. (2020). Accuracy of molecular simulation-based predictions of Koffvalues: a metadynamics study. *J. Phys. Chem. Lett.* 6373–6381.
- 58 Lopes, P.E.M., Huang, J., Shim, J. et al. (2013). Force field for peptides and proteins based on the classical Drude oscillator. *J. Chem. Theory Comput.* 9 (12): 5430–5449.
- 59 Shi, Y., Xia, Z., Zhang, J. et al. (2013). Polarizable atomic multipole-based AMOEBA force field for proteins. *J. Chem. Theory Comput.* 9 (9): 4046–4063.
- 60 Bruce, N.J., Ganotra, G.K., Kokh, D.B. et al. (2018). New approaches for computing ligand-receptor binding kinetics. *Curr. Opin. Struct. Biol.* 49: 1–10.
- 61 Ribeiro, J.M.L., Tsai, S.-T., Pramanik, D. et al. (2019). Kinetics of ligand-protein dissociation from all-atom simulations: Are we there yet? *Biochemistry* 58 (3): 156–165.

- 62 Bernetti, M., Masetti, M., Rocchia, W., and Cavalli, A. (2019). Kinetics of drug binding and residence time. *Annu. Rev. Phys. Chem.* 70: 143–171.
- 63 Nunes-Alves, A., Kokh, D.B., and Wade, R.C. (2020). Recent progress in molecular simulation methods for drug binding kinetics. *Curr. Opin. Struct. Biol.* 64: 126–133.
- 64 Limongelli, V. (2020). Ligand binding free energy and kinetics calculation in 2020. *WIREs Comput. Mol. Sci.* 8 (93): e1358.
- 65 Ahmad, K., Rizzi, A., Capelli, R. et al. (2022). Enhanced-sampling simulations for the estimation of ligand binding kinetics: current status and perspective. *Front. Mol. Biosci.* 9: 899805.
- 66 Wang, J., Do, H.N., Koirala, K., and Miao, Y. (2023). Predicting biomolecular binding kinetics: a review. *J. Chem. Theory Comput* 19 (8): 2135–2148. <https://doi.org/10.1021/acs.jctc.2c01085>.
- 67 Sohraby, F. and Nunes-Alves, A. (2023). Advances in computational methods for ligand binding kinetics. *Trends Biochem. Sci.* 48 (5): 437–449. <https://doi.org/10.1016/j.tibs.2022.11.003>.
- 68 Chen, Y.-C. (2015). Beware of docking! *Trends Pharmacol. Sci.* 36 (2): 78–95.
- 69 Mollica, L., Theret, I., Antoine, M. et al. (2016). Molecular dynamics simulations and kinetic measurements to estimate and predict protein-ligand residence times. *J. Med. Chem.* 59 (15): 7167–7176.
- 70 Bortolato, A., Deflorian, F., Weiss, D.R., and Mason, J.S. (2015). Decoding the role of water dynamics in ligand-protein unbinding: CRF1R as a test case. *J. Chem. Inf. Model.* 55 (9): 1857–1866.
- 71 Potterton, A., Hussein, F.S., Southey, M.W.Y. et al. (2019). Ensemble-based steered molecular dynamics predicts relative residence time of A2A receptor binders. *J. Chem. Theory Comput.* 15 (5): 3316–3330.
- 72 Wolf, S., Amaral, M., Lowinski, M. et al. (2019). Estimation of protein-ligand unbinding kinetics using non-equilibrium targeted molecular dynamics simulations. *J. Chem. Inf. Model.* 59 (12): 5135–5147.
- 73 Kokh, D.B. and Wade, R.C. (2021). G protein-coupled receptor-ligand dissociation rates and mechanisms from τ RAMD simulations. *J. Chem. Theory Comput.* 17 (10): 6610–6623.
- 74 Berger, B.-T., Amaral, M., Kokh, D.B. et al. (2021). Structure-kinetic relationship reveals the mechanism of selectivity of FAK inhibitors over PYK2. *Cell Chem. Biol.* 28 (5): 686–698.e7.
- 75 Tiwary, P. and Parrinello, M. (2013). From metadynamics to dynamics. *Phys. Rev. Lett.* 111 (23): 230602.
- 76 Casasnovas, R., Limongelli, V., Tiwary, P. et al. (2017). Unbinding kinetics of a p38 MAP kinase type II inhibitor from metadynamics simulations. *J. Am. Chem. Soc.* 139 (13): 4780–4788.
- 77 The PLUMED Consortium (2019). Promoting transparency and reproducibility in enhanced molecular simulations. *Nat. Methods* 16 (8): 670–673.
- 78 Votapka, L.W., Jagger, B.R., Heyneman, A.L., and Amaro, R.E. (2017). SEEKR: simulation enabled estimation of kinetic rates, a computational tool to estimate

- molecular kinetics and its application to trypsin-benzamidine binding. *J. Phys. Chem. B* 121 (15): 3597–3606.
- 79 Jagger, B.R., Ojha, A.A., and Amaro, R.E. (2020). Predicting ligand binding kinetics using a Markovian milestoning with Voronoi tessellations multiscale approach. *J. Chem. Theory Comput.* 16 (8): 5348–5357.
 - 80 Teo, I., Mayne, C.G., Schulten, K., and Lelievre, T. (2016). Adaptive multilevel splitting method for molecular dynamics calculation of benzamidine-trypsin dissociation time. *J. Chem. Theory Comput.* 12 (6): 2983–2989.
 - 81 Wolf, S. and Stock, G. (2018). Targeted molecular dynamics calculations of free energy profiles using a nonequilibrium friction correction. *J. Chem. Theory Comput.* 14: 6175–6182.
 - 82 Jarzynski, C. (1997). Equilibrium free-energy differences from nonequilibrium measurements: a master-equation approach. *Phys. Rev. E* 56 (5): 5018–5035.
 - 83 Bray, S., Dudgeon, T., Skyner, R. et al. (2022). Galaxy workflows for fragment-based virtual screening: a case study on the SARS-CoV-2 main protease. *J. Chem.* 14 (1): 1–13.
 - 84 Dror, R.O., Pan, A.C., Arlow, D.H. et al. (2011). Pathway and mechanism of drug binding to G-protein-coupled receptors. *Proc. Natl. Acad. Sci. U. S. A.* 108 (32): 13118–13123.
 - 85 Souza, P.C.T., Thallmair, S., Conflitti, P. et al. (2020). Protein–ligand binding with the coarse-grained Martini model. *Nat. Commun.* 11 (1): 1–11.
 - 86 Linker, S.M., Magarkar, A., Köfinger, J. et al. (2019). Fragment binding pose predictions using unbiased simulations and Markov-state models. *J. Chem. Theory Comput.* 15 (9): 4974–4981.
 - 87 Buch, I., Giorgino, T., and De Fabritiis, G. (2011). Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proc. Natl. Acad. Sci. U. S. A.* 108 (25): 10184–10189.
 - 88 Plattner, N. and Noé, F. (2015). Protein conformational plasticity and complex ligand-binding kinetics explored by atomistic simulations and Markov models. *Nat. Commun.* 6 (1): 7653.
 - 89 Plattner, N., Doerr, S., De Fabritiis, G., and Noé, F. (2017). Complete protein–protein association kinetics in atomic detail revealed by molecular dynamics simulations and Markov modelling. *Nat. Chem.* 9: 1005–1011.
 - 90 Miao, Y., Bhattarai, A., and Wang, J. (2020). Ligand Gaussian accelerated molecular dynamics (LiGaMD): characterization of ligand binding thermodynamics and kinetics. *J. Chem. Theory Comput.* 16: 5526–5547.
 - 91 Wang, Y.-T., Liao, J.-M., Lin, W.-W. et al. (2022). Structural insights into Nir-matrelvir (PF-07321332)-3C-like SARS-CoV-2 protease complexation: a ligand Gaussian accelerated molecular dynamics study. *Phys. Chem. Chem. Phys.* 24 (37): 22898–22904.
 - 92 Pan, A.C., Borhani, D.W., Dror, R.O., and Shaw, D.E. (2013). Molecular determinants of drug–receptor binding kinetics. *Drug Discov. Today* 18 (13–14): 667–673.
 - 93 Ansari, N., Rizzi, V., and Parrinello, M. (2022). Water regulates the residence time of Benzamidine in Trypsin. *Nat. Commun.* 13 (1): 1–9.

- 94 Schiebel, J., Gaspari, R., Wulsdorf, T. et al. (2018). Intriguing role of water in protein-ligand binding studied by neutron crystallography on trypsin complexes. *Nat. Commun.* 9 (1): 166.
- 95 Ribeiro, J.M.L. and Tiwary, P. (2018). Towards achieving efficient and accurate ligand-protein unbinding with deep learning and molecular dynamics through RAVE. *J. Chem. Theory Comput.* 15 (1): 708–719.
- 96 Brandt, S., Sittel, F., Ernst, M., and Stock, G. (2018). Machine learning of biomolecular reaction coordinates. *J. Phys. Chem. Lett.* 2144–2150.
- 97 Komp, E., Janulaitis, N., and Valleau, S. (2022). Progress towards machine learning reaction rate constants. *Phys. Chem. Chem. Phys.* 24 (5): 2692–2705.

4

Solvation Thermodynamics and its Applications in Drug Discovery

Kuzhanthaivelan Saravanan and Ramesh K. Sistla

thinkMolecular Technologies Pvt. Ltd., #03, 1st Cross, Reliaable Tranquil Layout, Bengaluru, Karnataka 560102, India

yo'pām āyatanam veda | āyatanavān bhavati

– taittirīya aruṇa praśna – 1.72

He who knows the position of water, secures his position.

4.1 Introduction

Water is the major constituent in all organisms. All life processes play out in the medium of water. Therefore, it is essential to understand the role of the aqueous medium in various life processes. The role of the solvent in the processes of protein folding [1, 2] and molecular recognition [3] is well documented in the literature. The hydrophobic effect [4] has been proposed to be the key reason for protein folding. A simplistic demonstration of the hydrophobic effect lies in the immiscibility of oil and water and the coming together of oil particles on a water surface. Extending this visualization further, the hydrophobic amino acids in a protein move away from the solvent and come together to form the hydrophobic core of the protein. This was demonstrated in a graphical manner through hydropathy plots proposed by Kyte and Doolittle [5]. The partitioning of the nonpolar amino acids into the lipid membranes and the formation of the hydrophobic core in globular proteins unambiguously suggest the role of the aqueous medium in which the proteins are present.

4.1.1 Protein Folding

Proteins are synthesized in the ribosomes inside the cells. Upon synthesis, the proteins immediately fold into three-dimensional structures and perform precise functions ranging from scaffolding to being carriers to performing catalysis. The protein folding phenomenon has been the subject of extensive research over the last several decades.

Computational Drug Discovery: Methods and Applications, First Edition.

Edited by Vasanthanathan Poongavanam and Vijayan Ramaswamy.

© 2024 WILEY-VCH GmbH. Published 2024 by WILEY-VCH GmbH.

The “thermodynamic hypothesis” for protein folding was suggested by Anfinsen [6], who postulated that, out of the numerous conformations that a protein can adopt, it chooses the conformation in the physiological milieu whose Gibbs-free energy is the lowest. The Gibbs free energy of a system has two components and is given as follows:

$$\Delta G = \Delta H - T\Delta S \quad (4.1)$$

where ΔG is the change in Gibbs free energy between two states, and the corresponding changes in enthalpy and entropy are denoted by ΔH and ΔS , respectively. T represents the temperature in Kelvin.

Enthalpy is the total energy of the system and comprises electrostatic and dispersion terms. These components of enthalpy manifest in a variety of inter-atomic interactions, such as hydrogen bonds, van der Waals interactions, aromatic π -stacking, and hydrophobic interactions, to name a few. For the folded conformation of the protein, ΔH is negative compared to the unfolded conformation. This essentially results from the favorable interactions between hydrophobic amino acids present in the core of a folded globular protein and also from the key hydrogen bonds and salt bridges that are formed in the secondary structures of the protein upon folding. The enthalpy is further boosted by the interactions of the exposed polar amino acids of the protein with the solvent. However, it must be borne in mind that the unfolded polypeptide chain makes several polar interactions with the solvent. When this chain folds, these waters are removed out into what is known as the bulk solvent, and the polar interactions are formed among the amino acid residues. Therefore, a net change in enthalpy during the process of protein folding could be modest [7].

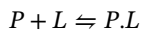
The entropy of a system is a measure of its randomness. It has its basis in the second law of thermodynamics. In simple terms, it means the number of degrees of freedom that are available to a system. In the case of the polypeptide chain in the aqueous medium, it means the number of degrees of freedom that each atom has. Water molecules near hydrophobic amino acids are highly constrained and hence have lower entropy. When the protein folds, these waters are removed to the bulk solvent, thereby enhancing the entropy of the solvent. The gain in entropy adds favorably to the Gibbs free energy of the system (protein + solvent).

Thus, even if the gain of enthalpy during the folding is modest, the contribution of the solvent entropy to Gibbs free energy is significant. Thus, it is the solvent that funds the energy needed for protein folding through the maximization of its entropy [7]. The entropy term is the sum of the entropy of the solvent, the conformational entropy of the protein, and the rotational/translational entropy. While the conformational and rotational/translational entropy of the protein might reduce during the folding, the enhancement in solvent entropy outdoes the reduction in the other two.

4.1.2 Protein–Ligand Interactions

The recognition of various proteins by their cognate ligands, cofactors, and other molecules and their complex formation are also influenced immensely by the solvent in which the molecules are present. The protein–ligand complexes are formed

only when the energy exchange between the protein, ligand, solvent, and any other essential component like ions present there is favorable. In other words, just as in the case of protein folding, the complex formation is favored when the Gibbs free energy of the whole system is negative [8]. The association and dissociation of a protein (P) and ligand (L) can be written in the following manner:



This reaction has an on rate of K_{on} , which represents the association of P and L , and an off rate of k_{off} signifying the dissociation of the ligand from the protein. The dissociation constant of the ligand K_d is given as

$$K_d = \frac{k_{off}}{K_{on}} = \frac{[P].[L]}{[P.L]} \quad (4.2)$$

This representation and estimation of K_d looks as if this phenomenon of the association of a protein and the ligand is purely binary and the solvent has little or no role to play. However, in reality, the role of the solvent is in fact rolled into both k_{off} and K_{on} and thus the solvent dictates the dissociation or association of a complex.

In physical terms, prior to complex formation, the protein and the ligand are individually solvated. Upon the formation of a collision complex, the ligand and the binding site undergo desolvation, and subsequently the protein and the ligand form a complex. The considerations of entropy and enthalpy between the protein and ligand are again on the lines similar to what we have seen in the protein folding in the preceding section. The solvent that interacted with the protein binding site and the ligand earlier, upon desolvation, is removed to the bulk, and its entropy increases. The association between the protein and the ligand, on the other hand, contributes to favorable interactions between complementary groups and causes an enthalpic gain. The gain in enthalpy of the protein–ligand system and the increase in the entropy of the solvent ensure an overall negative Gibbs free energy, which favors complex formation.

The process of a protein and ligand forming a complex is graphically and conceptually shown by the Born Haber cycle [9] shown below in Figure 4.1.

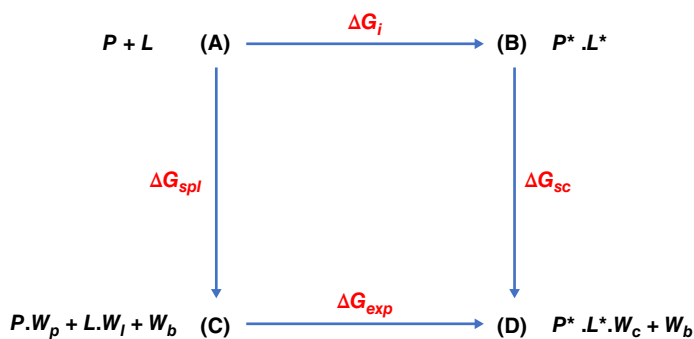


Figure 4.1 Born Haber cycle.

The Born Haber cycle provides a rigorous treatment of protein–ligand complex formation. In Figure 4.1, state **A** represents the protein and ligand, state **B** represents the unsolvated protein–ligand complex, state **C** represents the solvated states of the protein and ligand (*W* stands for water, its subscripts *p*, *l*, *c*, and *b* stand for protein, ligand, complex, and bulk, respectively) and state **D** represents the solvated protein–ligand complex and the bulk water. The transition from one state to another in the above process is associated with a change in Gibbs free energy. ΔG_i represents the intrinsic change in the free energy upon the complex formation ($A \rightarrow B$). ΔG_{exp} represents the change in free energy that would be measured experimentally by techniques such as isothermal titration calorimetry (ITC). The other two terms ΔG_{spl} ($A \rightarrow C$) and ΔG_{sc} ($B \rightarrow D$) represent the solvation-free energies (SFEs) of the individual components and the complex, respectively.

The process $A \rightarrow C$ represents the solvation of the individual components – the protein and the ligand. The process $B \rightarrow D$ represents the solvation of the protein–ligand complex. The energies estimated in these processes are solvation energies, and those values with the opposite sign represent the desolvation energies. In the process of a protein–ligand complex formation, **A** is the initial state and **D** is the final state. Free energy being dependent only on the state and not on the path the system takes, or, in other words, being a state function, the energy equation can be written as

$$\Delta G_i + \Delta G_{sc} = \Delta G_{spl} + \Delta G_{exp} \quad (4.3)$$

Rearranging,

$$\Delta G_{exp} = \Delta G_i + (\Delta G_{sc} - \Delta G_{spl}) \quad (4.4)$$

The term in the bracket in Eq. (4.4) directly relates to the free energy of solvation of the complex and that of the individual components. Thus, the observed free energy change upon complex formation is a sum of the intrinsic change in the free energy and the change in free energy of solvation.

The free energy is in turn composed of enthalpy (*H*) and entropy (*S*). As referenced in the earlier section, the enthalpy is a result of various intermolecular interactions, and the entropy signifies the order or lack thereof of a system. The enthalpic change of solvation/desolvation may not always be favorable (negative). For example, when the solvent molecules are near hydrophobic groups of the protein/ligand and upon complexation move into the bulk, there is a net positive change in the enthalpy. However, the movement of the solvent to the bulk results in enhanced entropy. This change in entropy contributes favorably to the Gibbs free energy of the system and makes complex formation favorable. It is well known that the free energy change is related to the affinity (rate constant).

Thus far, we have seen a qualitative treatment of the process of solvation and the role played by solvent in protein folding as well as protein–ligand complex formation. The quantitative treatment of the contribution of water molecules to the entropy of the system through experimental techniques can be very limiting. NMR techniques provide average properties of the solvent that rapidly exchange between the binding site and the bulk solvent. Knowledge of the positions of water molecules

by X-ray crystallography is highly resolution-dependent, and even in the best cases, only the structurally conserved water molecules can be assessed with confidence.

Docking the ligands in the binding sites of protein, and designing compounds based on the various interactions that are made by the ligand with the protein is a routine exercise in drug discovery. The ligands that dock in the binding site are ranked by the docking score, which is estimated using various scoring functions [10]. Ideally, the docking score computed from physics-based scoring functions should reflect the binding energy of the ligand. However, due to the limited accuracy of the forcefields employed, not considering the protein flexibility, and most importantly, not taking the solvent effect into account, the docking scores are not representative of the binding energy.

The effect of solvent is estimated in molecular modeling in two ways – (i) implicit solvent considerations and (ii) explicit solvent considerations. In either case, the goal is to incorporate the solvent effect into the estimation of the protein–ligand free energy.

In the case of the implicit solvent models, a dielectric continuum is considered in which the interactions between atoms of the protein and ligand (solute) take place. These interactions are influenced by the dielectric constant of the medium. Molecular mechanics Poisson–Boltzmann surface area (MM/PBSA) and molecular mechanics generalized Born surface area (MM/GBSA) [11] are popular endpoint free energy calculation methods that are more accurate compared to corresponding docking scores. Since it was first proposed [12] as an electrostatic interaction of a solute with the continuum, the implementations of the implicit solvent method have evolved and shown in several examples to result in better prediction of the binding affinity of the ligands [13]. As the name suggests, the medium in which the solute is present is represented by the dielectric, and there is no explicit consideration of solvent molecules in this approach. It is fast and can be employed in virtual screening campaigns. More recently, a machine-learning polarized continuum solvation model has been proposed [14]. The linear interaction energy is another method for estimating the ligand binding free energies [15]. This is based on forcefields and is a simplified method with several approximations, requiring only the intermolecular interactions between the ligand and the receptor to estimate the energy.

None of these methods treat the entropy associated with the solvent at all, and their limitations can be as such ascribed to that. In spite of this, these methods are computationally efficient and more accurate than docking scores and hence have great utility in large-scale virtual screening campaigns [16]. However, it has also been pointed out [17] that both the PBSA and GBSA performed poorly in estimating the solvation free energies (SFE) compared to explicit solvent methods. The main reason for this is the poor estimation of apolar contributions, which stem from solvent entropy.

High-resolution crystal structures reveal the presence of only strongly interacting water molecules. Other water molecules that have a transient presence at certain sites do not show up in the structures. This being the case, treating the two regions as a homogenous dielectric medium can introduce errors in the results of the application of continuum theory. There are several examples in the literature where conserved water molecules play a key role in the binding of ligands. The directionality of

the hydrogen bonds formed by the water molecules is an important determinant [18] in the potency of several molecules, which are not adequately treated by the continuum methods [19]. In this backdrop, it is imperative that the solvent be studied in an explicit manner rather than as a dielectric continuum.

Molecular dynamics (MD) [20] is a handy tool to treat water molecules explicitly and study their interactions with proteins and ligands. While the stability of the various protein–ligand interactions and conformational change of proteins is routinely studied by MD simulations, generally not much attention is placed on the estimations of the energetics of the water molecules, which are present in the protein pockets.

The enthalpic and entropic contributions of the water molecules are studied in a few different approaches – thermodynamic integration (TI), inhomogeneous solvation theory (IST), and reference interaction site models (RISM). The TI method computes the energy needed for the extraction of a water molecule from a particular position in the pocket.

The IST method draws information from the phase space of the solute–solvent complex generated from MD simulations. It considers the solute molecule to be central, evaluates the fluctuation of the density of water molecules, and estimates the enthalpies and entropies of these waters. Consequently, the implementation of IST is computationally intensive, but the reward is that the results are as accurate as the force field that is employed in these calculations. This approach has been implemented in some of the modern tools, such as Grid inhomogeneous solvation theory (GIST) [19, 21], Watermap [22], solvation thermodynamics of ordered water (STOW) [23], and solvation structure and thermodynamic mapping (SSTMap) [24]. Adequate sampling is needed to get accurate results with this method.

RISM, on the other hand, is a statistical mechanical integral approach [25]. This does not necessitate an extensive MD simulation and utilizes the optimized configuration of a solute–solvent complex. It then estimates for each solvent site a susceptibility function that is dependent on the positions and interactions of the solvent molecules with the solute and its polarizability. It solves the classic molecular Ornstein–Zernike equation that relates the correlation function between two atoms to the total correlation function of the solute and the solvent. Thus, it seeks to construct the molecular density distribution through atomic densities. The RISM approach has been implemented in several tools, viz. 3D-RISM, GCT, SZMAP, and WATsite.

A detailed description of each tool is beyond the scope of this chapter. In order to provide an insight into the contrasting techniques, in the upcoming section, a brief overview of the principles followed in the tools Watermap, GIST, and 3D-RISM will be presented, followed by case studies from the literature in Section 4.3.

4.2 Tools to Assess the Solvation Thermodynamics

To recap the earlier section, the Gibbs free energy has two components: enthalpy and entropy. The increase in solvent entropy has been pointed out to have a favorable

contribution to the overall Gibbs free energy of the system. The enthalpy component is computed by summing up various interatomic interactions using the parameters provided in the force fields. We have also discussed the inadequacies of continuum solvent models. Therefore, a reasonable treatment of the entropy of the system, and particularly the solvent entropy, is in order. We have touched upon the general methods that are employed to compute solvent entropy. Now in this section, we will deal briefly with the principles behind the tools Watermap [22], GIST [19], and 3D-RISM [26].

4.2.1 Watermap

Watermap [22] seeks to characterize the waters in the binding site as per their energies as happy (low energy) and unhappy (high energy) waters. It does so by estimating the enthalpy and the entropy associated with individual waters. MD simulation of a rigid solute in a solvent box is performed to study the behavior of individual waters in the binding site under the influence of the interatomic forces of the amino acids in the binding pocket. The protein molecule is constrained throughout the simulation, and the hydration of the binding sites is studied. The water molecules that hydrate the binding site are of three different types. The first kind are those that make a full complement of hydrogen bonds in the binding site. The second kind are those that are fully satisfied enthalpically but do not form hydrogen bonds optimally. The third are those that are held by weak forces in the binding site and thus have their degrees of freedom severely restricted. The second and third kinds of waters, when they are replaced by ligand groups that make complementary interactions, boost the binding energy by moving the constrained water into the bulk solvent and contribute substantially to the binding energy. There is no advantage in displacing waters that are optimally bound in the cavities, as there will be a significant loss of enthalpy upon displacement, which may not be compensated by an increase in entropy or may just be a zero-sum game. The first type of water is termed “happy water” and the second and third types are termed “unhappy water.”

Watermap divides the cavities in proteins into subvolumes. Each subvolume is termed a hydration site. A clustering algorithm scans through all the subvolumes and calculates the solvent density and solvent exposure at each point. The average number of water neighbors of each subvolume determines the solvent exposure at that point, and the degree of exposure is determined with respect to the solvent exposure of the bulk solvent. After identifying the hydration sites, the IST is used to determine the entropic cost of solvent ordering. Using this entropy calculation, the interaction energy of a water molecule at each hydration site with the rest of the system is computed.

The partial excess entropy of a given hydration site is computed through numerical integration, considering the orientational and spatial correlations of the water molecule at the given hydration site as per the following equation.

$$S^e = -\frac{k_b \rho_w}{\Omega} \int g_{sw}(r, \omega) \ln(g_{sw}(r, \omega)) dr d\omega \quad (4.5)$$

$$\approx k_b \rho_w \int g_{sw}(r) \ln(g_{sw}(r)) dr - \frac{k_b N_W^V}{\Omega} \int g_{sw}(\omega) \ln(g_{sw}(\omega)) d\omega \quad (4.6)$$

In this equation, r signifies the cartesian positions, while ω is the Euler angle orientation of the water molecules. The distribution functions (g) and the density of the bulk water (ρ) are considered. Other enhancements of this entropy estimation are possible by including higher-order terms, but they increase the computational cost.

The system interaction energy computed for each hydration site characterizes the water as high-energy (unhappy) or low-energy (happy) water. Displacement of an unhappy water by a favorable group on the ligand has been shown to explain the structure-activity relationships observed in a congeneric series of ligands [27]. An example of a prospective use of Watermap in a computational triage will be presented in the next section.

Some of the limitations of Watermap may stem from the consideration of the protein as rigid and not considering higher-order entropy terms. However, neither of these limitations is general in nature. In fact, Watermap reveals a lot more waters [22] than most high-resolution crystal structures and also provides the energetics associated with those waters thus providing a very important qualitative insight into the hydration of the binding site.

4.2.2 GIST

GIST is another implementation of the IST [19]. As the name suggests, it solves the equations of IST on a grid made of discrete cells called voxels in a region of interest in the protein. The values of the solvation entropies, enthalpies, and free energies in each voxel are computed. The summation of these parameters is then done on a trajectory obtained from the MD simulations of a protein in a chosen solvent box, which yields the solvation thermodynamics. Unlike Watermap, GIST does not limit the calculations to the high-density water hydration sites but rather estimates the thermodynamic parameters in every voxel and thus provides a smooth variation of the character of water at every position in the region of interest on the protein. Thus, the hydration thermodynamic information is independent of the density of water in a given region relative to the bulk solvent. This is especially useful to identify the sites that are partially occupied by water molecules. Mapping each voxel with the calculated parameters, GIST is able to compute free energy from the states.

The solvation entropy of a flexible solute is given as

$$\int p(q) \Delta S_{solv}(q) dq \quad (4.7)$$

where $p(q)$ is the Boltzmann probability and ΔS_{solv} is solvation entropy. q defines the coordinates of the system. The solvation entropy is a sum of the solute–water entropy and the water–water entropy.

$$\Delta S_{solv} = \Delta S_{sw} + \Delta S_{ww} \quad (4.8)$$

Ignoring the water–water entropy term, which deals with the bulk water and does not concern us in the region of the binding sites,

$$\Delta S_{solv} \approx \Delta S_{sw} \equiv k_B \frac{\rho^0}{8\pi^2} \int g_{sw}(r, \omega) dr d\omega \quad (4.9)$$

where k_B is the Boltzmann constant, g_{sw} is the pair correlation function between the solute–water and is a function of the cartesian coordinates (r) and the Euler angles (ω). The free energy would be a combination of the system interaction energy (enthalpy solute–water and water–water), solute–water translational entropy, and the solute–water rotational entropy (together the entropy of solvation). The above equation is discretized and the entropies are estimated for every voxel k and summed over the voxels in a region R .

The translational entropy for the voxel k is given as

$$\Delta S_{sw}^{trans}(r_k) \equiv k_B \rho^0 \int_k g(r) \ln g(r) dr d\omega \approx k_B \rho^0 V_k g(r_k) \ln g(r_k) \quad (4.10)$$

$$g(r_k) \equiv \frac{N_k}{\rho^0 V_k N_f} \quad (4.11)$$

The translational entropy over a region R is the sum of those over the voxels in the region and is given as

$$\Delta S_{sw}^{R,trans} \approx \sum_{k \in R} \Delta S_{sw}^{trans}(r_k) \quad (4.12)$$

The normalized entropy (per water in the region R containing n^R number of waters) would then become

$$\Delta S_{sw}^{R,trans,norm} \equiv \frac{\Delta S_{sw}^{R,trans}}{n^R} \quad (4.13)$$

The orientational entropy is also estimated similarly for each voxel, summed over all the voxels in a region R , and is given as

$$\Delta S_{sw}^{R,orient,norm} \equiv \frac{\Delta S_{sw}^{R,orient}}{n^R} \quad (4.14)$$

The solute–water interaction energy as well as the water–water interaction energy are also discretized for each voxel k , summed over a region R and finally normalized as given below.

$$\Delta E_{sw}(r_k) \equiv \int_k \Delta E_{sw}(r) dr \quad (4.15)$$

$$\Delta E_{sw}^R = \sum_{k \in R} \Delta E_{sw}(r_k) \quad (4.16)$$

$$\Delta E_{sw}^{R,norm} = \frac{\Delta E_{sw}^R}{n^R} \quad (4.17)$$

The water–water interactions are defined between voxels k and l

$$E_{ww}^{R,corr} = \sum_{k \in R} E_{ww}(r_k) - \frac{1}{2} \sum_{k \in R} \sum_{l \neq k, l \in R} E_{ww}(r_k, r_l) \quad (4.18)$$

$$E_{ww}^{R,norm} = \frac{\sum_{k \in R} E_{ww}(r_k)}{n^R} \quad (4.19)$$

From the above terms, the total energy and entropy for each voxel are calculated as follows.

$$\Delta E_{total}(r_k) \equiv \Delta E_{sw}(r_k) + \Delta E_{ww}(r_k), \quad (4.20)$$

$$\Delta S_{sw}^{total}(r_k) = \Delta S_{sw}^{trans}(r_k) + \Delta S_{sw}^{orient}(r_k), \quad (4.21)$$

These two terms yield the free energy of every voxel

$$\Delta G(r_k) = \Delta E_{total}(r_k) - T\Delta S_{sw}^{total}(r_k) \quad (4.22)$$

Thus, GIST is an elegant implementation of the IST, which splits the region into a grid made of voxels and discretizes the energy and entropy of each voxel. Hence, it maps out the regions of favorable and unfavorable free energies in the protein with respect to the solvent, providing insight into the design of ligands. More recent work [21] has extended the GIST framework to use chloroform as a solvent to simulate the membrane permeability of molecules, and further, the method has been generalized to include any rigid molecule as a solvent. This important enhancement extends the utility of the tool to important parameters other than protein–ligand potency optimization.

4.2.3 3D-RISM

The RISM in the context of a solute–water interaction was proposed by Kovalenko [25]. Subsequently, it has been implemented in several software packages as an approach to study solvation through the three-dimensional RISM (3D-RISM) technique [26]. As rigorous as the formulation of the IST method and its implementation are, the results are influenced by the amount of sampling that the solvent undergoes. The adequacy of sampling in a given simulation is always a matter of contention.

The 3D-RISM method attempts to circumvent this by taking a purely statistical mechanical approach to the problem and applying the integral equation theory. In this method, first, the rigid solute is subjected to a standard 3D-RISM calculation. Subsequently, solvent molecules are placed at different sites, and the solvent distribution function $g(r)$, total correlation function $h(r)$, direct correlation $c(r)$, and the local interaction potential $u(r)$ at every solvent site are calculated and iterated until they are self-consistent.

Both the positions and the orientations of the solvent molecules are optimized until a preset cutoff is reached. A local population function and the location of maximum probability are computed through iterations to identify the solvent distribution. After identifying the locations of the solvents through this, the orientational distribution is identified.

Based on the thorough characterization of the solvent sites and estimation of the distributions of solvent density, the energy and the entropy of each site are then calculated using the following equations.

$$E_n^1 = \rho_0 \sum_{\gamma} \int_{V_n} g_{\gamma}(r) u_{\gamma}(r) dr \quad (4.23)$$

$$S_n^{1,trans} = -\frac{\rho_0 k_B}{N_{rot}} \int_{V_n} g_{anchor}(r) dr \int_{\omega} g(\omega r) \ln g(\omega r) d\omega \quad (4.24)$$

Having briefly touched upon the methods employed in tools like Watermap, GIST, and 3D-RISM, case studies involving these tools from the literature are presented in the upcoming section. They are not meant to demonstrate the exhaustive use of these tools but rather to provide a flavor of their utilities.

4.3 Case Studies

4.3.1 Watermap

4.3.1.1 Background and Approach

Group-I p21-activated kinase 1 (PAK1) is essential for various cellular functions such as cytoskeletal organization, motility, mitosis, and angiogenesis. These roles make it an attractive therapeutic target for cancer [28], infectious diseases, and neurological disorders [29]. However, designing a highly selective PAK1 inhibitor is quite challenging because of the high homology of the kinase domain with other kinases.

In this work [30], a synergistic computational approach was used to repurpose the FDA-approved drugs from the Drugbank database [31] against PAK1. This synergistic approach includes molecular docking to understand the binding modes and potential affinity of the drug molecules revealed through noncovalent interaction energies. Since routine docking does not take into account the solvation effects, the authors utilized Watermap, which is a useful tool to predict the effects of explicit water molecules in the binding sites as described earlier. In this work, a short 2 ns MD simulation is performed using the Grand Canonical Monte Carlo (GCMC) sampling to predict structurally weak water clusters in the protein binding pocket.

The crystal structure of PAK1 kinase domain in complex with FRAX597 inhibitor (PDB id: 4EQC) with a resolution of 2.01 Å was selected for this study. The virtual screening process of the curated Drugbank molecules against PAK1 was performed using GLIDE molecular docking software in the Schrodinger suite. Out of 2162 FDA-approved drugs from the Drugbank database, 27 compounds were shortlisted based on the interactions that the docked poses made with the protein. These 27 compounds were then assessed with respect to the hydration site displacements predicted by Watermap in determining the binding affinity gains likely to be made by these compounds.

4.3.1.2 Results and Discussion

The Watermap calculations facilitated the identification of localized hydration sites around the binding cavity of PAK1, breaking down their thermodynamic energy profiles, viz. enthalpy (ΔH), entropy ($T\Delta S$), and differential binding energy ($\Delta\Delta G$). Based on $\Delta\Delta G$, the overlapping hydration sites on ligand functional groups could be categorized into displaceable ($\Delta\Delta G \gg 0$ and $\Delta H \gg 0$), replaceable ($\Delta H \ll 0$ and yet $\Delta\Delta G \gg 0$ or ≈ 0), and stable ($\Delta\Delta G \ll 0$) water molecules. Out of the 27 drug molecules shortlisted through virtual screening by docking, 20 drug molecules were

observed to displace a smaller number of high-energy hydration sites and hence may not offer better binding affinity to PAK1. Thus, the remaining seven molecules, namely, Mitoxantrone, Labetalol, Acalabrutinib, Sacubitril, Flubendazole, Trazodone, and Niraparib were shortlisted based on displacing more unfavorable hydration sites. The molecules were considered for exploring the implications of hydration site displacements in determining the binding affinity gains. The Watermap results indicated that the hinge region of PAK1 had two high-energy hydration sites (5.27 and 2.59 kcal/mol), and the remaining four high-energy hydration sites were concentrated in the back pocket of PAK1 (lining the gatekeeper residue) with hydration energies ranging from 3.93 to 6.44 kcal/mol. Except for Mitoxantrone, all other molecules were found positioned toward the hydrophobic back pocket of PAK1. It was observed that molecules Acalabrutinib, Flubendazole, and Trazodone had thermodynamically greater unfavorable hydration sites that were displaced, and therefore are anticipated to have stronger binding affinity gains to PAK1.

These shortlisted molecules were subjected to stability analysis using MD simulations, followed by molecular orbital and electrostatic surface potential analyses. The drug molecules Flubendazole, Niraparib, and Acalabrutinib scored better than others and are expected to have better binding affinity for PAK1. The three molecules that were identified thus from a large collection of >2000 molecules from the Drugbank would need further experimental testing, as noted by the authors.

The importance of this study stems from the fact that they have used computational triaging, including solvation thermodynamics, in a prospective manner. Additionally, the study also highlights the importance of drug repurposing. It is worth noting here that pandemics like Covid19 for which medication does not exist have benefited from drug repurposing efforts [32]. Therefore, this kind of computational triage can inform a drug repurposing effort.

A lot of other examples are present in the literature [27, 33], where Watermap has been used to explain the activity differences in a congeneric series. These have established the importance of solvation thermodynamics considerations in drug design.

4.3.2 Grid Inhomogeneous Solvation Theory (GIST)

4.3.2.1 Objective and Approach

In a structure-based drug discovery (SBDD) campaign, it is crucial to know the precise location of water molecules in order to predict the next candidate ligand molecule with optimized binding properties. For that, it is necessary to characterize the water structure of all end-states during the ligand-binding reaction. The configuration space of solvent molecules is strongly coupled to the configuration of solute molecules. In this study [34], GIST was used for the construction of a rational structure-activity relationship (SAR) based solely on solvent contributions. These solvent contributions are rationalized by building different physically motivated models that use data from MD simulations and associated GIST calculations in order to predict SFEs. The functional form of all these models was based on the previously described displaced-solvent functional [22, 35]. A highly congeneric

series of 53 and 12 ligands were selected for binding to thrombin and trypsin, respectively. For all protein–ligand complexes, high-resolution crystal structures are available along with free energies measured by ITC or surface plasmon resonance (SPR) [36]. The ligands were sorted into matched pairs such that the affinity difference between ligands within a given pair can predominantly be attributed to a difference in solvation. The resulting 186 pairs for thrombin are used for further parameterization and testing of the solvent functionals.

MD simulations were carried out for the *apo* protein, protein–ligand complexes, and individual ligands in solution. Properties like the solvent energy, entropy, density, and entropy of protein–ligand association processes were calculated using the GIST method from these simulations. The solute atoms in each MD simulation were restrained to a reference structure.

In order to address the inaccuracies of the solvation calculations arising out of a fixed protein consideration, it was considered that conformational flexibility played a major role in the *apo* protein than in the more stabilized protein–ligand complexes. Thus, the *apo* structure was simulated in an unrestrained MD, and the trajectory was clustered into unique conformations. The cluster representatives were used as input for GIST calculations. For the protein–ligand complexes as well as the unbound ligand molecules, only fully restrained MD simulations were carried out, keeping the complex spatially fixed to the conformation found in the crystal structure.

In this work, three different basic solvent functionals, viz. F4, F5, and F6, for the ligand in solution (L/F4, L/F5 and L/F6), protein (P/F4, P/F5 and P/F6), and the protein–ligand complex (PL/F4, PL/F5 and PL/F6), have been used. These solvent functionals use the raw solvent entropy, energy, and density data from the GIST calculations and differ in the weighting parameters [22]. High-resolution X-ray structures of thrombin and trypsin were utilized to estimate the energy and entropy distributions in the binding site, and the high energy/high entropy regions in both enzymes were identified that would contribute to an enhanced binding energy of the ligand. The functionals were trained based on this experimental data, and they were then used to estimate the binding free energy, which was then compared with the experimental values.

4.3.2.2 Results and Discussion

The models based on the ligand alone in aqueous solution resulted in the best agreement with the experimental data. The solvent functionals L/F4, L/F5, and L/F6 gave the highest correlation ($r = 0.80$, 0.83 , and 0.79 , respectively) and performed better than similar functionals that were trained using shuffled data ($r = 0.20$, 0.25 , and 0.24). The solvent functionals based on GIST data from the protein binding pocket, P/F4 to P/F6, performed well ($r = 0.77$, 0.79 , and 0.77) but were worse than the ones based on the ligands alone. Almost similar performance was observed for the ones based on the protein binding pocket. In particular, the F5 basic solvent functional performed the best. The F5 basic functional led to a slightly increased mean unsigned error range compared to the other basic functionals F4 and F6. In the second part of this work, contributions from both the protein–ligand

complex and the ligand molecule were considered in the same calculation. The performance of functionals PL-L/F4, PL-L/F5, and PL-L/F6 ($r = 0.42, 0.61$, and 0.65) was considerably worse than the corresponding functionals based on the individual displacement treatments. Interestingly, the predictive power of these functionals increased ($r = 0.40, 0.85$, and 0.76) when considering grid voxels up to $3.5 \text{ \AA}$ away from the surface of the ligand instead of only using grid voxels up to $3.0 \text{ \AA}$. When an additional layer of grid voxels up to $4.0 \text{ \AA}$ is taken, no increase in performance was observed ($r = 0.37, 0.84$, and 0.73). For the other functionals, using only GIST data from the ligand molecule, no such performance increase is observed. This grid size variation is insightful as it captures the importance of the first hydration shell and possibly highlights the importance of considering all the waters that participate in it.

The work also compared their results with those obtained from MM-GBSA and 3D-RISM techniques, both in terms of computational efficiency and accuracy of predictions (correlation). While in terms of computational efficiency, the method was equally good, it outperformed the other two in achieving better correlation with experimental binding free energy data for this set of compounds against thrombin and trypsin.

GIST provides useful insights into the protein binding pockets in terms of the continuous distribution of the hydrophobicity and polarity, which offers excellent insights into high energy and high entropy regions. This information can help in both drug design as well as estimating the druggability of a certain binding site and a target. However, the results obtained from this method may depend on the extent of training of the functionals, which could be highly dependent on the quality of the structure under consideration. The functionals are not transferable across targets.

4.3.3 Three-Dimensional Reference Interaction-Site Model (3D-RISM)

4.3.3.1 Objective and Background

Coronaviruses initiate cell entry via the binding of the receptor binding domain (RBD) of the viral spike protein to the receptor protein, angiotensin-converting enzyme 2 (ACE2), on the surface of cells [37]. This process triggers infection and proliferation, making the binding of spike proteins to ACE2 the most significant initiator of the SARS-CoV-2 infection. Various studies on the binding of SARS-CoV-2 have been reported and have pointed out the importance of hydrogen bonding between amino acids on the RBD–ACE2 binding interface, as well as the importance of the hydration structure changes and hydrogen-bond bridging afforded by water molecules [38]. In this work, the binding process of the SARS-CoV-2 spike protein RBD to ACE2 was investigated using MD and the three-dimensional reference interaction-site model (3D-RISM) theory to highlight the effect of solvation in the binding process.

The trajectory data set DESRES-ANTON- [10 857 295,10 895 671], available on the D. E. Shaw research website [39], was obtained via the accelerated weighted ensemble MD simulation of the RBD–ACE2 complex formation process (PDB: 6VW1).

Taking one of the trajectory outputs as the initial structure, MD simulations were performed for 10 ns in each window. To this output, 3D-RISM theory was employed coupled with the Kovalenko Hirata closure [25] to evaluate the correlation functions and the SFE every 500 ps. The number of grid points in the 3DRISM-KH calculations was 512 with a spacing of 0.5 Å.

4.3.3.2 Results and Discussion

The distance change between the ACE2 and the RBD during the simulation was monitored, and it was found that the binding initially reduced the distance between the two proteins, and subsequently a structural rearrangement of the complex took place. The protein structure energy and the SFE exhibited dramatic changes through the binding process. It was observed that a gradual decrease in the conformational energy as the distance between the proteins decreased was accompanied by an increase in the SFE. The reason behind this is stated to be the penalty of free energy caused by desolvation. The complex is stabilized by 9 hydrogen bonds and resulted in a stabilization energy of about -400 kcal/mol. Obviously, this stabilization energy is far more than the contribution of the hydrogen bonds and confirms the presence of a variety of other interactions, including the van der Waal interactions and electrostatic interactions between the two.

The SFE, which gives insight into the process of desolvation, had large fluctuations through the binding process and was investigated by monitoring the solvent-accessible surface area (SASA) and the partial molar volume (PMV). While the SASA decreased during the process, the PMV increased. The binding process results in the exclusion of the solvent from the first hydration shell of the individual proteins and thus contributes to the reduction in SASA. Water molecules present in the interface of RBD and ACE2 were identified using the placement algorithm. Water molecules bound to ACE2/LYS31, GLU35, and LYS353 and RBD/TYR489 and GLN498 were observed. Principal component analysis of protein structural change revealed that the ACE2 protein underwent a conformational change upon binding, while the RBD conformation was less varied.

The enterprising use of the solvent functional energy in the study of the protein-protein binding process and the useful observations therein certainly expand the scope of solvation thermodynamics considerations in drug discovery.

The advent of machine learning, deep learning, and artificial intelligence has opened up new ways of analysis by identifying interesting patterns of molecular behavior influenced by the changes in structure and interactions. A recent study [40] extended the traditional 3D-RISM method to develop an AI-based model using the SFE data of ~ 4000 complexes estimated from this approach. The SFE predictions made from this model have resulted in a good correlation with experimental values.

Another study [41] applied deep learning to the output of MD simulations of thousands of protein structures and analyzed networks formed by the waters around the complexes. Their study affirmed that desolvation and water-mediated interactions are important. Additionally, the enthalpically favorable networks of first shell waters around the solvent-exposed portions of the ligands also play an important role in protein-ligand binding.

4.4 Conclusion

The developments in the theoretical framework for considering explicit water-based solvation thermodynamics with methods such as IST and 3D-RISM, coupled with giant leaps in computing power with high-performance clusters and graphics processing units (GPUs), have brought solvation thermodynamics into the realm of regular practice in drug design campaigns. While the methods discussed in the previous sections might have some shortcomings, it is undeniable that they bring immense value to the understanding of SAR in drug discovery programs and use the learnings in the prospective design of compounds. The solvation considerations impact not only the on-target and off-target potency of the compounds but also their disposition in terms of permeability and solubility. These are extremely important in the lead optimization phase of a drug discovery program, as they have a huge impact on driving the pharmacokinetic exposures needed for the compounds to show efficacy. As noted at the end of the last section, deep learning and machine learning techniques can augment and enhance the utility of physics-based methods in accurate prediction of solvation effects.

Thus, it can be visualized that solvation thermodynamics will soon mature into a mainstay of drug design processes, impacting the discovery process all the way from target identification to lead optimization.

References

- 1 Yu, Y., Wang, J., Shao, Q. et al. (2016). The effects of organic solvents on the folding pathway and associated thermodynamics of proteins: a microscopic view. *Sci. Rep.* 6: 19500. <https://doi.org/10.1038/srep19500>.
- 2 Lucent, D., Vishal, V., and Pande, V.S. (2007). Protein folding under confinement: a role for solvent. *PNAS* 104 (25): 10430–10434.
- 3 Yoshida, N. (2017). *J. Chem. Inf. Model.* 57 (11): 2646–2656. <https://doi.org/10.1021/acs.jcim.7b00389>.
- 4 Kyte, J. (2003). *Biophys. Chem.* 100: 193–203.
- 5 Kyte, J. and Doolittle, R.F. (1982). *J. Mol. Biol.* 157 (1): 105–132.
- 6 Anfinsen, C.B. (1973). *Science* 181: 4096.
- 7 Rose, G.D. (2021). *Biochemistry* 60 (49): 3753–3761.
- 8 Xing, D., Li, Y., Xia, Y.L. et al. (2016). Insights into protein–ligand interactions: mechanisms, models, and methods. *Int. J. Mol. Sci.* 17 (2): 144.
- 9 Homans, S.W. (2007). *Top. Curr. Chem.* 272: 51–82.
- 10 Li, J., Fu, A., and Zhang, L. (2019). An overview of scoring functions used for protein–ligand interactions in molecular docking. *Interdiscip. Sci. Comput. Life Sci.* 11: 320–328.
- 11 Wang, E., Sun, H., Wang, J. et al. (2019). End-point binding free energy calculation with MM/PBSA and MM/GBSA: strategies and applications in drug design. *Chem. Rev.* 119 (16): 9478–9508.

- 12 Miertuš, S., Scrocco, E., and Tomasi, J. (1981). Electrostatic interaction of a solute with a continuum. A direct utilization of AB initio molecular potentials for the prediction of solvent effects. *Chem. Phys.* 55 (1): 117–129.
- 13 Hou, T., Wang, J., Li, Y. et al. (2011). Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. *J. Chem. Inf. Model.* 51 (1): 69–82.
- 14 Alibrakshi, A. and Hartke, B. (2021). *Nat. Commun.* 12: 3584.
- 15 Aqvist, J. and Marelus, J. (2001). The linear interaction energy method for predicting ligand binding free energies. *Comb. Chem. High Throughput Screen* 4 (8): 613–626.
- 16 King, E., Aitchison, E., Li, H. et al. (2021). Recent developments in free energy calculations for drug discovery. *Front. Mol. Biosci.* 8: 712085.
- 17 Zhang, J., Zhang, H., Wu, T. et al. (2017). Comparison of implicit and explicit solvent models for the calculation of solvation free energy in organic solvents. *J. Chem. Theory Comput.* 13 (3): 1034–1043.
- 18 Nair, S., Kumar, S.R., Paidi, V.R. et al. (2020). Optimization of nicotinamides as potent and selective IRAK4 inhibitors with efficacy in a murine model of psoriasis. *ACS Med. Chem. Lett.* 11 (7): 1402–1409.
- 19 Nguyen, C.N., Young, T.K., and Gilson, M.K. (2012). Erratum: “Grid inhomogeneous solvation theory: hydration structure and thermodynamics of the miniature receptor cucurbit [7] uril”. *J. Chem. Phys.* 137 (14): 044101.
- 20 Karplus and Kuriyan (2005). *PNAS* 102 (19): 6679–6685.
- 21 Waibl, F., Kraml, J., Hoerschinger, V.J. et al. (2022). Grid inhomogeneous solvation theory for cross-solvation in rigid solvents. *J. Chem. Phys.* 156 (20): 204101.
- 22 Abel, R., Young, T., Farid, R. et al. (2008). Role of the active-site solvent in the thermodynamics of factor Xa ligand binding. *J. Am. Chem. Soc.* 130 (9): 2817–2831.
- 23 Li, Z. and Lazaridis, T. (2012). *Methods Mol. Biol.* 819: 393–404.
- 24 Haider, K., Cruz, A., Ramsey, S. et al. (2018). Solvation structure and thermodynamic mapping (SSTMap): an open-source, flexible package for the analysis of water in molecular dynamics trajectories. *J. Chem. Theory Comput.* 14 (1): 418–425.
- 25 Kovalenko, A. and Hirata, F. (1998). *Chem. Phys. Lett.* 290: 237–244.
- 26 Imai, T., Kovalenko, A., and Hirata, F. (2004). Solvation thermodynamics of protein studied by the 3D-RISM theory. *Chem. Phys. Lett.* 395 (1–3): 1–6.
- 27 Wang, L., Berne, B.J., and Friesner, R.A. (2011). Ligand binding to protein-binding pockets with wet and dry regions. *PNAS* 108 (4): 1326–1330.
- 28 Vadlamudi, R.K. and Kumar, R. (2003). *Cancer Metastasis Rev.* 22: 385–393.
- 29 Meng, J., Meng, Y., Hanna, A. et al. (2005). Abnormal long-lasting synaptic plasticity and cognition in mice lacking the mental retardation gene Pak3. *J. Neurosci.* 25 (28): 6641–6650.
- 30 Biswal, J., Jayaprakash, P., Rayala, S.K. et al. (2021). Watermap and molecular dynamic simulation-guided discovery of potential PAK1 inhibitors using repurposing approaches. *ACS Omega* 6 (41): 26829–26845.

- 31 Wishart, D.S., Feunang, Y.D., Guo, A.C. et al. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* 46 (D1): D1074–D1082.
- 32 Smith, D.P., Oechsle, O., Rawling, M.J. et al. (2021). Expert-augmented computational drug repurposing identified baricitinib as a treatment for COVID-19. *Front. Pharmacol.* 12: 709856.
- 33 Haider, K. and Huggins, D.J. (2013). *J. Chem. Inf. Model.* 53: 2571–2586.
- 34 Hufner-Wulsdorf, T. and Klebe, G. (2020). *J. Chem. Inf. Model* 60: 1409–1423.
- 35 Nguyen, C.N., Cruz, A., Gilson, M.K. et al. (2014). Thermodynamics of water in an enzyme active site: grid-based hydration analysis of coagulation factor Xa. *J. Chem. Theory Comput.* 10 (7): 2769–2780.
- 36 Sander, A., Hufner-Wulsdorf, T., Heine, A. et al. (2019). Strategies for late-stage optimization: Profiling thermodynamics by preorganization and salt bridge shielding. *J. Med. Chem.* 62 (21): 9753–9771.
- 37 V'kovski, P., Kratzel, A., Steiner, S. et al. (2021). Coronavirus biology and replication: implications for SARS-CoV-2. *Nat. Rev. Microbiol.* 19 (3): 155–170.
- 38 Kobryn, A.E., Maruyama, Y., Velazquez-Martinez, C.A. et al. (2021). Modeling the interaction of SARS-CoV-2 binding to the ACE2 receptor via molecular theory of solvation. *New. J. Chem.* 45 (34): 15448–15457.
- 39 http://www.deshawresearch.com/resources_sarscov2.html
- 40 Osaki, K., Ekimoto, T., Yamane, T. et al. (2022). 3D-RISM-AI: a machine learning approach to predict protein–ligand binding affinity using 3D-RISM. *J. Phys. Chem. B* 126 (33): 6148–6158.
- 41 Mahmoud, A.H., Masters, M.R., Yang, Y. et al. (2020). Elucidating the multiple roles of hydration for accurate protein-ligand binding prediction via deep learning. *Commun. Chem.* 3 (1): 19.

5

Site-Identification by Ligand Competitive Saturation as a Paradigm of Co-solvent MD Methods

Asuka A. Orr and Alexander D. MacKerell Jr

University of Maryland, School of Pharmacy, Department of Pharmaceutical Sciences, 20 Penn Street, Baltimore, MD 21201, USA

5.1 Introduction

Computer-aided drug design (CADD) is an essential component of the repertoire of tools used in drug discovery. Ligand-based drug design (LBDD) approaches, most notably quantitative structure-activity relationships, have made important contributions to CADD from the 60's onward [1–9]. Alternatively, more widely used are structure-based drug design (SBDD) approaches, which rely on the availability of the 3D structure of the target macromolecule. SBDD approaches play roles in the full range of the steps to which CADD contributes to the drug design process. This ranges from database screening methods [10–13] for hit identification to lead optimization, using various approaches such as free energy perturbation [14–16], and addressing pharmacokinetic issues in drug development including contributions to improve absorption, disposition, metabolism, excretion, and toxicological (ADMET) considerations [17–20]. Multiple recent reviews of various CADD methods have been presented [21–27].

CADD methods are inherently limited due to the use of mathematical models to treat the complexity of ligand-protein interactions, including the extremely challenging problem of accounting for contributions from the environment. Two major challenges that counterbalance each other exist; the accuracy of the underlying Hamiltonian or potential energy function/model chemistry used to estimate the system energy and the ability to sample the full range of conformations that contribute to ligand–target interactions, including the configurational space required to calculate entropic contributions. Accordingly, molecular docking methods commonly used for virtual screening of larger databases of molecules in the hit identification stage typically use simplified energy functions that significantly approximate the contribution of the environment to ligand binding [21]. More rigorous treatment of the aqueous environment using continuum models, such as the Poisson–Boltzmann (PB) [28–30] or Generalized Born (GB) [31, 32] approaches, offers improved treatment of the solvent contributions in conjunction with MD

simulations to predict binding affinities of ligands [33–36]. Free-energy perturbation (FEP) [37, 38] methods offer a formally rigorous approach to calculate the absolute or relative free energy of binding of ligands to a target in the presence of an explicit solvent [16, 39–43]. However, PB, GB, and FEP methods require extensive MD simulations to obtain the needed ensemble of conformations, such that empirical potential energy functions in conjunction with force fields, which are inherently limited in their accurate treatment of molecular interactions for the wide range of drug-like molecules, are typically applied [16, 34]. Quantum mechanical (QM) models in principle could overcome the limitation of potential energy functions, but they are computationally demanding as well as limited in their ability to treat long-range dispersion contributions to ligand–target interactions, limiting the wide-spread utilization of QM models in CADD to date [44–46].

Co-solvent MD methods can overcome the above limitations to various extents [47]. In the so-called co-solvent or mixed MD methods, which build upon experimental methods such as the multiple solvent crystal structures (MSCS) method [48] and early computational approaches such as Goodford’s GRID [49] and the multiple copy simultaneous search (MCSS) [50] methods, a small solute is included in an explicit-solvent simulation system along with the target macromolecule. MD simulations are then run, from which the probability distribution of the solute molecule around the target is calculated. The resulting probability distribution, which is based on true free energy as stated above, is then used to inform compound design. This precomputation of the ensemble of conformations [51] required to calculate the probability distribution comes with a computational cost, typically on the order of days depending on the approach and system size. However, once the probability distribution is available, it may be used rapidly in ligand design, overcoming the need to recalculate an ensemble for each drug-like molecule as required in PB, GB, and FEP methods and thereby circumventing the need for extensive computational resources. In addition, while co-solvent MD methods use empirical potential energy functions, as the simulations only include the target molecule, water, and small solutes, their accuracy may be considered to be acceptable as such molecules are generally the most carefully optimized aspects of a force field. This partially overcomes the need to have highly refined parameters for the wide range of chemical spaces present in drug-like molecules.

The remainder of this chapter presents various details of different co-solvent MD methods. These methods include a variety of approaches such as MDmix [52–55], MixMD [56–63], probe-based MD [64–66], and the works of Yang and Wang [67–70]. In addition to the details presented below, we refer the reader to the review by Ghan-toka and Carlson [58] on co-solvent MD methods. Details are then presented on the Site Identification by Ligand Competitive Saturation (SILCS) method developed in our laboratory and commercialized by SilcsBio LLC. This includes methodological details, including how they contrast with other co-solvent methods, and the various ways in which the SILCS methods may be used in the drug development process. Specific case studies are then presented to illustrate the SILCS approach.

Several co-solvent methods have been developed with probability distributions of the solute binding pattern used for various types of analysis. A summary of

Table 5.1 Co-solvent technologies and their applications in CADD.

Authors/method	Targeted applications
Privat et al. (Fragment dissolved MD)	Binding site identification [71].
Fabritiis et al.	Protein cryptic site prediction [72, 73].
Favia et al.	Cryptic ligand binding pockets prediction [74].
Zariquiey et al. (Cosolvent Analysis Toolkit)	Hotspot identification [75].
Yang et al.	Binding hotspots identification [67–70] and druggable protein conformations [76].
Bahar et al.	Determination of allosteric protein druggability [77]; pharmacophore modeling [78].
Barril et al. (MDmix)	Binding sites determination, predicting druggability [52], water displaceability and functional group mapping of protein binding sites [53], protein–ligand docking and binding affinity prediction [54], and pharmacophore modeling [55].
Carlson et al. (MixMD)	Hotspots mapping [57, 59], binding site identification [60, 63], kinetics of co-solvent on/off rates and identification of allosteric binding sites [62], free energies and entropies of binding sites [61], and displaceable water site identification [60].
Gorfe et al. (pMD)	Active and allosteric binding sites identification in proteins [64] including membrane-bound drug targets [64, 65].
Caflisch et al.	Co-solvent on-off rates and binding affinities [79] and water displaceability by co-solvents [80].
Tan and Verma (LMMD)	Hydrophobic and halogen binding site identification [81–83] and use of MD to map multiple ligand binding sites [84].
Lill et al.	Rank-ordering of ligand binding and identification of binding locations of functional groups [85, 86].
Yanagisawa et al. (EXPROPER)	Evaluation of hotspots and binding sites on proteins [87].
Takemura et al. (ColDock)	Prediction of protein–ligand docking structures [88].

published co-solvent methods and their applications is presented in Table 5.1. Co-solvent methods are typically based on MD simulations of a single solute molecule in aqueous solution that encompasses the entire target macromolecule, which is a protein in the majority of cases (Table 5.2). The solutes, or co-solvents, that are used are typically of a molecular weight of 150 daltons or less. The small size of the solute is important as it allows for adequate diffusion of the solute to sample the full region around the protein within the time scale of the MD simulations. Such sampling is further facilitated by the use of multiple solutes in the simulation systems (Table 5.2). In addition, enhanced sampling techniques, including accelerated MD [93] and lambda dynamics [86], have been implemented in co-solvent

methods to facilitate solute sampling. However, the majority of efforts use standard MD simulations, typically spanning 20 to 200 ns, with multiple replicas performed to facilitate convergence. The importance of obtaining adequate convergence of the solute distributions cannot be understated. This is often assessed by comparing the calculated probability distributions of different portions of the MD simulations. This could be the first or second half of an extended MD simulation or, when multiple simulations are performed, calculating two sets of probability distributions from sets of simulations. Notably, the majority of co-solvent MD simulations can be performed with standard MD simulation packages such as AMBER [94], OpenMM [95], GROMACS [96], NAMD [97], and CHARMM [98], with available tools such as CPPTRAJ [99] used for calculation of the probability distributions and other analyses.

The solutes included in co-solvent simulations range from charged molecules, including acetate and methylammonium, to neutral polar molecules and to apolar molecules (Table 5.2). Polar neutral molecules include methanol, ethanol, acetamide, isopropylamine, acetic acid, isopropanol, and acetonitrile. Isopropanol was used in some early studies as it contains both polar alcohol and apolar aliphatic carbons. Apolar molecules include benzene, isobutane, and propane, as well as heterocycles such as pyridine and imidazole. In general, the majority of studies perform individual sets of simulations with a single solute molecule at various concentrations, with the probability distributions from the different sets of simulations combined for the various types of analysis listed in Table 5.1.

A critical consideration in co-solvent simulations is the concentration of the solutes included in the explicit aqueous solution. Typically, solute molecules are included in concentrations ranging from 0.1 to 1.5 M, though some workers use up to 3 and 12 M (Table 5.2). High concentrations of solute molecules may result in the denaturation of the target macromolecule [100]; however, low concentrations of solute molecules may result in slow convergence in the sampling of solute distribution around the full 3D space of the macromolecule. The undesired artifacts resulting from high concentrations of solute molecules can be circumvented through restraints to the target macromolecule. In such cases, restraints to the macromolecule can be balanced to maintain the structural integrity of the macromolecule while simultaneously allowing sufficient flexibility for potential binding sites to open. An additional concern potentially exasperated by the choice of solute concentration includes the potential for aggregation of hydrophobic species, ion pairing between solutes, and proper convergence of the probability distributions. Apolar solutes can aggregate with themselves when used in co-solvent simulations. This issue was specifically addressed by Guvench and coworkers with respect to the potential for solutes to cause protein denaturation during co-solvent simulations [100]. To avoid hydrophobic aggregation as well as ion pairing, repulsive potentials may be introduced between selected solutes. This allows for an effective “ideal solution” behavior, thereby facilitating the sampling of the solutes in the full 3D space of the protein. However, such repulsive potentials also limit the sampling of, for example, two benzene solutes directly adjacent to each other in a binding pocket, as previously discussed [84].

Table 5.2 Overview of solutes used and their concentrations in co-solvent technologies.

Authors/method	Solutes investigated	Solute concentration
Privat et al. (Fragment dissolved MD)	Single solute per simulated system [71]. Investigated solute ligands include ethyl 3-amino-4-methylbenzoate, 2-methyl-2-(4-morpholinyl)-1-butanamin, (3r)-piperidin-3-yl(piperidin-1-yl)methanone, n-methyl-1-(1-methyl-1h-imidazol-2-yl) methanamine, 5-hydroxy-2-aminobenzimidazole, 2-aminopyrimidine, 1-aminoisoquinoline, 3-chloro-1-benzothiophene-2-carboxylate, 3-(4-chloro-3,5-dimethylphenoxy)propanoic acid, dimethyl-sulfoxide-methyl-(methylsulfinyl) methyl sulfide, and 6-azaniumylhexanoate in water [71].	Solute concentration dependent on size of solute molecule [71]: ~0.01 to ~0.09 M
Fabritiis et al.	Single solute per simulated system [72, 73]. Investigated 129 solute fragments [72, 73].	Single solute molecule per protein system [72, 73].
Favia et al.	Single solute per simulated system [74]. Investigated solutes include acetic acid, isopropanol, and resorcinol in water [74].	5% (m/m) [74]: ~3 M.
Zariquiey et al. (Cosolvent Analysis Toolkit)	Single solute per simulated system [75]. Investigated solutes include acetamide, benzene, acetanilide, imidazole, and isopropanol in water [75].	10% (m/m) [75]: ~6 M.
Yang et al.	Single solute per simulated system [68, 69, 76]. Investigated solutes include isopropanol in water [68, 69, 76].	20% (v/v) [68, 69, 76]: ~3.5 M
Bahar et al.	Single solute per simulated system [77] or multiple solutes per simulated system [78]. Investigated solutes include isopropanol, acetamide, imidazole, acetate, isopropylamine, and isobutane in water [77, 78].	20 : 1 water: solute ratio [77, 78]: ~3 M.
Barril et al. (MDmix)	Single solute per simulated system [52–55]. Investigated solutes include isopropanol, ethanol, acetamide, methylammonium, acetate, and acetonitrile in water [52–55].	20% (v/v) [52–55]: ~3 to 5 M.
Carlson et al. (MixMD)	Single solute per simulated system [57, 59, 60, 63] or multiple solutes per simulated system [60–63]. Investigated solutes include isopropanol, acetonitrile, pyrimidine, imidazole, n-methylacetamide, methylammonium, and acetate in water [57, 59–63].	50% (w/w) [57, 59]: ~8.5 to ~12 M 2.5% (v/v) [61, 62]: ~0.4 M 5% (v/v) [60–63]: ~0.6 M to ~1.5 M

(continued)

Table 5.2 (Continued)

Authors/method	Solutes investigated	Solute concentration
Gorfe et al. (pMD)	Single solute per simulated system [64] or multiple solutes per simulated system [65]. Investigated solutes include isopropanol, isobutane, acetamide, acetate, urea, dimethylsulfoxide, and acetone in water [64, 65].	~20 : 1 water : solute ratio [64]: ~3 M. 140 : 1 water : solute ratio [65]: ~0.4 M.
Caflisch et al.	Single solute per simulated system [79, 80]. Investigated solutes include 4-hydroxy-2-butanone, dimethylsulfoxide, 5-diethylamino-2-pentanone, methyl sulphinyl-methyl sulfoxide, 5-hydroxy-2-pentanone, tetrahydrothiophene 1-oxide, methanol, and ethanol in water [79, 80].	Single solute molecule per protein system [79]. 0.44 M [80]
Tan and Verma (LMMD)	Single solute per simulated system [82, 83] or multiple solutes per simulated system [84]. Investigated solutes include benzene, chlorobenzene, methanol, acetaldehyde, methylammonium, and acetate in water [82–84].	0.1 to 0.4 M [82–84]
Lill et al.	Multiple solutes per simulated system [85, 86]. Investigated solutes include propane, formamide, acetaldehyde, benzene, fluoro-benzene, chloro-benzene, bromo-benzene, and iodo-benzene in water [85, 86].	0.25 M [85, 86]
Yanagisawa et al. (EXPROPER)	Single solute per simulated system [87]. Investigated solutes include 138 molecules [87].	0.25 M [87]
Takemura et al. (ColDock)	Single solute per simulated system [88]. Investigated solutes include dimethylsulfoxide, methylsulfinyl-methylsulfoxinide, ϵ -aminocaproic acid, 4-(4-bromo-1H-pyrazol-1-yl) piperidinium, transaminomethyl-cyclohexanoic acid, and FK506 (Tacrolimus) [88].	~0.065 to ~0.15 M [88]
MacKerell et al. (SILCS)	Multiple solutes per simulated system [89–92]. Investigated solutes include benzene, propane, methanol, imidazole, acetaldehyde, formamide, methylammonium, acetate, dimethylether, fluoroethane, trifluoroethane carbon, fluorobenzene fluorine, chloroethane, chlorobenzene, and bromobenzene in water [89–92].	1 M ^{a)} [89, 90] 0.25 M [90–92]

a) The ~1 M concentrations were used in SILCS simulations containing only benzene and propane as co-solvents in early studies. All SILCS simulations thenceforth generally consist of 8 co-solvent molecules, each at ~0.25 M.

A strength of the co-solvent MD methods is their ability to convert the probability distributions into free energies. As the probability distributions are based on occupancies on a 3D grid, the conversion yields a 3D free energy grid associated with the binding of a solute molecule to the protein. In this procedure, as described by Alvarez and Barril and also performed early on by Bahar and coworkers [53, 77], the number of counts/MD snapshot in each 3D voxel is normalized by the expected number of counts based on the concentration of the solute in the simulation system. They then performed a standard state correction associated with a 1 M concentration. In that study, to estimate the binding free energy of individual hotspots, the free energies were calculated over all voxels within 2 Å of the hotspot, and then the Boltzmann weighted average over those voxels was calculated. It should be noted that defining concentration in a finite simulation system is not unique, given the inaccessible volume occupied by the target macromolecule. For example, calculating the concentration of solutes based on the total simulation volume is typically not appropriate as a significant portion of the system is not accessible to solutes or water. Accordingly, it is necessary to calculate solute concentrations relative to the amount of water in the system [53, 101]. An exception to this occurs with the application of the SILCS technology for the calculation of functional group-free energy distributions in bilayers, where the use of the total simulation system is appropriate [102].

Beyond the free energy functional group affinity information in the form of 3D maps generated from co-solvent MD methods, there is the comprehensive nature of that information for the entire 3D space of the protein or other target molecule being studied. This allows for a variety of different types of analyses to be performed based on affinity information, with those analyses being able to be performed rapidly given the pre-computed nature of the 3D maps. Listed in Table 5.1 are various types of analyses performed to date. The widest use of the technology has been for the identification of novel binding or hotspots that offer the potential to be novel allosteric sites for the discovery of novel positive and negative allosteric modulators. However, the successful identification of novel sites is impacted by the treatment of protein conformational flexibility in the MD simulations. The use of rigid protein structures can limit the extent to which putative binding sites not present in crystal structures may open, while the total lack of restraints may potentially lead to denaturation events, especially when high solute concentrations are used [100].

Once binding sites are known, co-solvent methods can be used for various aspects of ligand discovery and design. They have been used for the determination of pharmacophore models, which may subsequently be used for large-scale in silico database screening. Additionally, such methods can also be used to identify displaceable water sites and determine free energies and entropies of binding sites, facilitating drug design. An interesting variation of co-solvent methods has been the extraction of kinetics associated with the on- and off-rates of the co-solvents [79]. While the many variations of the co-solvent MD technology have seen wide use, it is safe to say that their full potential has yet to be achieved. To highlight this, in the remainder of the chapter, we will provide an overview of the various physical phenomena to which the SILCS technology has been implemented, along with case studies to offer explicit examples of those approaches.

5.2 SILCS: Site Identification by Ligand Competitive Saturation

SILCS was originally developed in the year 2008 by our lab through discussions involving Olgun Guvench and Alex MacKerell. The method was first implemented in the program CHARMM [98], and associated workflows were written to extract and calculate functional group probability distributions, which could readily be visualized using VMD [103], Pymol [104], or any visualization package that accepts grid density maps, such as those from X-ray crystallography. SILCS and other co-solvent approaches are simple: simulate the distribution of solutes presenting specific functional groups around a protein or other macromolecule to identify where such groups would, or would not, interact. Barril and coworkers published the approach on a single solute on several protein targets at approximately the same time as the initial SILCS publication in 2009 that examined the BCL-6 oncoprotein [52, 89]. Notably, SILCS included multiple solutes, initially propane and benzene [89], while the other earlier efforts included only a single solute [52, 56, 57, 77, 82]. This difference represented a unique capability that, along with the use of a repulsive potential between solutes in the simulation system, allowed for the technology to receive a US patent in 2018 (US Patent Number 10,002,228). Additionally, the SILCS technology has been commercialized in the form of SilcsBio LLC, while the development of the technology continues in the MacKerell laboratory. This combination has allowed the SILCS technology to be applied to a range of problems, as listed in Table 5.3 and described below.

SILCS, as with the other co-solvent MD approaches [52, 56], was motivated by the need to identify the location of different types of functional groups on the full surface of a protein. In contrast to other methods, we wanted to map multiple types of functional groups from a single set of simulations through “competitive saturation” between the different solutes representing the functional groups and with water. This contrasts with other methods that initially only included a single solute competing with water [52, 56, 57, 77, 82]. The original embodiment simply included benzene, propane, and water, representing aromatic, aliphatic, and hydrogen bond donors and acceptors, respectively [89]. The analysis focused on the probability distributions of the solute and water functional groups and how they recapitulated the known location of peptides binding to the BCL-6 protein lateral groove [89]. However, given the known importance of specific treatment of water versus other hydrogen bond donors and acceptors, the collection of solutes was expanded to a total of 8 solute molecules, each at a concentration of ~ 0.25 M, in work by Raman et al. [90, 91]. These 8 solute molecules included benzene, propane, methanol, imidazole, acetaldehyde, formamide, methylammonium, and acetate, with acetaldehyde subsequently replaced with dimethylether given the presence of an aldehyde moiety on formamide and the role of ether oxygens in drug-like molecules [92]. In addition, also motivated by their presence in drug-like molecules, was the implementation of a collection of halogen-containing solutes, termed SILCS-X [106], as performed by other groups [82, 86]. In SILCS, these halogen-containing solutes include fluoroethane, trifluoroethane, fluorobenzene,

Table 5.3 SILCS methods and tools.

SILCS Method	Description
SILCS GCMC/MD Simulations	Combined oscillating excess chemical potential Grand Canonical Monte Carlo/Molecular Dynamics Simulations [105] for generation of SILCS FragMaps using either the standard [90, 91] or halogen (SILCS-X) [92] solutes [90, 91].
SILCS-MC Docking	Monte Carlo docking of ligands into the SILCS FragMaps [92, 106].
SILCS-MC Pose Refinement	Local Monte Carlo docking of ligands from known orientations into the SILCS FragMaps.
SILCS-Hotspots	Identification of fragment binding sites via comprehensive SILCS-MC docking on entire 3D space of target macromolecule [107]. Subsequent SILCS-MC docking of a subset of FDA-approved compounds to identify binding sites for drug-like molecules including allosteric sites.
SILCS-Pharmacophore	Generation of target-based pharmacophore features for large-scale database screening [108, 109].
SILCS-PPI	Evaluation of protein–protein interaction distribution [110].
SILCS-RNA	Optimized SILCS simulation approach for RNA [111].
SILCS-Biologics	Rational selection of excipients for biologic (e.g. mAbs) formulation [112, 113].
SILCS-BML	Optimization of SILCS FragMaps targeting experimental binding affinities using a Bayesian Markov Chain Monte Carlo Machine Learning approach [106].

chloroethane, chlorobenzene, and bromobenzene [106]. Thus, the SILCS method, through two sets of GCMC/MD simulations, can generate a comprehensive set of functional group affinity patterns representing the majority of those commonly seen in drug-like molecules [92].

While the initial motivation for the SILCS and the other cosolvent approaches was primarily site identification, the power of the approach is more fully utilized when the probability distributions from the simulations are normalized and converted to free energies as described above. This approach is also applied in the SILCS approach with additional normalization performed for the number of functional group atoms in each solute [106], yielding grid free energy (GFE) FragMaps. The GFE FragMaps represent the binding free energies of each functional group on the 3D grid relative to being in an aqueous solution, where the GFE values equal zero. Thus, simply overlaying a ligand on the collection of GFE FragMaps and assigning atoms in the ligand to functional group types allows a GFE score to be assigned to the classified atoms, which may then be summed to yield the ligand grid free energy (LGFE). LGFE scores are an approximation of the experimental ligand binding affinities as discussed by Ustach et al. [106] Given that the SILCS FragMaps are precomputed, calculation of the LGFE score is virtually instantaneous; in the context

of the SILCS-MC approach, MC sampling of translational, rotational, and dihedral rotational degrees of freedom of the ligands may be readily performed in minutes for ligand docking in a given binding site that includes a near exhaustive search of the binding site. Alternatively, if the bound orientation of a ligand or its parent is known, pose-refinement SILCS-MC may be performed from which the LGFE score of the ligand is obtained. This approach has been performed in a number of studies [12, 92, 106, 114–116], and in a comprehensive comparison, the use of SILCS-MC exhaustive docking has been shown to be comparable to highly computationally demanding FEP calculations for the prediction of relative affinities of ligand binding, with the SILCS-MC calculation being hundreds of times faster than the FEP approach [92]. Notably, the SILCS-MC docking approach does not require a known bound starting orientation of the ligand, as required by FEP. In addition, when experimental data is available on even a small set of the ligands (e.g. down to 10), the “best” SILCS-MC scoring regimen may be selected to improve the predictability of the LGFE scores. Moreover, the application of a Bayesian Markov Chain Monte Carlo Machine Learning approach (BML) [106] allows for optimization of the contribution of the different types of FragMaps to the LGFE scores, yielding predictive models that are systematically better than those from FEP methods [92].

A central theme in a medicinal chemistry campaign is developing an accurate and detailed structure-activity relationship (SAR) to lead the ligand design process. With SILCS, the use of the atomic GFE scores to yield the LGFE scores allows for the contribution of individual atoms as well as different chemical moieties to the binding affinity to be determined. This is powerful as, for example, the addition of a chemical moiety to a molecule may lead to a relatively small change in the overall binding affinity due to energetic gains of the added moiety being offset by less favorable contributions from other parts of the molecule. This was shown to occur in a series of allosteric inhibitors of heme oxygenase [117]. The use of GFE contributions has facilitated ligand design in a number of studies targeting proteins such as Bcl-6 [118], Mcl-1, and Bcl-xl [119].

A key development in the evolution of the SILCS technology was the implementation of the oscillating excess chemical potential, μ_{ex} , GCMC approach [105]. This was motivated by the need for adequate solute sampling in deep and fully occluded, cryptic binding pockets, such as those commonly seen in GPCRs and in nuclear receptors, respectively, in a thermodynamically correct fashion. As GCMC allows for the insertion and deletion, as well as rotations and translation, of solutes from an external bath into the simulation, system a proper equilibrium of the solutes and water in deep and occluded binding pockets may be achieved. From this, GFE and LGFEs can be obtained for these types of binding sites as compared to co-solvent MD methods, where the diffusion times and/or the need for the protein to partially unfold to adequately sample such sites are prohibitive. However, the use of standard GCMC, where μ_{ex} for water and the solutes are set to the experimental hydration free energy leads to very low acceptance rates for insertions and deletions, thereby requiring prohibitive amounts of GCMC sampling. The use of an oscillating μ_{ex} , where μ_{ex} is varied based on the target concentration of each solute or water, effectively acting as an umbrella potential in chemical potential space rather than conformational

space, overcomes this limitation, yielding adequate acceptance rates. The approach was initially implemented and shown to yield accurate relative binding affinities of ligands to the T4 lysozyme pocket mutant, which contains a totally occluded binding pocket [105]. The method was subsequently applied to nuclear receptors and the β_2 adrenergic receptor, a GPCR [120], and is now the standard in SILCS simulations in conjunction with the MD portion of the method, which is needed to further facilitate both solute and water sampling as well as conformational sampling of the protein or other target macromolecule. Recent efforts have ported the oscillating μ_{ex} GCMC method to GPUs, yielding significant speed enhancements, especially in larger simulation systems [121].

In the context of drug design and development, the SILCS technology represents an end-to-end resource. Qualitatively, visualization of the SILCS FragMaps may be used to facilitate the identification of possible binding sites, including cryptic and allosteric sites [117], facilitate decisions on what types of scaffolds can occupy a site, and then help the medicinal chemist determine the types of functional groups that may be added to a lead compound to improve the binding affinity while simultaneously considering synthetic accessibility. Quantitatively, the various applications are diverse. In the absence of known binding sites on a target macromolecule, the SILCS-Hotspots approach is of utility [107]. In SILCS-Hotspots, a library of fragment molecules common to drug-like compounds [122, 123] is comprehensively docked in the full 3D space of the target macromolecule and then subjected to 2 rounds of clustering to identify and rank putative fragment binding sites. This goes beyond simply using the solute binding locations typically performed in co-solvent methods, as the fragments are of a larger MW and more chemically diverse. Additionally, while other methods often assess the hotspots on macromolecules based on rigid fragments, during the SILCS-Hotspots docking, fragment conformation, as well as orientation are sampled. Due to the inclusion of macromolecule flexibility in the computation of SILCS FragMaps, SILCS-Hotspots are able to explore and potentially identify cryptic pockets [107]. The number of identified hotspots throughout the macromolecule with low LGFE scores can also indicate the propensity of the macromolecule to bind several classes of ligands at different sites. Once fragment binding sites are identified, the identification of binding sites for larger drug-like molecules can be performed through the identification of two or more adjacent hotspots followed by SILCS-MC docking of FDA-approved compounds into those sites. The average LGFE scores of the top 20 or 25 compounds are then obtained along with the relative solvent accessibility (rSASA) of those compounds in the presence and absence of the target macromolecule with putative binding sites typically having average LGFE scores <-10 kcal/mol and rSASA values $>60\%$, where 100% indicates full exclusion of the ligands from the solvent by the protein. The SILCS-Hotspots approach has been used for the identification of cryptic, allosteric sites on β -Glucosidase A [124] and on the β_2 adrenergic receptor, a GPCR [125, 126], and the identification of a site to block protein-protein interactions on the Ski8 complex [127].

When binding sites are known or have been identified as described in the preceding paragraph, SILCS offers effective tools for large-scale database screening. Utilizing the SILCS FragMaps, target-based pharmacophores may be generated

[108, 109]. The method typically generates 10 to 12 pharmacophore features for a binding site from which multiple 4 or 5-feature pharmacophore hypotheses are generated. These typically contain 2 to 3 apolar/hydrophobic features along with various polar or charged features. The pharmacophore hypotheses are used to screen pre-prepared virtual databases that contain multiple conformations and classified functional groups using programs such as Pharmer [128]. From this process up to 10 000 compounds for each hypothesis are selected and then subjected to SILCS-MC pose refinement based on the pharmacophore-docked orientations. The resulting LGFE scores are then used to rank the compounds with the top 1000 to 2000 selected. Bioavailability may then be predicted using various metrics, such as the 4D Bioavailability (4DBA) metric, which takes into account the terms in Lipinski's rule of 5 in a single scalar number [129]. The compounds may then be clustered based on chemical fingerprints from which final compounds from individual clusters may be selected for purchase and experimental assay. Once active hit compounds are identified, compounds similar to the hit compounds may be obtained from a virtual database and experimentally assayed to determine true lead compounds as well as initial SAR data to jump-start a ligand design effort [130]. A similar approach has been successfully applied to identify both agonists and antagonists of a number of proteins, including the $\beta 2$ adrenergic receptor [120], the A18 ribonucleotide binding protein [131], and the WDR61 protein in the Ski8 complex [127]. In the $\beta 2$ adrenergic receptor study, FragMaps for both the active and inactive forms of the GPCR were used to specifically perform database searching for agonists; 7 out of 15 compounds were shown to be active agonists [120].

When active compounds are known, the SILCS FragMaps may be used for ligand optimization. It should be emphasized that the same FragMaps, in some cases supplemented with the SILCS-X halogen FragMaps, used for binding site identification along with pharmacophore development and database screening are also used for optimization efforts, showing the utility and extreme computational efficiency of the SILCS method. If an experimental structure of the ligand-protein complex is available, then the experimental structure may be used in the ligand optimization, with the experimental structure overlaid on the FragMaps and subjected to local SILCS-MC pose refinement. Visual analysis of the overlap of the various ligand moieties and the FragMaps, as well as the GFE contributions of the individual atoms on the ligand, may be used to identify which regions of the molecule are contributing to binding versus acting as scaffolding elements. Such analysis, especially visualizing the FragMaps adjacent to the ligand, can be used to generate ideas concerning the types of chemical modifications to make for improved affinity. In many cases, multiple FragMap types occupy the same region of 3D space, information that can be used to modify ring systems in the context of scaffold hopping. For example, if there is a phenyl ring occupying an apolar or aromatic FragMap and there is also an H-bond acceptor FragMap in that region, then one may consider modifying the ring into a heterocycle such as a pyridine. As stated above, this approach in general is of direct utility to medicinal chemists, as they can readily visualize modifications that will improve experimental activity while simultaneously accounting for synthetic accessibility.

Once possible potential modifications or sites on which functional groups may be substituted are identified, the SILCS-MC approach may then be used to obtain LGFE scores for the modified species, allowing them to be prioritized for synthesis and experimental assay. As stated above, the SILCS-MC pose refinement can be applied if the orientation of the lead compounds is known and only local relaxation of the compounds is desired, as with FEP calculations. However, full SILCS-MC docking may also be performed, allowing for reorientation of the ligand in the binding site, a phenomenon that often occurs as seen with inhibitors of the Mcl-1 protein [106]. Again, full SILCS-MC docking has been shown to be competitive with FEP methods but significantly more computationally efficient. Moreover, the BML approach may be used to optimize the FragMaps to improve their predictability as a medicinal chemistry campaign proceeds with additional BML optimizations as additional experimental data becomes available. This allows for continual improvements in the predictability of SILCS as the program proceeds, including the ability to do individual BML optimizations for specific chemical scaffolds. Numerous studies have used SILCS FragMaps for ligand optimization for diverse proteins such as Mcl-1, Bcl-xl [119], Bcl-6, Erk [132], mGluR5 [133, 134], Heme oxygenase [117], and, interestingly, the ribosome [135].

In the context of drug-like molecule ligand design and development, the SILCS method has been recently extended to RNA [111]. Given the polyanionic nature of RNA, this required special approaches for the SILCS simulations, leading to the need to perform separate SILCS simulations for the apolar and polar neutral solutes and for the charged solutes as methylammonium-dominated sampling of the RNA molecule. Using the modified protocol, it was shown that the GFE FragMaps recapitulated the functional groups of a number of RNA-binding ligands on multiple targets as identified in crystallographic studies. In addition, using the approach, new binding sites were identified in combination with SILCS-Hotspots. However, as discussed in the study, the significant impact of ligand binding on RNA conformation combined with the minimal amount of quantitative data on ligand–RNA interactions make RNA a challenging target for computational methods. Nevertheless, SILCS-RNA appears to offer information content to meet this challenge beyond that accessible to many of the other available methods.

Other areas in which the SILCS method has been shown to be useful are in the area of drug liability through the development of models to predict the hERG blockade [136, 137], where the approach can identify portions of the ligands that make the largest contribution to hERG binding. Motivated by the ability to apply SILCS to membrane-bound GPCRs, the method was applied to lipid bilayers and shown to be capable of mapping functional group free energy profiles across the membranes [102]. The approach was able to yield good agreement between bilayer/water and experimental octanol/water partition coefficients, predict resistances of the solutes in the bilayers, and calculate free energy profiles for drug-like molecules across the bilayers (PAMPA and POPC/cholesterol) in a computationally accessible manner [102]. The latter capability is currently being used to develop models for the prediction of membrane permeabilities of drug-like molecules in conjunction with deep neural nets [138]. Interestingly, the SILCS calculated free energy profiles

of molecules across the bilayers represent absolute free energy profiles due to the use of GFE normalization based on the full volume of the simulation system, as the solutes and water are fully accessible to all regions of the simulation boxes.

Beyond small molecule-focused drug design, the SILCS technology has utility for studies of macromolecules themselves. The initial step in applying SILCS to proteins alone was its extension to calculate protein–protein interactions (PPI) [110]. SILCS-PPI uses a fast-Fourier transform (FFT) sampling approach in conjunction with the overlap of the SILCS FragMaps “receptor” protein with the distribution of functional groups on the “ligand” protein to score the orientations from which distribution of PPI orientations are obtained. The technology is competitive with available computational PPI methods, though the requirement to calculate the FragMaps makes it computationally demanding when only PPI analysis is needed. However, SILCS-PPI may be combined with SILCS-Hotspots to facilitate the formulation of biologics, including monoclonal antibodies (mAb). SILCS-Biologics [112, 113] combines the distribution of the probability of residues participating in PPI on the entire protein surface with the distribution of excipients, buffers, and monoions on the surface of the protein. This combination allows for excipients that may block PPI that contribute to aggregation or increased viscosity to be analyzed. In addition, information on excipients that may impact protein stability can be obtained. This combination of information may be used to facilitate the selection of excipients, especially in formulations that require a high protein concentration. In addition, analysis of the distribution of excipients, buffers, and monoions bound to the protein may be used to estimate the total effective charge and dipole of the protein [139], providing additional information of utility in biologics formulation.

The application of SILCS to proteins, including glycoproteins, has elucidated macromolecular interactions involved in immune function and downstream signaling. In one study, the interactions of the protein endoglycosidase S2 (EndoS2), a protein excreted by *Streptococcus pyogenes*, which deglycosylates the Fc of mAbs thereby limiting the host immune response, were investigated [140]. In the study, Fc glycans were docked to the carbohydrate binding module (CBM) and the glycoside hydrolase (GH) domains of EndoS2 using SILCS-MC, following which the remainder of the Fc was built from an ensemble of Fc-glycan conformations generated from extensive MD simulations. Following this docking and reconstruction procedure, MD simulations of selected docked complexes were performed from which details of the interaction of EndoS2 with the Fc were predicted. Notably, the study showed the importance of PPI between the Fc and EndoS2 rather than just interactions involving the glycan alone, an observation that was shown to be in agreement with subsequent experimental cryo-EM and crystallographic studies [141, 142]. In a second study, the role of clustering of the Fc γ RIIIa-Fc ϵ RI γ receptors upon multivalent binding of antibodies on phosphorylation events by the kinase LCK was investigated [143]. In the study, models of the transmembrane (TM) and intracellular (IC) regions of the Fc γ RIIIa-Fc ϵ RI γ complex in different spatial relationships mimicking different extents of clustering were investigated via MD simulations and SILCS-MC docking. Multiple long-time MD simulations of the complexes under the different extents of clustering were performed to generate large

ensembles of conformations of the receptors. SILCS-MC docking of the Tyr-based activation motifs (ITAMs) of the IC regions of the receptor onto LCK was then performed, following which the full IC, TM, and membrane bilayer-LCK complexes were reconstructed. From the resulting ensemble of complexes, information on how LCK can perform multiple phosphorylations on individual ITAMS and on ITAMS in different Fc γ RIIIa-Fc ϵ RI γ complexes was obtained. Among these conformations, a number to which the kinase Syk could bind as required for downstream signaling were identified. This approach nicely showed how downstream signaling may be facilitated by “concentration” effects where ITAMs coming close to each other due to receptor clustering can occur. In combination, both studies highlight how large ensembles of conformations of flexible biomolecules such as glycans and disordered peptides can be generated and then docked to proteins using SILCS, from which ensembles of putative interactions responsible for biological events may be identified.

A key contributor to the success of SILCS technology is the use of the CHARMM General Force Field (CGenFF) [144–147] and the CGenFF program [145, 146]. CGenFF provides coverage for a wide range of chemical groups within biomolecules and drug-like molecules. The CGenFF program included with the SILCS software package (SilcsBio LLC) rapidly generates FF parameters for a ligand of interest according to analogy to known small molecule CGenFF parameters. These topologies and parameters are generated within fractions of a second, enabling the analysis of a multitude of ligands. In the context of SILCS simulations, CGenFF and the CGenFF program provide the empirical force field of the standard SILCS and SILCS-X solute molecules. In SILCS-X and for halogen-containing ligands, CGenFF represents the σ -hole as a positive point charge, a “lone pair,” to improve the treatment of halogen bonding [148]. In conjunction with the CHARMM36 FF [149–157] and CHARMM TIP3P models [157, 158] describing the target macromolecule and surrounding water, CGenFF allows for highly accurate sampling of the solute-target interactions within the SILCS simulations. In the context of SILCS analyses performed after simulations, CGenFF is crucial for assigning FragMap types to solutes and ligands of interest based on an atom classification scheme for LGFE scoring. Additionally, in SILCS-MC, the detection of rotatable bonds is based on the topology of the ligand generated by the CGenFF program, and the intramolecular energy of the ligand is calculated using the CGenFF potential energy function.

5.3 SILCS Case Studies: Bovine Serum Albumin and Pembrolizumab

To illustrate the application of SILCS, here we provide two case studies of the SILCS method applied to bovine serum albumin (BSA) and Pembrolizumab, a humanized antibody used in cancer immunotherapy sold under the brand name Keytruda. In summary, given an initial structure of the target macromolecule, SILCS simulations are performed to sample the interaction pattern of solute molecules with

the target macromolecule, in this case, BSA and pembrolizumab, and ultimately calculate pre-computed GFE FragMaps for use in a wide range of analyses. These analyses include SILCS-MC to sample spatial and conformational sampling of ligands, SILCS-Hotspots to identify allosteric sites, SILCS-PPI to map protein–protein interactions, and SILCS-Biologics to assess excipients for formulation development of protein-based drugs. Details of these SILCS-based analyses and their applications to BSA and pembrolizumab are presented in the following sections.

5.3.1 SILCS Simulations

SILCS simulations are performed to sample the interaction of solutes of different chemical classes and water with the target macromolecule and pre-compute SILCS FragMaps for subsequent SILCS analyses. For the SILCS simulations of BSA, the initial coordinates of BSA were extracted from its crystal structure (PDB: 4F5S [159]). For the SILCS simulations of pembrolizumab, the initial coordinates of pembrolizumab were extracted from the crystal structure of the full-length pembrolizumab mAb (PDB: 5DK3 [160]). For BSA, the entire protein was solvated in a water box. For pembrolizumab, due to its size, the Fc and Fab portions of the mAb were separated and then solvated and simulated independently. Additionally, only one Fab portion of pembrolizumab was simulated for computational expediency as both Fab portions of pembrolizumab are sequentially identical and structurally very similar. For each system, the dimensions of the water box were independently set such that the atoms of the target protein were separated from the box edge by at least 12 Å on all sides. Eight solutes representing different chemical classes (benzene, propane, dimethylether, methanol, formamide, imidazole, acetate, and methylammonium) were added into the system at a concentration of ~0.25 M, to probe the functional group preferences of the target proteins. For each protein, 10 independent simulation systems were built with different distributions of water and solute molecules as well as rotated protein sidechain configurations for the solvent-accessible residues.

Prior to the SILCS simulations, each simulation system was subjected to a 5000-step steepest descent energy minimization followed by a 250 ps MD equilibration. During the energy minimization and equilibration steps, all non-hydrogen atoms of the proteins were constrained through a harmonic positional restraint with a force constant of 2.4 kcal/mol-Å². Following the energy minimization and equilibration steps, each system was subjected to 25 GCMC cycles of 200 000 GCMC steps each to re-equilibrate the solute and water molecules around the target protein. Subsequently, a production run of 100 GCMC/MD cycles was performed. In each GCMC/MD cycle, 200 000 GCMC steps followed by 1 ns of MD simulation were executed. During the production run, Cα backbone atoms of the target protein were lightly constrained through a harmonic positional restraint with a force constant of 0.12 kcal/mol-Å². Additionally, to prevent the aggregation of solute molecules, a repulsive intermolecular wall was applied to select solute pairs. During the MD simulations, the Nosé–Hoover method was used to maintain the temperature at 298 K, and pressure was maintained at 1 bar using the Parrinello–Rahman barostat.

The CHARMM36m protein force field [157], CGenFF [144–147], and CHARMM TIP3P water model [157, 158] were used to describe protein, solutes, and water during the simulations, respectively. The GCMC portion of the runs was performed using SILCS software (SilcsBio LLC), and MD was conducted using the GROMACS [96] program. Upon completion of the 100 GCMC/MD cycles, 100 ns of simulation data was extracted per simulation system for a cumulative 1 μ s simulation time (100 ns * 10 simulation systems) per protein system (BSA, pembrolizumab Fab, and pembrolizumab Fc).

5.3.2 FragMap Construction

Using the SILCS simulation snapshots, probability distributions of the solutes and water molecules in and around the protein are calculated to produce FragMaps. The FragMaps are determined by binning selected atoms of the solute molecules into 1 $\text{\AA} \times 1 \text{\AA} \times 1 \text{\AA}$ cubic volume elements (voxels) of a grid spanning the entire system for each simulation snapshot, extracted every 10 ps. The voxel occupancies calculated in the presence of the target macromolecule are divided by the value in bulk as determined by the number of each solute in the system relative to the number of waters, assuming 55 M water, to obtain a normalized occupancy. The solute concentrations may also be set to an assumed concentration of 0.25 M or based on the total volume of the simulation system [102]. The normalized occupancies are then converted to GFE values by using the Boltzmann transformation. The GFE represents the free energy of functional groups, and GFE FragMaps can indicate both favorable and unfavorable interactions with the target macromolecule through negative and positive free energies, respectively. Additionally, an “exclusion map” is also calculated based on the regions not sampled by water or any solute molecules during the entire duration of the SILCS simulations. This exclusion map accounts for the conformational flexibility of the macromolecule. As such, the initial, rigid macromolecule experimental structure is not used for subsequent SILCS calculations aside from visualization purposes.

BSA, pembrolizumab Fab, and pembrolizumab Fc FragMaps were independently determined. The resulting FragMaps for BSA and pembrolizumab are shown in Figures 5.1a,b, respectively. For pembrolizumab, the FragMaps of the Fab and Fc domains of the mAb were determined independently as their SILCS simulations were also performed independently for computational efficiency. As only one Fab domain of pembrolizumab was simulated for computational expediency, the FragMaps resulting from the simulated Fab were reoriented to fit the second Fab domain. The independent sets of FragMaps for pembrolizumab Fc and Fab domains are overlaid on the full-length pembrolizumab structure in Figure 5.1b for clarity. The BSA FragMaps show abundant apolar and positively charge binding regions, shown by the green and cyan-colored FragMaps in Figure 5.1a at the protein surface. Additionally, negatively charged binding regions, shown by the orange-colored FragMap in Figure 5.1a, are also observed in pockets within the protein structure. The pembrolizumab FragMaps show the presence of apolar binding regions, as shown by the green-colored FragMap in Figure 5.1b, distributed across the Fc

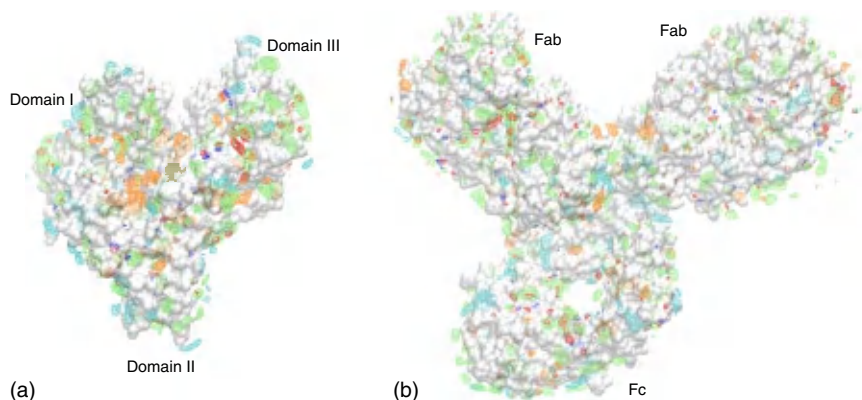


Figure 5.1 The SILCS FragMaps of (a) BSA and (b) pembrolizumab. BSA and pembrolizumab are shown in transparent surface representations. SILCS-FragMaps for generic apolar, generic H-bond donor, generic H-bond acceptor, negative, and positive groups are shown in green, blue, red, orange, and cyan mesh representations, respectively. Isocontour GFE FragMaps are shown at a contour level of -1.2 kcal/mol.

and Fab domains. Positively charged binding regions are primarily observed in the Fc domain of pembrolizumab due to the larger number of negatively charged residues in the Fc domain compared to the Fab domain. Additionally, a few pockets of hydrogen bond donor and acceptor binding regions in the Fab and Fc domains are also observed, as shown by the blue and red FragMaps in Figure 5.1b. Typically, hydrogen bond FragMaps occur at low contour levels, -0.6 kcal/mol vs. -1.2 kcal/mol shown in Figure 5.1, due to the balance of favorable solute–protein interactions and the desolvation penalty associated with such functional groups binding with the protein.

5.3.3 SILCS-MC

The SILCS FragMaps can be used to rapidly dock, score, and evaluate ligands for their binding to a target macromolecule through SILCS-MC. In the SILCS-MC method, selected atoms of the ligand are associated with a FragMap type, and a GFE score is assigned to each atom based on the value of the FragMap at that position. The atom FragMap type is translated from CGenFF atom types using an atom-classification scheme [106]. The coordinates of each atom are then used to determine its overlap with a FragMap voxel, with that atom being assigned the GFE value of the voxel. A ligand GFE (LGFE) is subsequently calculated based on a summation of the atomic GFE scores for classified atoms. The LGFEs serve as approximations to binding free energies but are not formal binding free energies due to additional factors, such as entropy loss of combining multiple smaller fragments into a larger ligand and the contribution of ligand and protein internal strain, among others, being omitted from the calculation.

The SILCS-MC docking procedure determines the most energetically favorable, or lowest LGFE, pose through a series of energy minimization, Markov chain MC,

and simulated annealing steps. The initial pose of the ligand may be randomly generated for blind docking or taken from a predetermined set of coordinates for pose refinement. For SILCS-MC docking, the ligand of interest is placed randomly in a sphere centered at a user-defined coordinate of a user-defined size, typically 5 or 10 Å radius, at which five independent SILCS-MC runs are performed to sample ligand docked poses. Each of the SILCS-MC runs involves a minimum of 50 and a maximum of 250 cycles of Monte Carlo/Simulated Annealing (MC/SA) sampling of the molecule within the user-defined search space sphere. In each cycle, 10 000 steps of MC at room temperature are followed by 40 000 steps of SA, lowering the temperature are performed with the molecule reoriented at the beginning of each cycle. Subsequently, the ligand-docked poses are scored by LGFE and ligand efficiency (LE), with the LGFE divided by the number of heavy atoms.

For the case study, SILCS-MC was performed to dock divanillin to BSA. As divanillin has been experimentally determined to bind to binding site I of BSA [161] and Trp 212 quenching is commonly used to determine ligand binding to binding site I [162–165], the C α coordinate of Trp 212 was used as the center of the 10 Å sphere search space. Figure 5.2 shows the most energetically favored, lowest LGFE,

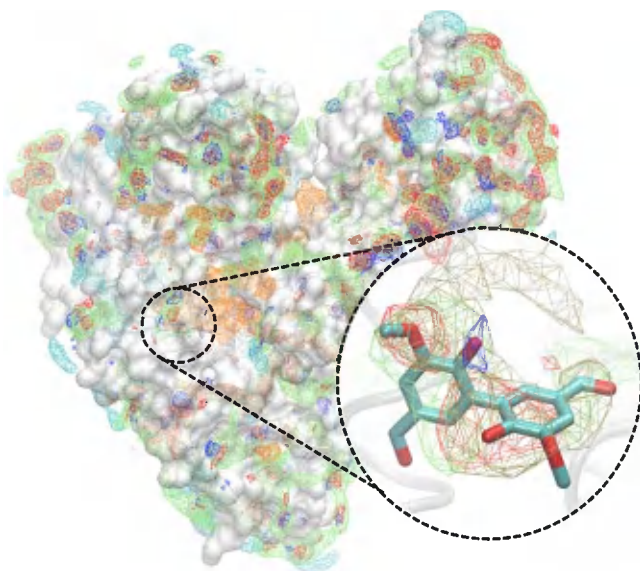


Figure 5.2 Most energetically favored conformation of divanillin binding to BSA in binding site I. Binding site I is encircled in black dotted lines with a zoomed-in view of divanillin's binding to BSA. BSA is shown in transparent surface representation and transparent gray tube representation within the zoomed-in view. SILCS-FragMaps for generic apolar, generic H-bond donor, generic H-bond acceptor, alcohol, negative, and positive groups are shown in green, blue, red, ochre, orange, and cyan mesh representation, respectively. Divanillin is shown in licorice representation. Isocontour GFE FragMaps are shown at a contour level of -0.5 kcal/mol for generic apolar, generic H-bond donor, generic H-bond acceptor, and alcohol groups and -1.2 kcal/mol for negative and positive groups. In the zoomed-in view, FragMaps for negative and positive groups are omitted, and the view is reoriented for clarity.

docked pose of divanillin binding to BSA. The LGFE of divanillin binding to BSA was -4.1 kcal/mol. In comparison, the LGFE of warfarin, a fluorescent marker of BSA binding site I, docked in the same manner using SILCS-MC was -2.3 kcal/mol. The predicted higher affinity, or more favorable LGFE, of divanillin compared to warfarin is in line with experiments, which showed that divanillin displaces warfarin within BSA binding site I [161].

5.3.4 SILCS-Hotspots

SILCS-Hotspots is an extension of SILCS-MC that identifies fragment-binding hotspots that are spatially distributed in and around the target molecule [107]. SILCS-Hotspots are performed by systematically partitioning the full 3D space of the target molecule into $14.14 \text{ \AA} \times 14.14 \text{ \AA} \times 14.14 \text{ \AA}$ subspaces in which fragments are independently docked using SILCS-MC. In each subspace, each fragment is randomly positioned in a sphere of $10 \text{ \AA}$ radius centered where the random variation of one rotatable bond of the fragment is generated through SILCS-MC. Subsequently, each fragment is subjected to 10 000 MC steps (at 300 K) followed by 40 000 MC annealing steps from 0 to 300 °K. This procedure is applied 1000 times for each fragment in each subspace. Subsequently, center-of-mass (COM)-based clustering, with a clustering radius of $3 \text{ \AA}$, is performed for each fragment to identify orientations with the highest neighbor population. An additional round of clustering using a clustering radius of $4 \text{ \AA}$ is then performed on all poses selected in the first round of clustering across all fragments to identify hotspots, which may be populated by multiple fragments. The clustering radii of the first and second clustering rounds may be adjusted, with larger clustering radii typically yielding fewer, more spatially separated identified hotspots. The LGFE of each fragment in each Hotspot site (centers of predicted fragment binding sites) is averaged to determine the average LGFE of the hotspot and hotspots with LGFE scores greater than -2 kcal/mol are typically discarded. Other metrics, such as the number of fragments in a hotspot, may be used to rank order hotspots in addition to the average LGFE score.

In the case study, SILCS-Hotspots were applied to BSA to identify potential binding sites. For this study, 135 low-molecular-weight compounds from the Astex MiniFragments probing library [123] were used as the SILCS-Hotspots fragments. The resulting Hotspots identified are shown in Figure 5.3. As shown in Figure 5.3, the hotspots encompass the entire BSA protein, including interior pockets. The presence of multiple, energetically favorable adjacent hotspots throughout BSA indicates that BSA has a high propensity to bind several classes of ligands at different sites, in line with previous studies on serum albumin [169–171]. Aligning experimentally resolved structures of BSA bound to ligands (PDB IDs 4JK4 [166], 4OR0 [167], and 6QS9 [168]) with the structure used in the SILCS simulations and analyses shows that the experimentally determined binding sites of 3,5-diiodosalicylic acid (DIU), naproxen (NPS), and *R*-ketoprofen (JGE) are captured by the top (most favorable LGFE) 15 hotspots. Close-up views of the experimentally resolved binding of DIU, NPS, and JGE to BSA in relation to the hotspots and FragMaps are encircled in black dotted lines in Figure 5.3. These results confirm the ability of SILCS-Hotspots

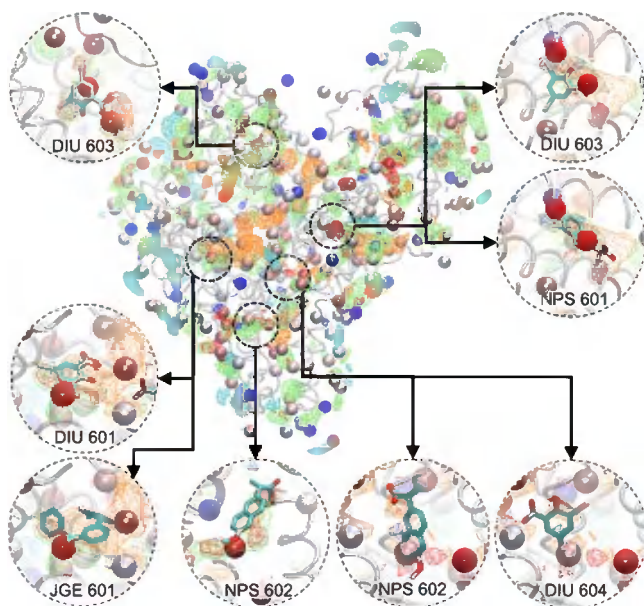


Figure 5.3 The SILCS-Hotspots and SILCS FragMaps of BSA, along with experimentally resolved binding sites of DIU, NPS, and JGE according to crystallography [166–168]. The crystallographic orientations of DIU, NPS, and JGE overlaid on BSA, the SILCS-Hotspots, and SILCS FragMaps, encircled in black dotted lines, are zoomed-in and reoriented for clarity. BSA is shown in gray, transparent tube representation; the SILCS-Hotspots are shown in VDW representation and are colored by their LGFE scores, with red indicating the most favorable and blue indicating the least favorable (-2 kcal/mol being the lowest LGFE shown); SILCS-FragMaps for generic apolar, generic H-bond donor, generic H-bond acceptor, alcohol, negative, and positive groups are shown in green, blue, red, ochre, orange, and cyan mesh representation, respectively.

to identify binding sites that correspond to known, experimentally resolved ligand binding sites. Note that in all cases, the ligand binding sites are occupied by multiple hotspots. This pattern is common for sites that are suitable for the binding of larger, drug-like molecules, including allosteric modulators [107].

5.3.5 SILCS-PPI

SILCS-PPI uses FragMaps, protein functional group probability grids, and FFTs [172] to perform protein–protein docking from which patterns of protein–protein interactions are identified. The protein functional group probability grid maps (PPGMaps) are extracted from the SILCS simulations and subsequently assigned to the corresponding FragMap types. The assignment is done such that the spatial overlap of the receptor protein FragMaps and the ligand–protein PPGMaps provides a rapid estimation of the protein receptor–protein–ligand interaction. SILCS-PPI performs protein–protein docking by maximizing the complementarity between the FragMaps of one protein and the PPGMaps through an FFT-based algorithm [110]. During the docking process, the receptor FragMaps and PPGMaps are fixed in space,

and the ligand FragMaps and PPGMaps are translated and rotated systematically over all possible orientations. The docked poses are scored using protein grid free energies (PGFEs), which are calculated based on the overlap of FragMaps and PPGMaps. PPI preference (PPIP) maps are subsequently calculated using a two-step, COM and orientation-based clustering analysis of all docked poses. After the clustering, per-residue PPIP is computed as the number of contacts between the receptor and ligand–protein atoms, with any non-hydrogen atom within a 5 Å cutoff considered in contact, and summed over the top 2000 docked poses sorted by PGFE score. Each per-residue PPIP value is subsequently normalized by the maximum per-residue PPIP value, resulting in a PPIP score. The PPIP scores range from 0 to 1 with higher PPIP scores indicating that a residue is more likely to be involved in a PPI.

In the case study, regions of high PPIP in BSA and pembrolizumab were identified using SILCS-PPI. For the case study, BSA-BSA self-PPI and pembrolizumab Fab-Fc, Fc-Fc, and Fab-Fab were considered. As the pembrolizumab Fc and Fab domains were simulated independently, the full-length pembrolizumab structure (PDB: 5DK3 [160]) was overlaid on the receptor Fab or Fc, and any docked poses in which the ligand Fab/Fc resulted in steric clashes were discarded. In this way, poses that are sterically inaccessible in the full-length pembrolizumab were excluded. The predicted PPIP maps of BSA and pembrolizumab are shown in Figure 5.4. For

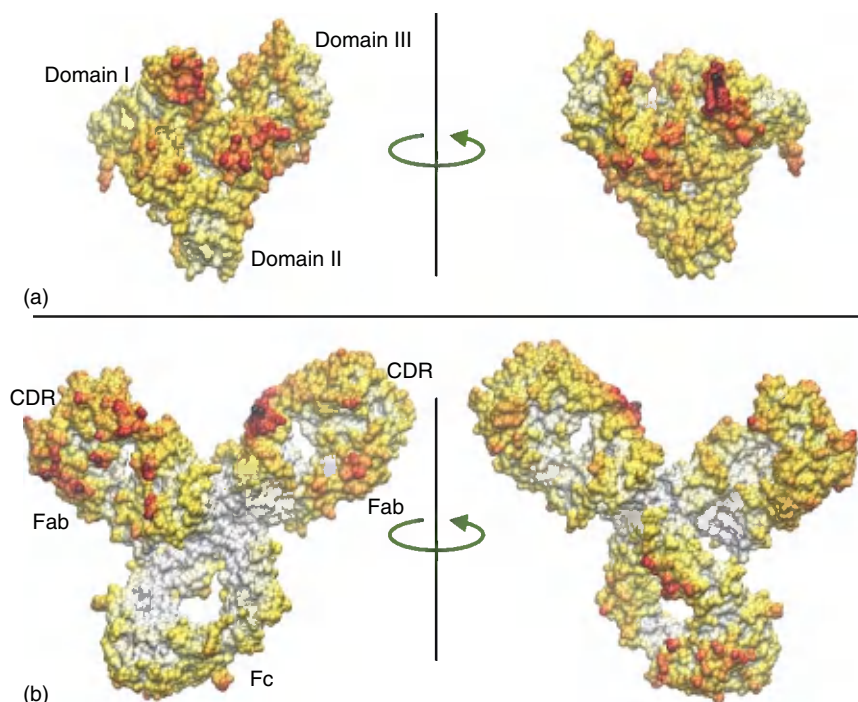


Figure 5.4 Predicted PPI preference maps of (a) BSA and (b) the full pembrolizumab (Fab and Fc). BSA and pembrolizumab are shown in surface representation with the highest PPI preference regions colored dark red and the lowest PPI preference regions colored white.

BSA, the strongest PPIP regions are in domains I and III (Figure 5.4a). The higher interaction preference of BSA in domains I and III is consistent with experimentally derived crystal structures of BSA dimers [159, 166–168] and experiments suggesting that BSA dimers are stabilized by residues Cys 34 and 513 [173], which are located at or near regions with predicted high PPIP. Interestingly, pembrolizumab does not show a particularly strong PPIP in its complementary determining region (CDR) over other domains of the mAb (Figure 5.4b). The regions with the strongest PPIP are distributed along the sides of the Fab and Fc domains of pembrolizumab and are not concentrated at the CDR. The relatively low PPIP of the CDR may explain experimental data showing the low propensity of pembrolizumab to aggregate even in refrigerated conditions, and the ability of pembrolizumab to retain its functional ability to bind with PD-1 after two weeks stored in saline solution at refrigerated conditions [174]. It is worth reiterating that the reported PPIP values are relative values and cannot be directly compared across different protein systems. Thus, comparisons of which proteins may be more prone to aggregation cannot be directly inferred from their self-PPIP values. Nevertheless, these PPIP maps may be used to inform the introduction of mutations to individual protein therapeutics to enhance their stability.

5.3.6 SILCS-Biologics

SILCS-Biologics combines SILCS-PPI and SILCS-Hotspots to guide excipient selection for therapeutic protein formulations. In SILCS-Biologics, SILCS-PPI is used to compute a protein–protein self-interaction map, and SILCS-Hotspots are used to identify binding sites of a set of excipients of interest. Subsequently, combining PPI self-interaction and excipient hotspot maps, SILCS-Biologics produces a range of data that can be processed through data science and machine learning approaches to predict various experimental properties. For example, the number of excipient binding sites overlapping with regions with high predicted PPIP has been shown to correlate with the experimentally determined viscosity for several excipients, including amino acids and sugars [112, 113]. Additionally, the number of binding sites with predicted high binding affinity may predict relative protein stability [112, 113].

In the case study, SILCS-Biologics was applied for pembrolizumab to examine how excipients interact with high PPIP regions of the mAb. For this case study, excipients histidine and sucrose were investigated as they are included in the commercial formulation of pembrolizumab (available at 25 mg/mL), Keytruda. Figure 5.5 shows the sites where the excipients are predicted to bind on the surface of pembrolizumab. According to the predicted binding sites, histidine and sucrose cooperatively bind to high PPIP regions of pembrolizumab (Figure 5.5). Histidine and sucrose exclusively bind to portions of pembrolizumab, particularly on the Fc domain of the mAb. The binding sites of histidine and sucrose covering high PPIP regions are hypothesized to prevent self-PPI, which would normally lead to aggregation and increased viscosity. Such atomistic-level understanding of how excipients interact with protein therapeutics in conjunction with how protein therapeutics may self-interact can

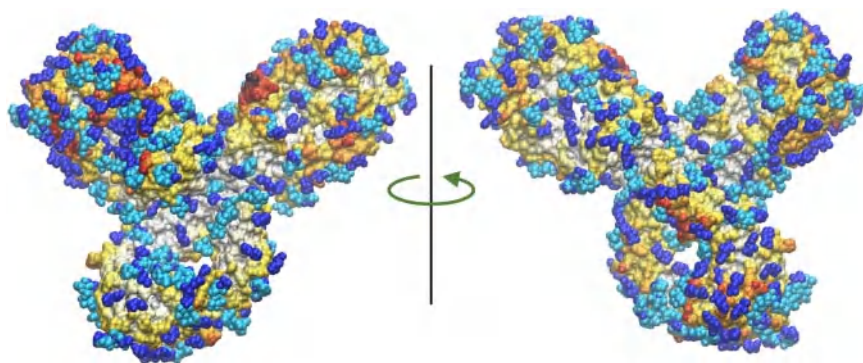


Figure 5.5 Predicted excipient binding sites and PPI preference map of the full pembrolizumab (Fab and Fc). Histidine and sucrose molecules are shown in blue and cyan VDW representations, respectively. Pembrolizumab is shown in surface representation with the highest PPI preference regions colored dark red and the lowest PPI preference regions colored white.

demystify the selection and optimization of excipient formulation to maximize protein stability and minimize aggregation and viscosity.

5.4 Conclusion

Co-solvent MD methods have become useful tools in CADD and have been successfully applied to a wide variety of macromolecular targets. These methods are advantageous to many other CADD methods as protein flexibility and competition with water are incorporated into the resulting predictions. SILCS, now commercialized in the form of SilcsBio LLC, represents one of the most extensively developed co-solvent MD methods, with the MacKerell laboratory making continual improvements and extensions of the SILCS technology. From one set of SILCS simulations, affinity patterns for diverse functional groups in the form of FragMaps are generated in and around the target macromolecule. These FragMaps are then used as the basis for a wide range of SILCS analyses, which include, among others, ligand docking through SILCS-MC, identification of allosteric binding sites through SILCS-Hotspots, protein–protein docking through SILCS-PPI, and protein therapeutic formulation through SILCS-Biologics. The included case studies for BSA and pembrolizumab show how a single pre-computed FragMaps of a target macromolecule can be used for a wide range of applications and analyses. Overall, the wide variety of analyses possible with SILCS sets it apart from other co-solvent methods and suggests that the technology may be expanded beyond its current uses.

Conflict of Interest

ADM Jr. is co-founder and Chief Scientific Officer of SilcsBio LLC.

Acknowledgments

The authors acknowledge financial support from NIH GM131710 and R44GM 130198 and computational resources provided by the Computer-Aided Drug Design (CADD) Center at the University of Maryland, Baltimore, as well as the Extreme Science and Engineering Discovery Environment (XSEDE).

References

- 1 Hansch, C., Maloney, P.P., Fujita, T., and Muir, R.M. (1962). Correlation of biological activity of phenoxyacetic acids with hammett substituent constants and partition coefficients. *Nature* 194 (4824): 178–180.
- 2 Hansch, C. and Fujita, T. (1964). P- Σ -II analysis. A method for the correlation of biological activity and chemical structure. *J. Am. Chem. Soc.* 86 (8): 1616–1626.
- 3 Schultz, T.W., Lin, D.T., and Arnold, L.M. (1991). Qsars for monosubstituted anilines eliciting the polar narcosis mechanism of action. *Sci. Total Environ.* 109–110: 569–580.
- 4 Aptula, A.O., Netzeva, T.I., Valkova, I.V. et al. (2002). Multivariate discrimination between modes of toxic action of phenols. *Quant. Struct.-Activity Relat.* 21 (1): 12–22.
- 5 Ma, Q.-S., Yao, Y., Zheng, Y.-C. et al. (2019). Ligand-based design, synthesis and biological evaluation of xanthine derivatives as Lsd1/Kdm1a inhibitors. *Eur. J. Med. Chem.* 162: 555–567.
- 6 Mirabello, C. and Wallner, B. (2020). Interlig: improved ligand-based virtual screening using topologically independent structural alignments. *Bioinformatics* 36 (10): 3266–3267.
- 7 Jia, X., Ciallella, H.L., Russo, D.P. et al. (2021). Construction of a virtual opioid bioprofile: a data-driven Qsar modeling study to identify new analgesic opioids. *ACS Sustain. Chem. Eng.* 9 (10): 3909–3919.
- 8 Bajad, N.G., Swetha, R., Singh, R. et al. (2022). Combined structure and ligand-based design of dual Bace-1/Gsk-3 β inhibitors for Alzheimer's disease. *Chem. Pap.* .
- 9 Perron, Q., Mirguet, O., Tajmouati, H. et al. (2022). Deep generative models for ligand-based de novo design applied to multi-parametric optimization. *J. Comput. Chem.* 43 (10): 692–703.
- 10 Koes, D.R. and Camacho, C.J. (2012). Zincpharmer: pharmacophore search of the zinc database. *Nucleic Acids Res.* 40 (Web Server issue): W409–14.
- 11 Ke, Y.-Y., Singh, V.K., Coumar, M.S. et al. (2015). Homology modeling of Dfg-in Fms-like tyrosine kinase 3 (Flt3) and structure-based virtual screening for inhibitor identification. *Sci. Rep.* 5 (1): 11702.
- 12 Parvaiz, N., Ahmad, F., Yu, W. et al. (2021). Discovery of β -lactamase Cmy-10 inhibitors for combination therapy against multi-drug resistant enterobacteriaceae. *PLoS ONE* 16 (1): e0244967.

- 13 Tabrez, S., Zughaibi, T.A., Hoque, M. et al. (2022). Targeting glutaminase by natural compounds: structure-based virtual screening and molecular dynamics simulation approach to suppress cancer progression. *Molecules* 27 (15): 5042.
- 14 Wang, L., Wu, Y., Deng, Y. et al. (2015). Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *J. Am. Chem. Soc.* 137 (7): 2695–2703.
- 15 Zhang, H., Jiang, W., Chatterjee, P., and Luo, Y. (2019). Ranking reversible covalent drugs: from free energy perturbation to fragment docking. *J. Chem. Inf. Model.* 59 (5): 2093–2102.
- 16 Cournia, Z., Allen, B.K., Beuming, T. et al. (2020). Rigorous free energy simulations in virtual screening. *J. Chem. Inf. Model.* 60 (9): 4153–4169.
- 17 Nikiforov, P.O., Blaszczyk, M., Surade, S. et al. (2017). Fragment-sized ethr inhibitors exhibit exceptionally strong ethionamide boosting effect in whole-cell mycobacterium tuberculosis assays. *ACS Chem. Biol.* 12 (5): 1390–1396.
- 18 Kessler, D., Gmachl, M., Mantoulidis, A. et al. (2019). Drugging an undruggable pocket on Kras. *Proc. Natl. Acad. Sci.* 116 (32): 15823–15829.
- 19 Abdul-Hammed, M., Adedotun, I.O., Falade, V.A. et al. (2021). Target-based drug discovery, admet profiling and bioactivity studies of antibiotics as potential inhibitors of Sars-Cov-2 main protease (Mpro). *Virusdisease* 32 (4): 642–656.
- 20 El Aissouq, A., Bouachrine, M., Ouammou, A., and Khalil, F. (2022). Homology modeling, virtual screening, molecular docking, molecular dynamic (Md) simulation, and admet approaches for identification of natural anti-parkinson agents targeting mao-B protein. *Neurosci. Lett.* 786: 136803.
- 21 Guedes, I.A., Pereira, F.S.S., and Dardenne, L.E. (2018). Empirical scoring functions for structure-based virtual screening: applications, critical aspects, and challenges. *Front. Pharmacol.* 9: 1089.
- 22 Maia, E.H.B., Assis, L.C., de Oliveira, T.A. et al. (2020). Structure-based virtual screening: from classical to artificial intelligence. *Front. Chem.* 8: 343.
- 23 Fischer, A., Smieško, M., Sellner, M., and Lill, M.A. (2021). Decision making in structure-based drug discovery: visual inspection of docking results. *J. Med. Chem.* 64 (5): 2489–2500.
- 24 Hussain, W., Rasool, N., and Khan, Y.D. (2021). Insights into machine learning-based approaches for virtual screening in drug discovery: existing strategies and streamlining through Fp-Cadd. *Curr. Drug Discov. Technol.* 18 (4): 463–472.
- 25 Sabe, V.T., Ntombela, T., Jhamba, L.A. et al. (2021). Current trends in computer aided drug design and a highlight of drugs discovered via computational techniques: a review. *Eur. J. Med. Chem.* 224: 113705.
- 26 Giordano, D., Biancaniello, C., Argenio, M.A., and Facchiano, A. (2022). Drug design by pharmacophore and virtual screening approach. *Pharmaceuticals (Basel)* 15 (5).
- 27 Lee, J.W., Maria-Solano, M.A., Vu, T.N.L. et al. (2022). Big data and artificial intelligence (Ai) methodologies for computer-aided drug design (Cadd). *Biochem. Soc. Trans.* 50 (1): 241–252.

- 28 Warwicker, J. and Watson, H.C. (1982). Calculation of the electric potential in the active site cleft due to α -helix dipoles. *J. Mol. Biol.* 157 (4): 671–679.
- 29 Klapper, I., Hagstrom, R., Fine, R. et al. (1986). Focusing of electric fields in the active site of Cu-Zn superoxide dismutase: effects of ionic strength and amino-acid modification. *Proteins: Struct. Funct. Bioinf.* 1 (1): 47–59.
- 30 Nicholls, A. and Honig, B. (1991). A rapid finite difference algorithm, utilizing successive over-relaxation to solve the poisson-boltzmann equation. *J. Comput. Chem.* 12 (4): 435–445.
- 31 Constanciel, R. and Contreras, R. (1984). Self consistent field theory of solvent effects representation by continuum models: introduction of desolvation contribution. *Theor. Chim. Acta* 65 (1): 1–11.
- 32 Still, W.C., Tempczyk, A., Hawley, R.C., and Hendrickson, T. (1990). Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* 112 (16): 6127–6129.
- 33 Genheden, S. and Ryde, U. (2012). Comparison of End-Point Continuum-Solvation Methods for the Calculation of Protein-Ligand Binding Free Energies. *Proteins* 80 (5): 1326–1342.
- 34 Genheden, S. and Ryde, U. (2015). The Mm/Pbsa and Mm/Gbsa methods to estimate ligand-binding affinities. *Expert Opin. Drug Discovery* 10 (5): 449–461.
- 35 Wang, E., Sun, H., Wang, J. et al. (2019). End-point binding free energy calculation with Mm/Pbsa and Mm/Gbsa: strategies and applications in drug design. *Chem. Rev.* 119 (16): 9478–9508.
- 36 Orr, A.A., Yang, J., Sule, N. et al. (2020). Molecular mechanism for attractant signaling to dhma by *E. coli* Tsr. *Biophys. J.* 118 (2): 492–504.
- 37 Landau, L.D. (1938). *Statistical Physics*. Oxford: Clarendon.
- 38 Zwanzig, R.W. (1954). High-temperature equation of state by a perturbation method. I. Nonpolar gases. *J. Chem. Phys.* 22 (8): 1420–1426.
- 39 Hirono, S. and Kollman, P.A. (1990). Calculation of the relative binding free energy of 2'gmp and 2'amp to ribonuclease T1 using molecular dynamics/free energy perturbation approaches. *J. Mol. Biol.* 212 (1): 197–209.
- 40 Mutyala, R., Reddy, R.N., Sumakanth, M. et al. (2007). Calculation of relative binding affinities of fructose 1,6-bisphosphatase mutants with adenosine monophosphate using free energy perturbation method. *J. Comput. Chem.* 28 (5): 932–937.
- 41 Jiang, Z.-Y., Lu, M.-C., Xu, L.L. et al. (2014). Discovery of potent Keap1-Nrf2 protein-protein interaction inhibitor based on molecular binding determinants analysis. *J. Med. Chem.* 57 (6): 2736–2745.
- 42 Clark, A.J., Gindin, T., Zhang, B. et al. (2017). Free energy perturbation calculation of relative binding free energy between broadly neutralizing antibodies and the Gp120 glycoprotein of Hiv-1. *J. Mol. Biol.* 429 (7): 930–947.
- 43 Cournia, Z., Chipot, C., Roux, B. et al. (2021). Free energy methods in drug discovery—introduction. In: *Free Energy Methods in Drug Discovery: Current State and Future Directions*, vol. 1397, 1–38. American Chemical Society.
- 44 Mucs, D. and Bryce, R.A. (2013). The application of quantum mechanics in structure-based drug design. *Expert Opin. Drug Discovery* 8 (3): 263–276.

- 45 Cavasotto, C.N., Adler, N.S., and Aucar, M.G. (2018). Quantum chemical approaches in structure-based virtual screening and lead optimization. *Front. Chem.* 6.
- 46 Bryce, R.A. (2020). What next for quantum mechanics in structure-based drug discovery? In: *Quantum Mechanics in Drug Discovery* (ed. A. Heifetz), 339–353. New York, NY: Springer US.
- 47 Bissaro, M., Sturlese, M., and Moro, S. (2020). The rise of molecular simulations in fragment-based drug design (Fbdd): an overview. *Drug Discov. Today* 25 (9): 1693–1701.
- 48 Allen, K.N., Bellamacina, C.R., Ding, X. et al. (1996). An experimental approach to mapping the binding surfaces of crystalline proteins. *J. Phys. Chem.* 100 (7): 2605–2611.
- 49 Goodford, P.J. (1985). A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* 28 (7): 849–857.
- 50 Joseph-McCarthy, D., Hogle, J.M., and Karplus, M. (1997). Use of the multiple copy simultaneous search (Mcscs) method to design a new class of picornavirus capsid binding drugs. *Proteins* 29 (1): 32–58.
- 51 Raman, E.P., Lakkaraju, S.K., Denny, R.A., and MacKerell, A.D. Jr., (2017). Estimation of relative free energies of binding using pre-computed ensembles based on the single-step free energy perturbation and the site-identification by ligand competitive saturation approaches. *J. Comput. Chem.* 38 (15): 1238–1251.
- 52 Seco, J., Luque, F.J., and Barril, X. (2009). Binding site detection and druggability index from first principles. *J. Med. Chem.* 52 (8): 2363–2371.
- 53 Alvarez-Garcia, D. and Barril, X. (2014). Molecular simulations with solvent competition quantify water displaceability and provide accurate interaction maps of protein binding sites. *J. Med. Chem.* 57 (20): 8530–8539.
- 54 Arcon, J.P., Defelipe, L.A., Modenutti, C.P. et al. (2017). Molecular dynamics in mixed solvents reveals protein–ligand interactions, improves docking, and allows accurate binding free energy predictions. *J. Chem. Inf. Model.* 57 (4): 846–863.
- 55 Arcon, J.P., Defelipe, L.A., Lopez, E.D. et al. (2019). Cosolvent-based protein pharmacophore for ligand enrichment in virtual screening. *J. Chem. Inf. Model.* 59 (8): 3572–3583.
- 56 Lexa, K.W. and Carlson, H.A. (2011). Full protein flexibility is essential for proper hot-spot mapping. *J. Am. Chem. Soc.* 133 (2): 200–202.
- 57 Lexa, K.W. and Carlson, H.A. (2013). Improving protocols for protein mapping through proper comparison to crystallography data. *J. Chem. Inf. Model.* 53 (2): 391–402.
- 58 Ghanakota, P. and Carlson, H.A. (2016). Driving structure-based drug discovery through cosolvent molecular dynamics: miniperspective. *J. Med. Chem.* 59 (23): 10383–10399.
- 59 Ung, P.M., Ghanakota, P., Graham, S.E. et al. (2016). Identifying binding hot spots on protein surfaces by mixed-solvent molecular dynamics: Hiv-1 protease as a test case. *Biopolymers* 105 (1): 21–34.

- 60 Graham, S.E., Leja, N., and Carlson, H.A. (2018). Mixmd probeview: robust binding site prediction from cosolvent simulations. *J. Chem. Inf. Model.* 58 (7): 1426–1433.
- 61 Ghanakota, P., DasGupta, D., and Carlson, H.A. (2019). Free energies and entropies of binding sites identified by mixmd cosolvent simulations. *J. Chem. Inf. Model.* 59 (5): 2035–2045.
- 62 Chan, W.K.B., DasGupta, D., Carlson, H.A., and Traynor, J.R. (2021). Mixed-solvent molecular dynamics simulation-based discovery of a putative allosteric site on regulator of G protein signaling 4. *J. Comput. Chem.* 42 (30): 2170–2180.
- 63 Smith, R.D. and Carlson, H.A. (2021). Identification of cryptic binding sites using mixmd with standard and accelerated molecular dynamics. *J. Chem. Inf. Model.* 61 (3): 1287–1299.
- 64 Prakash, P., Hancock, J.F., and Gorfe, A.A. (2015). Binding hotspots on K-Ras: consensus ligand binding sites and other reactive regions from probe-based molecular dynamics analysis. *Proteins: Struct. Funct. Bioinf.* 83 (5): 898–909.
- 65 Sayyed-Ahmad, A. and Gorfe, A.A. (2017). Mixed-probe simulation and probe-derived surface topography map analysis for ligand binding site identification. *J. Chem. Theory Comput.* 13 (4): 1851–1861.
- 66 Sayyed-Ahmad, A. (2018). Hotspot identification on protein surfaces using probe-based md simulations: successes and challenges. *Curr. Top. Med. Chem.* 18 (27): 2278–2283.
- 67 Yang, C.-Y. and Wang, S. (2010). Computational analysis of protein hotspots. *ACS Med. Chem. Lett.* 1 (3): 125–129.
- 68 Yang, C.-Y. and Wang, S. (2011). Hydrophobic binding hot spots of Bcl-XI protein–protein interfaces by cosolvent molecular dynamics simulation. *ACS Med. Chem. Lett.* 2 (4): 280–284.
- 69 Yang, C.-Y. and Wang, S. (2012). Analysis of flexibility and hotspots in Bcl-XI and Mcl-1 proteins for the design of selective small-molecule inhibitors. *ACS Med. Chem. Lett.* 3 (4): 308–312.
- 70 Yang, C.-Y. (2015). Identification of potential small molecule allosteric modulator sites on Il-1r1 ectodomain using accelerated conformational sampling method. *PLoS ONE* 10 (2): e0118671.
- 71 Privat, C., Granadino-Roldan, J.M., Bonet, J. et al. (2021). Fragment dissolved molecular dynamics: a systematic and efficient method to locate binding sites. *Phys. Chem. Chem. Phys.* 23 (4): 3123–3134.
- 72 Martinez-Rosell, G., Harvey, M.J., and De Fabritiis, G. (2018). Molecular-simulation-driven fragment screening for the discovery of new Cxcl12 inhibitors. *J. Chem. Inf. Model.* 58 (3): 683–691.
- 73 Martinez-Rosell, G., Lovera, S., Sands, Z.A., and De Fabritiis, G. (2020). Play-molecule crypticscout: predicting protein cryptic sites using mixed-solvent molecular simulations. *J. Chem. Inf. Model.* 60 (4): 2314–2324.
- 74 Kimura, S.R., Hu, H.P., Ruvinsky, A.M. et al. (2017). Deciphering cryptic binding sites on proteins by mixed-solvent molecular dynamics. *J. Chem. Inf. Model.* 57 (6): 1388–1401.

- 75 Zariquiey, F.S., de Souza, J.V., and Bronowska, A.K. (2019). Cosolvent analysis toolkit (Cat): a robust hotspot identification platform for cosolvent simulations of proteins to expand the druggable proteome. *Sci. Rep.* 9 (1): 1–14.
- 76 Kalenkiewicz, A., Grant, B.J., and Yang, C.Y. (2015). Enrichment of drug-gable conformations from apo protein structures using cosolvent-accelerated molecular dynamics. *Biology (Basel)* 4 (2): 344–366.
- 77 Bakan, A., Nevins, N., Lakdawala, A.S., and Bahar, I. (2012). Druggability assessment of allosteric proteins by dynamics simulations in the presence of probe molecules. *J. Chem. Theory Comput.* 8 (7): 2435–2447.
- 78 Lee, J.Y., Krieger, J.M., Li, H., and Bahar, I. (2020). Pharmmaker: pharmacophore modeling and hit identification based on druggability simulations. *Protein Sci.* 29 (1): 76–86.
- 79 Huang, D.Z. and Caflisch, A. (2011). The free energy landscape of small molecule unbinding. *PLoS Comput. Biol.* 7 (2).
- 80 Huang, D., Rossini, E., Steiner, S., and Caflisch, A. (2014). Structured water molecules in the binding site of bromodomains can be displaced by cosolvent. *ChemMedChem* 9 (3): 573–579.
- 81 Tan, Y.S., Śledź, P., Lang, S. et al. (2012). Using ligand-mapping simulations to design a ligand selectively targeting a cryptic surface pocket of polo-like kinase 1. *Angew. Chem.* 124 (40): 10225–10228.
- 82 Tan, Y.S., Spring, D.R., Abell, C., and Verma, C. (2014). The use of chlorobenzene as a probe molecule in molecular dynamics simulations. *J. Chem. Inf. Model.* 54 (7): 1821–1827.
- 83 Tan, Y.S., Spring, D.R., Abell, C., and Verma, C.S. (2015). The application of ligand-mapping molecular dynamics simulations to the rational design of peptidic modulators of protein–protein interactions. *J. Chem. Theory Comput.* 11 (7): 3199–3210.
- 84 Tan, Y.S. and Verma, C.S. (2020). Straightforward incorporation of multiple ligand types into molecular dynamics simulations for efficient binding site detection and characterization. *J. Chem. Theory Comput.* 16 (10): 6633–6644.
- 85 Yang, Y., Mahmoud, A.H., and Lill, M.A. (2018). Modeling of halogen–protein interactions in co-solvent molecular dynamics simulations. *J. Chem. Inf. Model.* 59 (1): 38–42.
- 86 Mahmoud, A.H., Yang, Y., and Lill, M.A. (2019). Improving atom-type diversity and sampling in cosolvent simulations using lambda-dynamics. *J. Chem. Theory Comput.* 15 (5): 3272–3287.
- 87 Yanagisawa, K., Moriwaki, Y., Terada, T., and Shimizu, K. (2021). Exprorer: rational cosolvent set construction method for cosolvent molecular dynamics using large-scale computation. *J. Chem. Inf. Model.* 61: 2744–2753.
- 88 Takemura, K., Sato, C., and Kitao, A. (2018). Coldock: concentrated ligand docking with all-atom molecular dynamics simulation. *J. Phys. Chem. B* 122 (29): 7191–7200.
- 89 Guvench, O. and MacKerell, A.D. Jr., (2009). Computational fragment-based binding site identification by ligand competitive saturation. *PLoS Comput. Biol.* 5 (7): e1000435.

- 90 Raman, E.P., Yu, W., Guvench, O., and MacKerell, A.D. (2011). Reproducing crystal binding modes of ligand functional groups using site-identification by ligand competitive saturation (Silcs) simulations. *J. Chem. Inf. Model.* 51 (4): 877–896.
- 91 Raman, E.P., Yu, W., Lakkaraju, S.K., and MacKerell, A.D. Jr., (2013). Inclusion of multiple fragment types in the site identification by ligand competitive saturation (Silcs) approach. *J. Chem. Inf. Model.* 53 (12): 3384–3398.
- 92 Goel, H., Hazel, A., Ustach, V.D. et al. (2021). Rapid and accurate estimation of protein-ligand relative binding affinities using site-identification by ligand competitive saturation. *Chem. Sci.* 12: 8844–8858.
- 93 Hamelberg, D., Mongan, J., and McCammon, J.A. (2004). Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. *J. Chem. Phys.* 120 (24): 11919–11929.
- 94 Case, D.A., Cheatham, T.E. 3rd, Darden, T. et al. (2005). The amber biomolecular simulation programs. *J. Comput. Chem.* 26 (16): 1668–1688.
- 95 Eastman, P., Swails, J., Chodera, J.D. et al. (2017). Openmm 7: rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.* 13 (7): e1005659.
- 96 Van der Spoel, D., Lindahl, E., Hess, B. et al. (2005). Gromacs: fast, flexible, and free. *J. Comput. Chem.* 26 (16): 1701–1718.
- 97 Phillips, J.C., Braun, R., Wang, W. et al. (2005). Scalable molecular dynamics with NAMD. *J. Comput. Chem.* 26 (16): 1781–1802.
- 98 Brooks, B.R., Brucoleri, R.E., Olafson, B.D. et al. (1983). Charmm: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* 4 (2): 187–217.
- 99 Roe, D.R. and Cheatham, T.E. (2013). Ptraj and Cpptraj: software for processing and analysis of molecular dynamics trajectory data. *J. Chem. Theory Comput.* 9 (7): 3084–3095.
- 100 Foster, T.J., MacKerell, A.D. Jr., and Guvench, O. (2012). Balancing target flexibility and target denaturation in computational fragment-based inhibitor discovery. *J. Comput. Chem.* 33 (23): 1880–1891.
- 101 Goel, H., Hazel, A., Yu, W. et al. (2022). Application of site-identification by ligand competitive saturation in computer-aided drug design. *New J. Chem.* 46 (3): 919–932.
- 102 Lind, C., Pandey, P., Pastor, R.W., and MacKerell, A.D. Jr., (2021). Functional group distributions, partition coefficients, and resistance factors in lipid bilayers using site identification by ligand competitive saturation. *J. Chem. Theory Comput.* 17 (5): 3188–3202.
- 103 Humphrey, W., Dalke, A., and Schulten, K. (1996). Vmd: visual molecular dynamics. *J. Mol. Graph.* 14: 33–38.
- 104 DeLano, W.L. (2002). Pymol: an open-source molecular graphics tool. *CCP4 Newsletter Protein Crystallogr.* 40 (1): 82–92.
- 105 Lakkaraju, S.K., Raman, E.P., Yu, W., and MacKerell, A.D. Jr., (2014). Sampling of organic solutes in aqueous and heterogeneous environments using

- oscillating μ ex grand canonical-like monte carlo-molecular dynamics simulations. *J. Chem. Theory Comput.* 10 (6): 2281–2290.
- 106 Ustach, V.D., Lakkaraju, S.K., Jo, S. et al. (2019). Optimization and evaluation of site-identification by ligand competitive saturation (Silcs) as a tool for target-based ligand optimization. *J. Chem. Inf. Model.* 59 (6): 3018–3035.
 - 107 MacKerell, A.D. Jr., Jo, S., Lakkaraju, S.K. et al. (2020). Identification and characterization of fragment binding sites for allosteric ligand design using the site identification by ligand competitive saturation hotspots approach (silcs-hotspots). *Biochim. Biophys. Acta, Gen. Subj.* 1864 (4): 129519.
 - 108 Yu, W., Lakkaraju, S.K., Raman, E.P., and MacKerell, A.D. Jr., (2014). Site-identification by ligand competitive saturation (Silcs) assisted pharmacophore modeling. *J. Comput. Aided Mol. Des.* 28 (5): 491–507.
 - 109 Yu, W., Lakkaraju, S.K., Raman, E.P. et al. (2015). Pharmacophore modeling using site-identification by ligand competitive saturation (silcs) with multiple probe molecules. *J. Chem. Inf. Model.* 55 (2): 407–420.
 - 110 Yu, W., Jo, S., Lakkaraju, S.K. et al. (2019). Exploring protein-protein interactions using the site-identification by ligand competitive saturation methodology. *Proteins: Struct. Funct. Bioinf.* 87 (4): 289–301.
 - 111 Kognole, A.A., Hazel, A., and MacKerell, A.D. (2022). Silcs-Rna: toward a structure-based drug design approach for targeting rnas with small molecules. *J. Chem. Theory Comput.* 18 (9): 5672–5691.
 - 112 Jo, S., Xu, A., Curtis, J.E. et al. (2020). Computational characterization of antibody-excipient interactions for rational excipient selection using the site identification by ligand competitive saturation (silcs)-biologics approach. *Mol. Pharm.* 17: 4323–4333.
 - 113 Somani, S., Jo, S., Thirumangalathu, R. et al. (2021). Toward biotherapeutics formulation composition engineering using site-identification by ligand competitive saturation (silcs). *J. Pharm. Sci.* 110 (3): 1103–1110.
 - 114 Yu, W., Weber, D.J., Shapiro, P., and MacKerell, A.D. (2020). Developing kinase inhibitors using computer-aided drug design approaches. In: *Next Generation Kinase Inhibitors: Moving Beyond the Atp Binding/Catalytic Sites* (ed. P. Shapiro), 81–108. Cham: Springer International Publishing.
 - 115 Young, B.D., Yu, W., Rodríguez, D.J.V. et al. (2021). Specificity of molecular fragments binding to S100b versus S100a1 as identified by nmr and site identification by ligand competitive saturation (Silcs). *Molecules* 26 (2).
 - 116 Jiang, W., Zhang, H., Yichun, L. et al. (2022). Binding free energies of piezo1 channel agonists at protein-membrane interface. *bioRxiv* 2022.06.27.497657.
 - 117 Heinzl, G.A., Huang, W., Yu, W. et al. (2016). Iminoguanidines as allosteric inhibitors of the iron-regulated heme oxygenase (Hemo) of *pseudomonas aeruginosa*. *J. Med. Chem.* 59 (14): 6929–6942.
 - 118 Cheng, H., Linhares, B.M., Yu, W. et al. (2018). Identification of thiourea-based inhibitors of the B-cell lymphoma 6 Btb domain via nmr-based fragment screening and computer-aided drug design. *J. Med. Chem.* 61: 7573–7588.

- 119 Lanning, M.E., Yu, W., Yap, J.L. et al. (2016). Structure-based design of N-substituted 1-hydroxy-4-sulfamoyl-2-naphthoates as selective inhibitors of the Mcl-1 oncoprotein. *Eur. J. Med. Chem.* 113: 273–292.
- 120 Lakkaraju, S.K., Yu, W., Raman, E.P. et al. (2015). Mapping functional group free energy patterns at protein occluded sites: nuclear receptors and G-protein coupled receptors. *J. Chem. Inf. Model.* 55: 700–708.
- 121 Zhao, M., Kognole, A.A., Jo, S. et al. (2023). GPU-specific algorithms for improved solute sampling in grand canonical Monte Carlo simulations. *J. Comput. Chem.* 44 (20): 1719. <https://doi.org/10.1002/jcc.27121>.
- 122 Taylor, R.D., MacCoss, M., and Lawson, A.D. (2014). Rings in drugs: miniper-spective. *J. Med. Chem.* 57 (14): 5845–5859.
- 123 O'Reilly, M., Cleasby, A., Davies, T.G. et al. (2019). Crystallographic screening using ultra-low-molecular-weight ligands to guide drug design. *Drug Discov. Today* 24 (5): 1081–1086.
- 124 Gomes, A., da Silva, G.F., Lakkaraju, S.K. et al. (2021). Insights into glucose-6-phosphate allosteric activation of β -glucosidase A. *J. Chem. Inf. Model.* 61 (4): 1931–1941.
- 125 Shah, S.D., Lind, C., De Pascali, F. et al. (2022). In silico identification of a β 2 adrenergic receptor allosteric site that selectively augments canonical β 2ar-Gs signaling and function. *FASEB J.* 36 (S1).
- 126 Shah, S.D., Lind, C., De Pascali, F. et al. (2022). In silico identification of a β 2-adrenoceptor allosteric site that selectively augments cannical β 2args signaling and function. *PNAS* .
- 127 Weston, S., Baracco, L., Keller, C. et al. (2020). The Ski complex is a broad-spectrum, host-directed antiviral drug target for coronaviruses, influenza, and filoviruses. *Proc. Natl. Acad. Sci.* 117 (48): 30687–30698.
- 128 Koes, D.R. and Camacho, C.J. (2011). Pharmer: efficient and exact pharma-cophore search. *J. Chem. Inf. Model.* 51 (6): 1307–1314.
- 129 Oashi, T., Ringer, A.L., Raman, E.P., and MacKerell, J.A.D. (2011). Auto-mated selection of compounds with physicochemical properties to maximize bioavailability and druglikeness. *J. Chem. Inf. Model.* 51: 148–158.
- 130 Macias, A.T., Mia, Y., Xia, G. et al. (2005). Lead validation and sar develop-ment via chemical similarity searching; application to compounds targeting the Py + 3 site of the Sh2 domain of P56lck. *J. Chem. Inf. Model.* 45: 1759–1766.
- 131 Solano-Gonzalez, E., Coburn, K.M., Yu, W. et al. (2021). Small molecules inhibitors of the heterogeneous ribonuclear protein A18 (Hnrrnp A18): a regula-tor of protein translation and an immune checkpoint. *Nucleic Acids Res.* 49 (3): 1235–1246.
- 132 Samadani, R., Zhang, J., Brophy, A. et al. (2015). Small molecule inhibitors of Erk-mediated immediate early gene expression and proliferation of melanoma cells expressing mutated braf. *Biochem. J.* 467: 425–438.
- 133 He, X., Lakkaraju, S.K., Hanscom, M. et al. (2015). Acyl-2-aminobenzimidazoles: a novel class of neuroprotective agents targeting Mglur5. *Bioorg. Med. Chem.* 23 (9): 2211–2220.

- 134 Lakkaraju, S.K., Mbatia, H., Hanscom, M. et al. (2015). Cyclopropyl-containing positive allosteric modulators of metabotropic glutamate receptor subtype 5. *Bioorg. Med. Chem. Lett.* 25 (11): 2275–2279.
- 135 Glassford, I., Teijaro, C.N., Daher, S.S. et al. (2016). Ribosome-templated azide-alkyne cycloadditions: synthesis of potent macrolide antibiotics by in situ click chemistry. *J. Am. Chem. Soc.* 138 (9): 3136–3144.
- 136 Mousaei, M., Kudaibergenova, M., MacKerell, A.D. Jr., and Noskov, S. (2020). Assessing Herg1 blockade from bayesian machine-learning-optimized site identification by ligand competitive saturation simulations. *J. Chem. Inf. Model.* 60 (12): 6489–6501.
- 137 Goel, H., Yu, W., and MacKerell, A.D. Jr., (2022). Herg blockade prediction by combining site identification by ligand competitive saturation and physico-chemical properties. *MDPI Chem.* 4: 630–645.
- 138 Pandey, P. and MacKerell, A. (2023). Combining SILCS and artificial intelligence for high-throughput prediction of drug molecule passive permeability. *J. Chem. Inf. Model.* 63 (18): 5903–5915. <https://doi.org/10.1021/acs.jcim.3c00514>.
- 139 Orr, A.A., Tao, A., Guvench, O., and MacKerell, A.D. Jr., (2023). Site identification by ligand competitive saturation-biologics approach for structure-based protein charge prediction. *Mol. Pharmaceutics.* 20 (5): 2600–2611. <https://doi.org/10.1021/acs.molpharmaceut.3c00064>.
- 140 Aytenfisu, A.H., Deredge, D., Klontz, E.H. et al. (2021). Insights into substrate recognition and specificity for IgG by endoglycosidase S2. *PLoS Comput. Biol.* .
- 141 Sudol, A.S.L., Butler, J., Ivory, D.P., and Crispin, M. (2022). Extensive substrate recognition by the streptococcal antibody-degrading enzymes ides and endos. *BioRxiv* <https://doi.org/10.1101/2022.06.19.496714>.
- 142 Trastoy, B., Du, J., Cifuentes, J. et al. (2022). Mechanism of antibody-specific deglycosylation and immune evasion by streptococcal IgG-specific endoglycosidases. *Res. Square* <https://doi.org/10.21203/rs.3.rs-1774503/v1>.
- 143 Chong, G. and MacKerell, A.D. Jr., (2022). Spatial requirements for itam signaling in an intracellular natural killer cell model membrane. *Biochim. Biophys. Acta Gen. Subj.* 1866 (11): 130221.
- 144 Vanommeslaeghe, K., Hatcher, E., Acharya, C. et al. (2010). Charmm general force field: a force field for drug-like molecules compatible with the charmm all-atom additive biological force fields. *J. Comput. Chem.* 31 (4): 671–690.
- 145 Vanommeslaeghe, K. and MacKerell, A.D. Jr., (2012). Automation of the charmm general force field (Cgenff) I: bond perception and atom typing. *J. Chem. Inf. Model.* 52 (12): 3144–3154.
- 146 Vanommeslaeghe, K., Raman, E.P., and MacKerell, A.D. Jr., (2012). Automation of the charmm general force field (Cgenff) II: assignment of bonded parameters and partial atomic charges. *J. Chem. Inf. Model.* 52 (12): 3155–3168.
- 147 Yu, W., He, X., Vanommeslaeghe, K., and MacKerell, A.D. Jr., (2012). Extension of the charmm general force field to sulfonyl-containing compounds and its utility in biomolecular simulations. *J. Comput. Chem.* 33 (31): 2451–2468.

- 148 Soteras Gutierrez, I., Lin, F.Y., Vanommeslaeghe, K. et al. (2016). Parametrization of halogen bonds in the charmm general force field: improved treatment of ligand-protein interactions. *Bioorg. Med. Chem.* 24 (20): 4812–4825.
- 149 Guvench, O., Hatcher, E., Venable, R.M. et al. (2009). Charmm additive all-atom force field for glycosidic linkages between hexopyranoses. *J. Chem. Theory Comput.* 5 (9): 2353–2370.
- 150 Klauda, J.B., Venable, R.M., Freites, J.A. et al. (2010). Update of the charmm all-atom additive force field for lipids: validation on six lipid types. *J. Phys. Chem. B* 114 (23): 7830–7843.
- 151 Raman, E.P., Guvench, O., and MacKerell, A.D. (2010). Charmm additive all-atom force field for glycosidic linkages in carbohydrates involving furanoses. *J. Phys. Chem. B* 114 (40): 12981–12994.
- 152 Denning, E.J., Priyakumar, U.D., Nilsson, L., and Mackerell, A.D. Jr., (2011). Impact of 2'-hydroxyl sampling on the conformational properties of Rna: update of the charmm all-atom additive force field for Rna. *J. Comput. Chem.* 32 (9): 1929–1943.
- 153 Guvench, O., Mallajosyula, S.S., Raman, E.P. et al. (2011). Charmm additive all-atom force field for carbohydrate derivatives and its utility in polysaccharide and carbohydrate-protein modeling. *J. Chem. Theory Comput.* 7 (10): 3162–3180.
- 154 Best, R.B., Zhu, X., Shim, J. et al. (2012). Optimization of the additive charmm all-atom protein force field targeting improved sampling of the backbone Φ , Ψ and side-chain X1 and X2 dihedral angles. *J. Chem. Theory Comput.* 8 (9): 3257–3273.
- 155 Hart, K., Foloppe, N., Baker, C.M. et al. (2012). Optimization of the charmm additive force field for DNA: improved treatment of the Bi/Bii conformational equilibrium. *J. Chem. Theory Comput.* 8 (1): 348–362.
- 156 Mallajosyula, S.S., Guvench, O., Hatcher, E., and MacKerell, A.D. (2012). Charmm additive all-atom force field for phosphate and sulfate linked to carbohydrates. *J. Chem. Theory Comput.* 8 (2): 759–776.
- 157 Huang, J., Rauscher, S., Nawrocki, G. et al. (2017). Charmm36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* 14 (1): 71–73.
- 158 Durell, S.R., Brooks, B.R., and Ben-Naim, A. (1994). Solvent-induced forces between two hydrophilic groups. *J. Phys. Chem.* 98 (8): 2198–2202.
- 159 Bujacz, A. (2012). Structures of bovine, equine and leporine serum albumin. *Acta Crystallogr. D Biol. Crystallogr.* 68 (Pt 10): 1278–1289.
- 160 Scapin, G., Yang, X., Prosser, W.W. et al. (2015). Structure of full-length human anti-Pd1 therapeutic IgG4 antibody pembrolizumab. *Nat. Struct. Mol. Biol.* 22 (12): 953–958.
- 161 Venturini, D., de Souza, A.R., Caracelli, I. et al. (2017). Induction of axial chirality in divanillin by interaction with bovine serum albumin. *PLoS ONE* 12 (6): e0178597.

- 162 Papadopoulou, A., Green, R.J., and Frazier, R.A. (2005). Interaction of flavonoids with bovine serum albumin: a fluorescence quenching study. *J. Agric. Food Chem.* 53 (1): 158–163.
- 163 Zhao, H., Ge, M., Zhang, Z. et al. (2006). Spectroscopic studies on the interaction between riboflavin and albumins. *Spectrochim. Acta A Mol. Biomol. Spectrosc.* 65 (3): 811–817.
- 164 Cheng, Z. and Zhang, Y. (2008). Spectroscopic investigation on the interaction of salidroside with bovine serum albumin. *J. Mol. Struct.* 889 (1): 20–27.
- 165 Meti, M.D., Nandibewoor, S.T., Joshi, S.D. et al. (2015). Multi-spectroscopic investigation of the binding interaction of fosfomycin with bovine serum albumin. *J. Pharm. Anal.* 5 (4): 249–255.
- 166 Sekula, B., Zielinski, K., and Bujacz, A. (2013). Crystallographic studies of the complexes of bovine and equine serum albumin with 3,5-diiodosalicylic acid. *Int. J. Biol. Macromol.* 60: 316–324.
- 167 Bujacz, A., Zielinski, K., and Sekula, B. (2014). Structural studies of bovine, equine, and leporine serum albumin complexes with naproxen. *Proteins* 82 (9): 2199–2208.
- 168 Castagna, R., Donini, S., Colnago, P. et al. (2019). Biohybrid electrospun membrane for the filtration of ketoprofen drug from water. *ACS Omega* 4 (8): 13270–13278.
- 169 Karush, F. (1950). Heterogeneity of the binding sites of bovine serum albumin. *J. Am. Chem. Soc.* 72 (6): 2705–2713.
- 170 Fasano, M., Curry, S., Terreno, E. et al. (2005). The extraordinary ligand binding properties of human serum albumin. *IUBMB Life* 57 (12): 787–796.
- 171 Velez Rueda, A.J., Benítez, G.I., Sommese, L.M. et al. (2022). Structural and evolutionary analysis unveil functional adaptations in the promiscuous behavior of serum albumins. *Biochimie* 197: 113–120.
- 172 Katchalski-Katzir, E., Shariv, I., Eisenstein, M. et al. (1992). Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Natl. Acad. Sci. U. S. A.* 89 (6): 2195–2199.
- 173 Ameseder, F., Biehl, R., Holderer, O. et al. (2019). Localised contacts lead to nanosecond hinge motions in dimeric bovine serum albumin. *Phys. Chem. Chem. Phys.* 21 (34): 18477–18485.
- 174 Sundaramurthi, P., Chadwick, S., and Narasimhan, C. (2020). Physicochemical stability of pembrolizumab admixture solution in normal saline intravenous infusion bag. *J. Oncol. Pharm. Pract.* 26 (3): 641–646.

Part II

Quantum Mechanics Application for Drug Discovery

6

QM/MM for Structure-Based Drug Design: Techniques and Applications

Marc W. van der Kamp and Jaida Begum

University of Bristol, School of Biochemistry, Cantock's Close, Bristol BS8 1TS, United Kingdom

6.1 Introduction

Hybrid quantum mechanics/molecular mechanics (QM/MM) modeling is a multi-scale modeling concept whereby a molecular system is divided into two different regions. One (typically small) “active” region is described using a quantum mechanical method, so that electronic changes can be captured, including polarization, charge transfer, and bond rearrangement. Another region, typically large and surrounding the smaller QM region, is treated using an MM force field. This allows a computationally efficient but detailed description of, for example, biomolecular macromolecules (e.g. enzymes/receptors and the surrounding solvent) with their small molecule ligands (e.g. substrates, inhibitors). QM/MM approaches, therefore, have many potential applications in computational biomolecular science, including those relevant to drug design, such as the calculation of spectroscopic properties [1] and pK_a values [2, 3], scoring and pose refinement in small molecule docking (see Section 6.2.2), and prediction of ligand binding affinities [4]. In particular, however, QM/MM has emerged as a reliable computational method to investigate chemical reactions in enzymes [5, 6]. Indeed, in the seminal work from 1976 that introduced the QM/MM concept [7], Warshel and Levitt applied it to perform energy calculations on an enzyme reaction mechanism (hen egg white lysozyme). Several years later, the QM/MM approach was then used for geometry optimization [8], and soon after also for QM/MM molecular dynamics simulation [9], and applied to study enzyme reactions [10].

Since those early days, QM/MM calculations have become increasingly popular, especially for applications on enzymes and indeed simulations of enzyme reactions [5, 11], with over 100 research articles that detail the use of QM/MM methods for enzymes appearing each year since 2007. Such studies have addressed fundamental general questions regarding enzyme catalysis, questions regarding the precise nature of enzymatic reaction mechanisms, and revealed differences in the efficiency of catalysis between enzyme variants (e.g. related to enzyme design), among other insights. Naturally, when such studies are performed on enzymes that are drug

targets, the QM/MM studies can assist structure-based drug design by gaining detailed mechanistic knowledge (e.g. key catalytic interactions in active sites), especially for the transient chemical transition states, which can inspire the design of tight-binding ligands. In addition, QM/MM studies can also provide insights into drug activation or breakdown, enzyme-mediated adverse reactions, and the effectivity of so-called warheads for covalent inhibitors. Over the past decades, several reviews have highlighted the use of QM/MM in relation to drug design and development [12–16], including the use of QM/MM studies that aided the synthesis of new covalent inhibitors [17].

In this chapter, we first introduce the QM/MM approach, alongside QM/MM modeling methods that can be used for modeling protein–ligand interactions as well as reactions. Then, we highlight examples of where QM/MM studies have provided insights into existing covalent drugs as well as covalent inhibitors with potential as drugs in several important therapeutic areas, including cancer treatments based on tyrosine kinase inhibition, emerging resistance of bacteria against treatment with β -lactam antibiotics, and potential treatments with covalent inhibitors of viral infections such as SARS-CoV-2.

6.2 QM/MM Approaches

6.2.1 Combined Quantum Mechanical/Molecular Mechanical Energy Calculations

The principle of QM/MM approaches is to treat a small part of the system, the QM region (e.g. a drug and key surrounding residues in the binding site), with a quantum mechanical (QM) method (describing the electronic structure of molecules), and the rest of the system with a molecular mechanical (MM) method (where interactions between atoms are described using a simple potential energy function, a “force field,” with no direct inclusion of electrons). The QM treatment then allows modeling of the electronic rearrangements, including polarization, charge transfer, and importantly, the breaking and making of chemical bonds. Simultaneous inclusion of the environment with a more computationally efficient MM treatment allows consideration of the influence of this environment on the QM region, e.g. including polarization and effects on the energetics of reactions. It is also possible to model the latter without directly representing the electronic structure, such as with the empirical valence bond (EVB) approach [18, 19] (which combines MM representations of different states along a reaction) and is sometimes also described as a QM/MM approach, but will not be discussed here. There are several reviews where QM/MM methods as applied to enzymes have been described in detail (e.g. Ref. [6]). Here, we give a brief overview of the main approaches and considerations.

The most commonly employed QM/MM approach uses an additive scheme, where the total energy of the system is calculated as:

$$E = E_{QM}(\mathbf{R}_{QM}, \mathbf{R}_{MM}) + E_{QM/MM}(\mathbf{R}_{QM}, \mathbf{R}_{MM}) + E_{MM}(\mathbf{R}_{MM}) \quad (6.1)$$

where E_{QM} is the energy of the QM subsystem $\mathbf{R}_{QM}$ polarized by the MM environment, E_{MM} is the MM force field energy for $\mathbf{R}_{MM}$ and $E_{QM/MM}$ is the interaction energy between the QM and MM parts. The first two terms in Eq. (6.1) depend on both the QM and MM parts. In the first term, the wave function (or electronic density in the case of density functional theory [DFT]) is obtained in the presence of MM point charges. The second term includes the electrostatic force of the QM electron density that is acting on the MM atoms. Van der Waals interactions between the QM and MM regions are typically modeled through the use of a (standard) Lennard-Jones function, with QM atoms being assigned MM parameters for this purpose. When using polarizable force fields, it is further possible to allow polarization of the MM environment by the QM region [20], although this is not yet commonly employed. The additive approach was implemented in the CHARMM simulation package already in the late '80s [21], and in Amber in the late '00s [22], including “internal” support for several popular semi-empirical QM methods, as well as interfaces to existing QM programs. Additive QM/MM calculations can also be performed with Schrödinger’s QSite program [23] (using Jaguar for QM calculations) [24]. More recently, implementations of additive QM/MM have also been developed for nanoscale molecular dynamics (NAMD) [25] and Gromacs [26]. In addition, many QM programs include options to perform QM/MM calculations with the additive approach, such as NWChem [27], ORCA [28], and CP2K [29], and independent interface programs that can combine various QM and MM software packages also exist, such as ChemShell [30, 31], PUPIL [32], and QMCube (QM³) [33].

An alternative to the additive QM/MM approach is a subtractive scheme, where the total energy of the system is described using:

$$E = E_{QM}(\mathbf{R}_{QM}) + E_{MM}(\mathbf{R}_{QM}, \mathbf{R}_{MM}) - E_{MM}(\mathbf{R}_{QM}) \quad (6.2)$$

where $E_{MM}(\mathbf{R}_{QM}, \mathbf{R}_{MM})$ indicates the energy of the total system at the MM level. Note that here, the QM region is also described by an MM force field, and thus the choice of suitable MM parameters (including for different states in a reaction) should be an important consideration. This scheme is employed by the ONIOM method [34] implemented in Gaussian, for example. Details and applications of this subtractive approach, including the important difference between so-called “mechanical embedding” and “electrostatic embedding,” are discussed by Ryde [35]. Notably, “electrostatic embedding,” where the QM system is polarized by the MM partial charges, is important to include the influence of the biomolecular environment (and therefore for calculations of reasonable accuracy, unless very large QM regions are used). As indicated above, implementations of the additive QM/MM scheme essentially always include such “electrostatic embedding.”

With an expression for the energy of a QM/MM system (as in Eq. 6.1 or Eq. 6.2), it is now possible to calculate gradients and forces for all atoms, and thus perform energy minimization and molecular dynamics simulation (using the Born-Oppenheimer approximation, i.e. movement of the QM atoms is propagated classically). Thus, many “standard” simulation techniques now become available, with the benefit that covalent bond making and breaking, alongside other electronic effects, can be included in the QM region.

6.2.2 QM/MM Methods for the Evaluation of Non-Covalent Inhibitor Binding

The potential benefits of more accurate inclusion of ligand descriptions and, in particular, polarization effects in small molecule-target binding have led to the development of QM/MM approaches to help predict binding conformation, evaluate binding energies, and even correct calculations of drug binding kinetics based on MM molecular dynamics (MD) simulation [36]. For small molecule docking, QM/MM methods can aid in the scoring of previously generated poses, refinement of such poses, as well as pose generation itself. For recovering native poses from co-crystal structures, Cho et al. demonstrated that (re)calculating ligand charges based on QM (B3LYP/6-31G*) in the MM protein environment can lead to significantly higher accuracy in predicted binding modes [37]. More recently, a similar approach to include QM/MM calculations in the docking protocol, with both the ligand and residues in the binding site treated QM, was shown to improve results in tests on GPCR co-crystal structures [38]. Ligand charges derived from QM calculations in the (MM) protein environment were also shown to improve both docking geometry and scoring for protein–ligand complexes that involve halogen bonding [39]. Further, a version of the attracting cavities docking approach with on-the-fly QM/MM calculations (using the efficient SCC-DFTB QM method) showed clear benefits for metalloenzymes: significant improvements were obtained for zinc metalloproteins and heme proteins (the latter with covalent binding between ligand and heme iron) [40]. Docking pose refinement (after standard generation) using short MM MD simulations and QM/MM minimization was also shown to improve pose accuracy and ranking, including for the heme protein cytochrome c peroxidase [41].

For a detailed evaluation of relative binding energies in small molecule series (e.g. in lead optimization), methods involving conformational sampling of the protein–ligand system are often used. In alchemical free energy perturbation (FEP) simulations, the application of QM/MM is not straightforward due to the nonphysical intermediate states involved with, e.g. insertion or deletion of atoms. However, applying an MM to QM/MM free energy correction to the end states (sometimes referred to as a “reference potential” or “book-ending” approach) is perfectly feasible, and methods continue to be actively developed and tested [42–46]. Early tests based on a structurally diverse set of fructose 1,6-bisphosphatase inhibitors indicate that such QM/MM-based free-energy perturbation approaches have potential for structure-based drug design [47].

The popular MM-PB(GB)SA end-point method is commonly applied based on molecular dynamics sampling of protein–ligand complexes. It provides an intermediate in terms of accuracy and computational resource between docking-based scoring and alchemical FEP [48]. After the (MM-based) sampling, binding energy calculations can make use of QM/MM, resulting in a QM/MM-PB(GB)SA approach, which can improve the resulting binding affinity ranking. For example, for ranking of inhibitors of polo-like kinase 1, it was found that when ligands were ranked using QM/MM-GBSA using semi-empirical QM methods instead of MM-GBSA, the ranking of congeneric compounds was significantly improved [49]. An

alternative to QM/MM-GBSA or -PBSA, with a similar philosophy of incorporating polarization, etc., is the application of QM calculations in conjunction with a continuum solvent model following MM MD sampling. Examples include the so-called MM/QM-COSMO method [50], which was applied to obtain insights into phosphopeptide binding to the BRCA1 C-terminal domain [51]. A similar approach, named SQM/COSMO [52], applies semi-empirical QM calculations and was recently shown to be sufficiently computationally efficient to be used in virtual screening, improving on other scoring functions [53].

The above-mentioned QM polarized ligand docking approach originating from Cho et al. [37] has been implemented in Schrödinger's Glide software as QM-polarized ligand docking (QPLD) [54], which has helped popularize this approach. The combination of QPLD with QM/MM-PBSA scoring (without further MD-based conformational sampling) has been shown to be advantageous, for example, in correctly ranking and obtaining insight into halogen-substituted ligands targeting phosphorylase kinases [55]. QM/MM in rescoring can be advantageous also when standard (MM-based) docking (and refinement) is performed. For example, for the identification of Myt1 kinase inhibitors, QM/MM-GBSA rescoring was shown to improve upon docking scores or MM-GBSA rescoring in the classification of active and inactive inhibitors [56]. The predictive power of such rescoring was indicated by testing five compounds predicted as active, of which three showed significant inhibition of Myt1. Further examples demonstrate the direct use of (QM/MM) pose refinement and QM/MM-PBSA scoring in guiding the synthesis of inhibitors of human *O*-GlcNAc hydrolase [57] (a target for Alzheimer's disease) and glycogen phosphorylase [58] (a target for type 2 diabetes).

As described in this section, the ability of QM/MM methods to capture polarization (and charge transfer) effects in protein–ligand interactions is often beneficial for the evaluation of protein–ligand binding. It is possible that the continuing active development of polarizable force fields [59], including automation of small molecule parameterization for such force fields [60, 61], will yield similar fundamental advantages as QM/MM or QM-based binding energy evaluation approaches [62], i.e. a more accurate description of polarization, at a significantly lower computational cost. Nevertheless, QM(/MM) approaches may still be required when electronically complex phenomena, such as sigma holes, cation- π interactions, or metal coordination, are important [63], as these remain challenging also for polarizable force field approaches.

6.2.3 QM/MM Reaction Modeling

Modeling of reactions in biomolecular systems can yield important structural insights (e.g. transition state structures, which can provide inspiration for effective enzyme inhibitors) and energetic insights (e.g. related to differences in reactivity within a series of covalent drug leads). For QM/MM modeling of reactions in biomolecular systems, typically either a QM/MM potential energy profile or minimum energy pathway (MEP) is (first) obtained, or QM/MM molecular dynamics simulations are performed, biased along the reaction path, to obtain free energy

profiles. Many techniques are available for both options. In this section, we will briefly describe some of the more commonly applied techniques for biomolecular systems and related considerations.

As for small molecule systems, a potential energy profile (or MEP) can be obtained by minimizing the energy of the system at several points along a reaction coordinate (or “collective variable”). This reaction coordinate could be based directly on (a combination of) geometric features describing bond making and breaking, or methods that optimize an MEP based on providing initial (reactant) and final (product) states can be used. The latter includes Nudged Elastic Band optimization (with the Climbing Image variant being able to provide a “true” transition state, the saddle point on the potential energy surface), which has been adapted specifically for QM/MM [64, 65], and the similar Replica Path method [66]. (A brief overview of several reaction path methods and their application in biomolecules is included in Ref. [21].) Due to the large configuration space available in biomolecular systems, one should be aware of changes in conformations along the optimized reaction pathway, which may not be directly related to the reaction. For example, when relying on iterative minimizations along a geometric reaction coordinate, sudden “jumps” in the QM/MM potential energy can be caused by a small change (e.g. rotation of a water molecule or side chain) further away from the active site where bond changing takes place. To avoid this from happening, part of the MM region could be constrained, and minimizations backwards and forwards along the reaction may resolve such jumps. Notably, due to the many possible conformations available to biomolecular (e.g. drug-target) systems, a single optimized reaction path (or a single set of stationary points optimized along this path) may not be representative, or indeed not allow for a confident prediction of differences between related systems (e.g. different covalent drugs, enzyme variants). Indeed, many different starting conformations may need to be considered (and reaction energies exponentially averaged) to obtain converged reaction barrier energies [67]. The direct output from optimized reaction paths, alongside structures, will be the QM/MM potential energies. Hybrid DFT functionals are suitable for high-accuracy structure optimization, which can then allow energies to be calculated with “gold standard” wavefunction methods, such as CCSD(T) or variants thereof. Performing such single-point energy calculations is also popular for correcting energy profiles obtained with more approximate methods (e.g. semi-empirical QM corrected by hybrid DFT). To estimate *free* energy profiles from minimum energy paths (beyond estimating entropic contributions through frequency calculations of optimized stationary points) one can use QM/MM-FEP approaches, which incorporate conformational sampling of the MM environment. These range from approximate to more sophisticated, e.g. from keeping the QM region completely fixed, to re-introducing some sampling in the QM region [68, 69], to re-optimizing the reaction path [70].

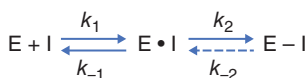
With QM/MM molecular dynamics simulations, conformations can be directly sampled along the reaction path by using enhanced sampling techniques (without the need of calculating a potential energy profile). Due to the need to compute forces every femtosecond, these simulations are significantly more computationally demanding than optimizing potential energy profiles and thus may require the

use of semi-empirical QM methods or more limited sampling (e.g. with DFT functionals). A commonly used enhanced sampling approach is umbrella sampling, where (mild) restraints are applied to particular reaction coordinate values (or “windows”) along the reaction. The sampling can then be subsequently unbiased using statistical methods to obtain the potential of mean force (or free energy profile). Popular methods for this include the well-known Weighted Histogram Analysis Method [71], Umbrella Integration [72], and the more recently developed Dynamic Histogram Analysis Method [73, 74]. The position along the reaction coordinate can be determined by sampling along one or two reaction coordinates (generating 1D or 2D free energy profiles along these) or by approaches that “dynamically” optimize the minimum free energy path, such as finite-temperature string methods [75–78]. The advantage of the latter approach is that, in principle, a minimum free energy path can be determined in the space of a large number of interdependent collective variables (distances, angles, etc.). A path collective variable combining several relevant degrees of freedom for the reaction process can then be used, which allows multi-event mechanisms with (partially) concerted pathways to be captured accurately. Other enhanced sampling approaches commonly used for QM/MM reaction simulations are steered MD [79] and metadynamics [80]. The latter is also popular in combination with Car-Parinello MD (CP-MD) in a QM/MM scheme for enzyme reaction simulations. CP-MD provides an alternative to the common “Born-Oppenheimer MD” approach, where the QM energy and forces come from a converged SCF calculation in each step. Instead, in CP-MD, wave functions are treated as fictitious dynamic variables and follow the motion of the nuclei “on the fly.” In the latter, a plane-wave basis is (almost always) used together with a full DFT functional (e.g. PBE, BLYP), which may limit the accuracy. QM/MM modeling of enzyme reactions with CP-MD was established in the 2000s, and an efficient implementation (with GROMACS for the MM part) is now available via MiMiC [81].

Once free energy profiles of enzyme reactions are obtained, the resulting reaction barrier can, in principle, be directly compared to an (apparent) activation free energy obtained from an experimental apparent first-order reaction rate constant (k_{cat} , or similarly, $k_{\text{off}}/k_{\text{on}}$ rates if these are dominated by the chemical reaction studied), using transition state theory (TST) [82]. Most approaches for QM/MM reaction modeling will have some shortcomings to arrive at highly accurate reaction barriers, such as limited conformational sampling or limited accuracy of the QM method employed. One strategy is to model the QM region using cheap, approximate methods (e.g. semi-empirical QM) during conformational sampling, and then correct the resulting barrier through the use of high-level QM(/MM) calculations. For such an approach, one should ensure that the more approximate methods sample the same mechanism [83, 84]. In favorable cases, QM/MM simulations can thereby calculate enzyme reaction barriers in quantitative agreement with experimental kinetics [85]. It is important to note, however, that qualitative agreement for the reaction barrier is often sufficient to distinguish between alternative mechanisms, determine differences in reactivity for drug series, or assess the effect of mutations; differences between barriers for comparable systems are often still captured accurately (see, e.g. Section 6.3.2).

6.3 Applications of QM/MM for Covalent Drug Design and Evaluation

Historically, the pharmaceutical industry was reluctant to focus its efforts on covalent drug design [86]. Given that toxicity (together with efficacy) is the leading cause of drug attrition in clinical development [87], covalent inhibitors were disfavored due to safety concerns: covalent binding may occur on off-target proteins. However, attitudes toward covalent drugs have changed in the past 3 decades, and the FDA has approved over 50 covalent drugs to date [88, 89]. The majority of these are treatments for cancer as well as bacterial and viral infections [89]. In this section, we discuss examples of the use of QM/MM simulations for obtaining a detailed understanding of key targets and their covalent inhibitors, which can be used in the development of covalent drugs. These calculations can provide information on the non-covalent interactions as well as the formation and breakdown of covalent bonds between the drug and its target. Many covalent inhibitors have two distinct motifs: one for providing specific recognition of the target through non-covalent interactions and one that provides reactivity to form the covalent link, commonly referred to as the “warhead.” Related to this, the binding process can be described using a two-step model, where first an enzyme-inhibitor non-covalent complex is established, before the covalent link is formed (Scheme 6.1). The chemical step leading to covalent inhibition can be described either as reversible or irreversible, depending on the magnitude of rate k_{-2} , i.e., the free energy barrier for bond breaking. To aid the design of covalent inhibitors, covalent docking is already a recognized tool [89], useful, for example, as a high-throughput screening method and to suggest binding modes (see Chapter 25). For more detailed insights into the energetics and the process of covalent drug binding, atomistic simulations can be used. Whereas the energetics of non-covalent complex formation can be captured by MM (or QM/MM) approaches, differences in the rate constants for covalent complex formation will typically require QM/MM reaction simulations.



Scheme 6.1 General two-step model for covalent inhibition.

E = enzyme, I = covalent inhibitor, $E \cdot I$ = non-covalent complex, $E - I$ = covalent complex. In addition to the non-covalent steps (i.e. k_1 and k_{-1}), the rate constant k_2 is relevant for irreversible covalent inhibition, whereas the balance between k_2 and k_{-2} determines reversible covalent inhibition.

6.3.1 Covalent Tyrosine Kinase Inhibitors for Cancer Treatment

Receptor tyrosine kinases (RTKs) are key regulators of cellular processes and can thereby govern cell growth and metastasis in cancer [90]. It is therefore no surprise that inhibitors of RTKs have been actively developed over the past decades. Several covalent inhibitors, binding to a cysteine in the RTK active site, have been approved

as drugs and are in clinical use for the epidermal growth factor receptor (EGFR, e.g. afatinib, neratinib, osimertinib) and Bruton tyrosine kinase (BTK, e.g. acalabrutinib, ibrutinib). In this section, we discuss how QM/MM modeling has provided insights relevant for covalent drug design and understanding emerging resistance for these two RTKs, after introducing some of the biological context.

The binding of cognate ligands at the extracellular site of the transmembrane protein EGFR can trigger signaling networks leading to cellular proliferation, differentiation, and survival [91]. Mutations in EGFR, which can cause it to be in a prolonged state of activation [92], are associated with various types of cancer, including non-small-cell lung cancer (NSCLC), which is responsible for 85% of all lung cancers [93]. Interfering with EGFR signaling using small-molecule inhibitors is therefore an appealing strategy for oncogenic treatment. First-generation ATP-competitive inhibitors for EGFR, such as gefitinib and erlotinib, were developed for NSCLC. While initial treatments were positively received, secondary drug resistance mutations, such as T790M, significantly reduced the clinical efficacy of these reversible non-covalent drugs [94, 95]. This led to the development of second-generation inhibitors such as afatinib and dacomitinib to inhibit EGFR T790M [96]. These compounds contain a distinctive electrophilic acrylamide warhead, which reacts with the nucleophilic Cys797 in the ATP-binding site to form a covalent adduct [97, 98]. Although these drugs can arrest EGFR T790M activity, selectivity issues arise as reactions also occur with Cys797 in wild type (WT) EGFR, leading to unwanted side effects.

Soon after detailed kinetic investigations of several covalent EGFR inhibitors were reported [97], Capoferri et al. [99] investigated the reaction mechanism for a prototypical irreversible covalent inhibitor with an acrylamide warhead against WT EGFR. In their approach, QM/MM (SCC-DFTB/ff99SB) MD umbrella sampling with a path collective variable was used to simulate the mechanism of covalent binding to the targeted Cys797. The simulations revealed that Asp800 likely acts as a general acid/base catalyst. Once it deprotonates Cys797, a concerted mechanism occurs in which the C α of the acrylamide inhibitor is protonated by Asp800, with concomitant formation of the saturated β -substituted product. The covalently bound product was calculated to be ~ 12 kcal/mol more stable than the non-covalent complex, emphasizing that the binding is spontaneous and exergonic, consistent with experimental findings [97, 100, 101]. The semi-empirical SCC-DFTB method, known to typically underestimate reaction barriers [102–104], predicted a reaction barrier of 14.6 kcal/mol, lower than the barrier of ~ 20 kcal/mol derived from the experimental rate [97]. The authors further concluded that desolvation of the Cys797 thiolate is key, suggesting that intrinsic acrylamide reactivity should only have a minor impact on the potency of inhibitors. Overall, this study represents the likely mechanism and energetics involved in second-generation EGFR inhibitors, which are strongly exergonic upon binding (i.e. k_2 is greater than k_{-2} , see Scheme 6.1) [105]. The issue with such reactive, irreversible inhibitors is that specificity for the cancer-related EGFR variant is often low, leading to toxicity. For this reason, third-generation EGFR inhibitors use a more weakly reactive warhead, leading to a reversible process (k_{-2} is not negligible). Indeed, such third-generation

inhibitors as osimertinib, lazertinib, and rociletinib do essentially not target WT EGFR [106, 107]. Unfortunately, however, further EGFR mutations can confer resistance to these third-generation inhibitors. One such mutation is L718Q [108], but the precise molecular mechanism by which glutamine at position 718 reduces the efficacy of osimertinib, for example, was not known. This motivated the work of Callegari et al. [105], where a combination of QM/MM reaction simulations, binding affinity calculations (using the WaterSwap method [109]), and MM MD simulations were used to investigate possible effects on the Cys797 pK_a , the reaction mechanism, binding affinity, and non-covalent conformational dynamics of osimertinib. Throughout, comparison between EGFR T790M and EGFR T790M/L718Q was performed to identify which properties were significantly affected by the L718Q mutation (Figure 6.1). Initially, both approximate pK_a calculations on MM MD

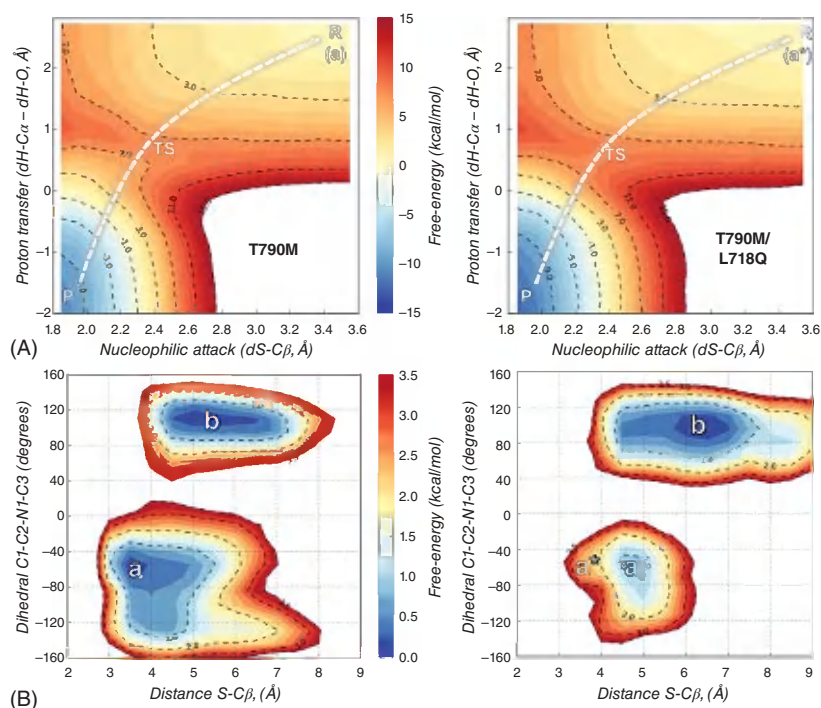


Figure 6.1 Assessment of osimertinib interactions with EGFR T790M (left) and EGFR T790M/L718Q (right). (A) Free energy surfaces based on 2D QM/MM (SCC-DFTB/ff99SB) umbrella sampling simulations indicate similar reaction barriers between the non-covalent reactant (R) and the covalently bound (Cys797 alkylated) state (P). The proton transfer reaction coordinate is defined as $d(H_{Asp800}-C\alpha_{acrylamide}) - d(H_{Asp800}-O_{Asp800})$ and nucleophilic attack as $d(S_{Cys797}-C\beta_{acrylamide})$. (B) 2D free energy surfaces in the space of $d(S_{Cys797}-C\beta_{acrylamide})$ and the acrylamide $C_1-C_2-N_1-C_3$ dihedral angle. Free energies are calculated from the frequency distribution of conformations observed in four independent 300 ns MM MD simulations for each complex. For EGFR T790M/L718Q, **a*** represents the region of reactive conformations of osimertinib (which approximately corresponds to basin **a** for EGFR T790M). Source: Reproduced with permission of Callegari et al. [105]/ Royal Society of Chemistry / Public Domain CC BY 3.0.

snapshots and QM/MM (SCC-DFTB/ff99SB) umbrella sampling simulations of the Cys797-S⁻/Asp800-COOH to Cys797-SH/Asp800-COO⁻ reaction were performed to ascertain that the resistance to osimertinib is not due to an increase in the Cys797 pK_a. Equally, the Cys-alkylation reaction energetics are not altered significantly, as indicated by the energy profiles obtained through 2D QM/MM umbrella sampling (Figure 6.1A and B). The binding affinity of osimertinib prior to the reaction was also not affected. However, MM MD simulations (4x 300 ns) revealed that there is a significant difference in the preferred binding conformation of osimertinib in the EGFR active site. In EGFR T790M, configurations in which the nucleophile (Cys797 S) and electrophile (acrylamide C β) are within reacting distance occur for >20% of the simulation time. This reduces to only 0.7% for EGFR T790M/L718Q, representing a free energy difference of ~3 kcal/mol (Figure 6.1A and B). This is primarily due to the stabilization of an alternative conformational “basin” (b), through hydrogen bonding of the Gln718 side-chain -NH₂ moiety with the acrylamide carbonyl. The combination of methods thus allows a detailed explanation of how the L718Q mutation leads to osimertinib resistance. This highlights how detailed multiscale modeling can help rationalize the effects of a mutation in the binding site on interactions with a covalent drug.

Not only can QM/MM simulations (combined with other MM modeling techniques) be used to obtain detailed insight into the origins of resistance to existing drugs in EGFR variants, but they can also contribute to the design of new covalent drugs for EGFR. For example, Castelli et al. used QM/MM simulations in their detailed study comparing heteroarylthioacetamide derivatives as EGFR inhibitors [110]. The main aim was to select alternative (weaker) warheads than acrylamide, which may help design selective inhibitors. To compare warhead effectivity, QM/MM simulations were performed to compare a chloroacetamide and two different 2-(imidazol-2-ylthio)acetamide derivatives as warheads. QM/MM (PDDG-PM3/ff99SB) steered MD was used to obtain approximate free energy profiles for the Cys797 alkylation reaction, with Cys797 and Asp800 (alongside the inhibitor) in the QM region. Initial simulations indicated that protonation of the imidazole ring, which is in close proximity to Asp800, is required to form a stable Cys797 alkylation product. The subsequent free energy profiles (obtained using the Jarzynski equality) suggested that only one of the two 2-(imidazol-2-ylthio)acetamide derivatives, without N-methylation on the imidazole, led to an energetically favorable reaction (similar to that for the chloroacetamide warhead). The simulations were thus used to prioritize the synthesis and testing of an EGFR-targeting compound with this warhead. This compound showed promising results in several different assays (with negligible reactivity with cysteine in solution), indicating that the imidazol-2-ylthioacetamide warhead could be used for more selective targeting of cysteines in kinases.

Bruton's tyrosine kinase (BTK) is another RTK that plays an essential role in multiple signaling pathways [111]. In the B-cell receptor signal transduction pathway, for which BTK's function is best understood [112], it is responsible for normal B-cell survival and proliferation. Downregulation of BTK activity causes immunodeficiency diseases [113], whereas overexpression of BTK leads to

autoimmune disorders [114] and various blood cancers [115]. Targeting BTK using small molecule inhibitors is therefore a therapeutic strategy for treating pathologies involved in the upregulation of BTK [116]. The first BTK inhibitor developed for the treatment of multiple B-cell cancers was ibrutinib [117]. Ibrutinib irreversibly inactivates BTK through its acrylamide warhead, which is responsible for binding covalently to Cys481 in the active site. Unlike EGFR, in which an aspartate base activates the nucleophilic cysteine residue [99], BTK does not have a suitable residue to act as a base in close proximity to the targeted cysteine. Hence, the mechanism of covalent binding is different, but not known, which prompted Voice et al. [118] to investigate this using QM/MM simulations. The semi-empirical DFTB3 method was used for the QM region, after benchmarking indicated that this method can predict reaction pathways for thiol addition mechanisms that are structurally in close agreement with higher-level methods such as ω B97-XD and MP2. QM/MM MD umbrella sampling at the DFTB3/ff14SB level was then used to investigate several possible covalent bond formation mechanisms of ibrutinib with BTK. This indicated that the most plausible mechanism, in good agreement with experimental kinetics [119], proceeds through three key steps: (i) activation of the nucleophilic Cys481 by the carbonyl oxygen atom of the Michael acceptor, (ii) enol-intermediate formation as a result of the nucleophilic attack of the cysteine thiolate onto the electrophilic warhead, and (iii) solvent-assisted tautomerization from the enol to the covalently bound keto product. The final step was indicated to be rate-limiting in BTK, with a water molecule present for solvent-assisted tautomerization. The reaction free energy of ~ -37 kcal/mol is consistent with the irreversible inhibition of BTK by ibrutinib. The authors further note that changing the heterocyclic core and/or warhead should not improve the covalent reactivity (i.e. k_{inact} rates). Voice et al. [118] thus suggest that an alternative strategy for tuning reactivity and increasing specificity could be to design inhibitors that modulate the Cys481 pK_{a} and/or its conformational behavior. This emphasizes the importance of considering the influence of the protein environment, as this may govern the reactivity of inhibitors. Indeed, Voice et al. previously showed that although ligand-only metrics, such as predicted proton affinity and reaction energies, work well for predicting the reactivity of small reactive fragments (e.g. acrylamide warheads), they fail for larger drug-like compounds [120]. This was attributed to the lack of such metrics to incorporate binding conformations, solvation, and intermolecular interactions.

Awoonor-Williams and Rowley [121] characterized both the non-covalent and covalent components of the covalent binding of a *t*-butyl inhibitor with a cyanoacrylamide warhead to BTK, using a combination of methods. First, they used constant-pH MD to predict the pK_{a} of Cys481, which defines the cost of thiolate formation (avoiding having to model a specific deprotonation mechanism). Second, absolute binding free energies for the non-covalent interactions were calculated using alchemical free-energy perturbation at the MM level, indicating that Van der Waals dispersion plays an important role. Conformational energies were also factored in, as the ligand can adopt multiple states in the bulk solution compared to when it is bound to the enzyme. Third, covalent bond formation was simulated using QM/MM MD umbrella sampling (ω B97X-D3BJ/def2-TZVP for the

QM region, CHARMM36 for MM), indicating an activation barrier of 3.4 kcal/mol. Finally, subtractive ONIOM QM/MM calculations were used to estimate the relative stability of the covalent adduct compared to the non-covalent complex with Cys481 as a thiol (indicating a ~ 31 kcal/mol difference, consistent with irreversible inhibition). Similar to the work by Callegari et al. on EGFR [105], this is another example where combining several different MM and QM/MM calculations for the relevant processes involved in covalent-drug binding shows promise for their use in the evaluation and development of inhibitors.

6.3.2 Evaluation of Antibiotic Resistance Conferred by β -Lactamases

Antimicrobial resistance (AMR) in bacteria is one of the leading health threats of the twenty-first century, affecting healthcare systems and economies. It has been estimated that in 2019, ~ 1.3 million people died due to bacterial AMR (with ~ 5 million deaths associated with it) [122], a number that was previously projected to soar to 10 million per year by 2050 if current trends continue [122, 123]. Several of the leading pathogenic infections for these AMR-related deaths are caused by Gram-negative bacteria, which are usually treated with β -lactam antibiotics. These antibiotics exert their antibacterial activity by covalently inhibiting penicillin-binding proteins, which are crucial enzymes involved in bacterial cell wall maintenance. The main cause of resistance toward β -lactam drugs in Gram-negative bacteria, which are the most frequently prescribed antibiotics worldwide, is the presence of β -lactamases (BLs). BLs can be split into two main categories: Class A, C, and D are serine BLs (SBLs) and Class B are metallo-BLs. Collectively, they can efficiently catalyze the breakdown of all four classes of β -lactams: penicillins, cephalosporins, carbapenems, and monobactams. To combat β -lactamase-mediated antibiotic resistance, so-called combination therapy is now available for treating complicated infections. This involves co-administering a β -lactamase inhibitor along with a β -lactam antibiotic, thus restoring the antibacterial action. Examples of clinically approved β -lactamase inhibitors include classical inhibitors (with a β -lactam core) sulbactam, tazobactam, and clavulanic acid, and nonclassical inhibitors (without a β -lactam core) avibactam, relebactam, and vaborbactam.

QM/MM simulations have played an important role in establishing the detailed mechanisms of β -lactam antibiotic breakdown in the different β -lactamase classes. The mechanism for all SBLs, similar to serine hydrolases, involves an acylation step to form a covalently bound acyl-enzyme intermediate (where the β -lactam ring is broken), followed by deacylation to release the inactive antibiotic. The details of these steps differ between the classes, however. For the Class A enzymes, QM/MM simulations helped confirm that Glu166 (absent in Class C or D enzymes) is acting as the base in deacylation [124]. This was based on potential energy profiles obtained for the reaction between the archetypal TEM-1 β -lactamase and (benzyl)penicillin, using AM1/CHARMM22 energy minimizations corrected by B3LYP/6-31G+(d). In acylation, there are two feasible options in Class A SBLs for the base that abstracts the proton from Ser70 (which performs the nucleophilic attack to form the covalent acyl-enzyme intermediate): Glu166 and Lys73 (in its neutral state). Initially, findings

based on potential energy surfaces from different groups and techniques led to different conclusions: either Glu166 acts as the base (AM1/CHARMM27 optimizations using the additive QM/MM scheme corrected by single point B3LYP/6-31G(d) calculations) [125, 126], or Lys73 is preferred as the base (ONIOM calculations with structure optimization at the HF/3-21G-OPLS-AA level and energy calculations at the MP2/6-31 + G(d)-OPLS-AA level) [127], with a slightly higher barrier indicated for Glu166. Subsequent additive QM/MM calculations at higher levels of theory (structure optimization with B3LYP/6-31+G(d)-CHARMM27 and energies with SCS-MP2/aug-cc-pVTZ-CHARMM27) [128] indicated that Glu166 was the more likely base to remove the proton from Ser70 (via a bridging water molecule), with Lys73 stabilizing the transition state for this step. However, for the acylation reaction between KPC-2 and the β -lactamase inhibitor avibactam, several different recent QM/MM studies report that, in this case, Lys73 is the likely base [129–131]. It is therefore possible that the nature of the base in Class A SBL acylation differs depending on the enzyme/drug combination.

The mechanisms involved in Class B metallo- β -lactamases are still debated, complicated by the involvement of the (typically) two zinc ions in the active site and uncertainties regarding the detailed structures of on-pathway intermediates, but QM/MM studies have contributed to insights here also, further highlighting the mutual benefits between protein crystallography and QM/MM simulation (see, e.g. discussion in Ref. [5]). In Class C SBLs (less extensively studied than Class A and D SBLs), QM/MM studies on the AmpC enzyme from *Escherichia coli* with cephalotin (metadynamics using CPMD with PBE and plane-wave basis sets, with GROMOS as MM force field) indicate that Lys67 (equivalent to Lys73 in Class A SBLs) acts as the base in acylation instead of Tyr150, which was proposed to be involved in protonation of the β -lactam ring N atom [132]. The same authors also studied deacylation for this system and concluded that Tyr150 performs a similar role as Glu166 in Class A SBLs in this step (as established previously using a different QM/MM approach on the deacylation of penicillin by P99) [133]: it abstracts a proton from the deacylating water (DW) during its nucleophilic attack [134]. Comparison between their two studies led to the prediction that the acylation reaction is the rate-determining step in cephalothin hydrolysis for *E. coli* AmpC. Notably, Class D SBLs feature an unusual carboxylated lysine (Lys73), which was proposed to act as the base in both acylation and deacylation. QM/MM studies based on umbrella sampling simulation at the PM3-PDDG/ff12SB level [135], alongside the inability of Lys73 mutants to support deacylation, helped confirm this.

Establishing the mechanistic details of the reactions involved in β -lactamases is important to gain further understanding of why certain antibiotics are less effective than others and how β -lactamase-mediated resistance against antibiotics can arise. This knowledge can then aid in the further (re)design of antibiotics and β -lactamase inhibitors to help combat such resistance. Once mechanisms are known, QM/MM reaction simulation protocols that correctly capture the difference in β -lactamase efficiency between enzyme-antibiotic combinations can provide such insights in fine atomic detail. For the breakdown of the important carbapenem antibiotics by Class A SBLs, deacylation is expected to be rate-limiting [136].

Chudyk et al. therefore focused on this step and showed how for eight different Class A SBLs, QM/MM MD simulations can correctly distinguish between those that can efficiently break down meropenem and those that cannot [137]. These simulations were based on SCC-DFTB/ff12SB umbrella sampling simulations along two reaction coordinates, one describing the nucleophilic attack of the DW onto the acyl oxygen and another the proton transfer between Glu166 and the DW, thus arriving at the tetrahedral intermediate. Although the semi-empirical SCC-DFTB method (required for the extensive conformational sampling performed in this work) again underestimates the absolute barriers for this reaction, the difference in barriers between carbapenemases (KPC-2, SFC-1, NMC-A, and SME-1) and carbapenem-inhibited SBLs (TEM-1, SHV-1, BlaC, and CTX-M-16) was very clear, consistent with kinetic data. Subsequently, it was shown that sampling can be significantly reduced, by focusing umbrella sampling just on an approximate minimum free energy path and reducing simulation length. This led to a computationally efficient assay to assess the carbapenem hydrolysis efficiency of Class A SBL enzymes [138]. To establish the origins of the difference in carbapenem hydrolysis in these eight enzymes, Chudyk et al. later conducted a detailed analysis and performed further “computational experiments” by simulating the reaction (using the same QM/MM MD protocols) with specific restraints or mutations [139]. This indicated that efficiency for carbapenem hydrolysis by Class A SBLs is influenced by a range of factors, including optimal stabilization by the oxyanion hole, the presence or absence of the active site Cys69-Cys238 disulfide bridge, the orientation of the 6 α -hydroxyethyl group of the carbapenem scaffold (including its interaction with Asn132), and the interaction of Asn170 with Glu166, the base in deacylation. Disrupting any of these factors away from their optimal values can lead to loss of carbapenemase activity, whereas the introduction of single individual factors (e.g. by conformational constraints or mutation) is not sufficient to introduce carbapenemase activity in carbapenem-inhibited Class A SBLs. The identified interactions could be exploited in the development of new β -lactam antibiotics able to evade resistance conferred by Class A carbapenemases. The transferability of the QM/MM MD umbrella sampling simulation protocols was further highlighted by their application to reaction simulations of KPC-2 and TEM-1 acyl enzymes formed by interaction with the classical covalent SBL inhibitor clavulanic acid [140]. This revealed that, of several possible adducts, the decarboxylated *trans*-enamine species is responsible for inhibition.

An alternative approach to using QM/MM MD reaction simulations and analysis to obtain insights into the breakdown efficiency is to obtain many different potential energy profiles and analyze their structural features to determine which acyl-enzyme conformations/interactions lead to efficient breakdown. The analysis of such features can be done using machine learning (ML) approaches, as was recently demonstrated for the deacylation of imipenem by the Class A GES-5 SBL [141], based on 500 semi-automatically generated pathways per reaction, obtained using DFTB3/CHARMM36 optimization and B3LYP-D3/6-31+G**-CHARMM36 energy calculations. Interpretation of the edge-conditioned graph convolutional neural network trained to predict the QM/MM barriers based on the initial acyl

enzyme conformation, helped provide insights into the difference in deacylation efficiency between the Δ^2 and Δ^1 tautomers of the pyrroline ring. As established in the '80s [142, 143] (and assumed in other simulation studies), the Δ^1 tautomer is significantly more stable; hence, deacylation will occur with the Δ^2 tautomer. Among the features that determine the reactivity of Δ^2 tautomer acyl enzyme conformations, the importance of the 6 α -hydroxyethyl orientation for efficient deacylation was again highlighted.

Aside from the Class A carbapenemases described above, another common route for carbapenem resistance is hydrolysis by Class D SBLs. Within this diverse class, OXA-48-like enzymes have emerged as a particular concern for clinical microbiologists [144]. These plasmid-borne SBLs, which are often difficult to detect, typically hydrolyze carbapenems fairly efficiently, while retaining activity against other antibiotic substrates (such as oxacillin) [145]. Differences in the efficiency of carbapenem hydrolysis do exist, however, with OXA-48 (and related variants) showing higher efficiency for hydrolyzing imipenem than meropenem and other 1 β -methyl carbapenems (such as ertapenem and doripenem) [145]. This difference was investigated by Hirvonen et al. using MM MD simulations and QM/MM MD umbrella sampling [146]. MM MD conformational sampling (with ff14SB) of the OXA-48 acyl enzymes with imipenem and meropenem revealed that the 6 α -hydroxyethyl moiety can adopt three distinct orientations. Subsequently, free energy profiles for the deacylation step (formation of the tetrahedral intermediate), known to be rate-limiting, were calculated using DFTB3/ff14SB umbrella sampling with each of the three distinct 6 α -hydroxyethyl orientations, as well as different hydration of the carboxylate group attached to Lys73 (carboxy-Lys73) for each (Figure 6.2a). As expected (and indicated by the benchmarking calculations performed), reaction barriers are underestimated. However, the difference in reaction barriers between imipenem and meropenem ($\Delta\Delta^\ddagger G$) agrees very well with experimental kinetics, which indicate this should be between 1.4 and 3.5 kcal/mol. For the active site conformations with the lowest barriers, the difference is 2.2 kcal/mol (Figure 6.2a), rising to 2.8 kcal/mol when the lower likelihood of sampling the 6 α -hydroxyethyl group in orientation I for imipenem (based on 5x 200 ns MM MD simulations) is considered. The barriers with different active site conformations demonstrate that increasing hydration around the carboxy-Lys73 base impairs deacylation: an additional water molecule next to the DW increases the barrier by at least 2 kcal/mol (Figure 6.2a). Further inspection of reaction simulations with the 6 α -hydroxyethyl group in orientation I provides an explanation for the difference in deacylation efficiency between imipenem (lacking a 1 β -methyl group) and meropenem. In imipenem deacylation, the DW donates hydrogen bonds to both carboxy-Lys73 and the 6 α -hydroxyethyl moiety (Figure 6.2b). In meropenem, the latter interaction is reversed; the DW now accepts a hydrogen bond from the 6 α -hydroxyethyl moiety instead. This difference, leading to a higher barrier as well as a change in the location of the transition state (Figure 6.2b), is likely due to a modification of the hydrogen bonding network around the 6 α -hydroxyethyl, ultimately caused by the presence of the 1 β -methyl group. Hirvonen et al. also used similar QM/MM (DFTB3/ff14SB) reaction simulations to compare the efficiency of

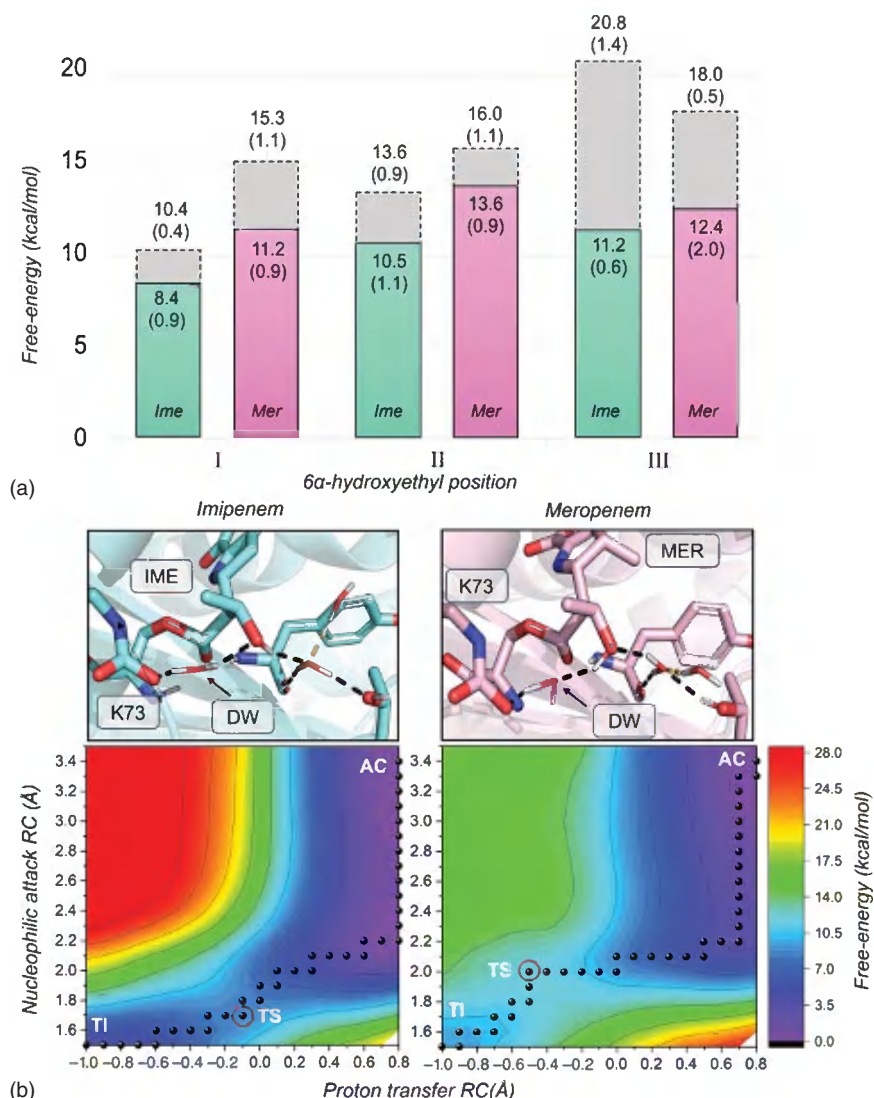


Figure 6.2 Dissection of the deacylation of imipenem and meropenem by the Class D β -lactamase OXA-48 using QM/MM (DFTB3/ff14SB) reaction simulations. (a) Free energy barriers obtained from 2D umbrella sampling for the three different 6 α -hydroxyethyl orientations observed in MM MD. Each bar includes the barrier obtained with a single water molecule hydrogen bonded to the carboxy-Lys73 (lowest barrier, outlined as *solid black line*) or with two water molecules (highest barrier, outlined as *dashed lines*). Each barrier is derived from three individual umbrella sampling runs, with standard deviations in parenthesis. Ime = imipenem, Mer = meropenem. (b) Free energy surfaces and transition state locations for 6 α -hydroxyethyl orientation I (lowest energy barriers in panel A), with alternative active site hydrogen bond configurations. DW = deacylating water, AC = acyl enzyme, TS = transition state, TI = tetrahedral intermediate. *Left*: Free energy surface for imipenem deacylation. The DW is donating a hydrogen bond to the carbapenem 6 α -hydroxyl group. *Right*: Free energy surface for meropenem deacylation. The carbapenem 6 α -hydroxyl donates a hydrogen bond to the DW. Source: Reproduced with permission of Hirvonen et al. [146]/ American Chemical Society / Public Domain CC BY 3.0.

the important cephalosporin antibiotic ceftazidime [147]. Whereas OXA-48 does not hydrolyze ceftazidime (and therefore will not confer significant resistance to bacterial infections treated with it), the variant OXA-163 does [145]. After distinguishing between reactive and nonreactive acyl-enzyme complex conformations of ceftazidime with DFTB3/ff14SB umbrella sampling, the barriers for deacylation obtained again indicated relative differences ($\Delta\Delta^\ddagger G$) consistent with experimental kinetics for OXA-48, OXA-163, and OXA-181 (another OXA-48-like enzyme not capable of ceftazidime hydrolysis). Further investigations of these simulations then indicated that subtle changes in the active site interactions lead to differences in hydration of carboxy-Lys73, which cause the difference in deacylation efficiency (rather than, e.g. differences in stabilization by the oxyanion hole). Overall, these works on dissecting the efficiency of β -lactam antibiotic breakdown by SBLs indicate how detailed QM/MM reaction simulations (combined with MM MD simulation) that consider alternative active site configurations provide detailed insights into the molecular origins of antibiotic resistance.

6.3.3 Covalent SARS-CoV-2 Inhibitors: Mechanism and Insights for Design

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) was first identified in December 2019 [148, 149] and then rapidly spread worldwide, with the World Health Organization declaring a global pandemic in March 2020. The virus had claimed the lives of more than 6.5 million people worldwide as of October 2022 [150]. Despite the success of the rapid development of COVID-19 vaccines, vaccination does not eradicate severe illness due to the emergence of COVID-19 variants [151, 152]. It is thus of high importance to develop effective treatment strategies that can prevent disease progression. One such strategy is to develop effective inhibitors of the main protease, M^{pro} (also known as 3-chymotrypsin-like protease, 3CL $^{\text{pro}}$), a key enzyme involved in the viral replication and transcription processes [153].

The mechanism of the cysteine protease M^{pro} is expected to feature a two-step acylation–deacylation process [154], similar to the serine β -lactamases, for example. Several groups investigated the details of this mechanism in SARS-CoV-2 M^{pro} soon after its structure was determined. Ramos-Guzman et al. [155] simulated the reaction with a representative peptide substrate at the B3LYP-D3/6-31G*-ff14SB QM-MM level. After characterizing the Michaelis complex between M^{pro} and the peptide, the adaptive string method [77] was used to perform umbrella sampling for both the acylation and deacylation reaction steps, using the relevant distances between the substrate and the Cys145/His41 catalytic dyad as collective variables. This indicated a free energy barrier of 14.6 kcal/mol for acylation and a concerted but asynchronous mechanism, where the catalytic dyad (Cys145 $^-$ /His41H $^+$) forms an ion pair, followed by the nucleophilic addition of the cysteine residue and protonation of the amide nitrogen at the P1' position of the substrate by His41. This mechanism, featuring an initial Cys145 $^-$ /His41H $^+$ ion pair formation, was also supported for a similar peptide substrate, using transition state optimization followed by intrinsic reaction coordinate optimizations with the subtractive ONIOM method [156], with geometries optimized at the B3LYP/6-31G(d,p)-ff14SB

level and energy calculations up to DLPNO-CCSD(T)/cc-pVTZ (with estimation of free energies using zero-point energy, thermal, and entropic corrections). However, the first QM/MM reaction simulations of SARS-CoV-2 M^{Pro}, performed by Swiderek and Moliner [157], indicated that in acylation, proton transfer between Cys145 and His41 may be concerted with nucleophilic attack of Cys145 to the carbonyl atom of the substrate. This was based on QM/MM umbrella sampling along two reaction coordinates with AM1/ff03, corrected to M06-2X/6-31+G(d, p)-MM by comparison of potential energy surfaces. The three studies agree that for M^{Pro}, the Cys145⁻/His41H⁺ ion pair is higher in energy than the corresponding neutral state, but whether or not this ion pair is a stable intermediate preceding nucleophilic attack by the Cys145 thiolate may depend on the substrate: in Refs. [155, 156], a serine is in position P1' (which is expected to stabilize the ion pair through its side-chain hydroxyl), whereas in Ref. [157], the chromophore 7-amino-4-carbamoylmethylcoumarin is present there.

Several groups used QM/MM reaction simulations early in the pandemic to provide insights into M^{Pro} inhibitor design [158–161], often based on the structure of SARS-CoV-2 in complex with a known covalent coronavirus M^{Pro} inhibitor, initially deposited already in January 2020 [162]. This peptidyl inhibitor, named N3, features a Michael acceptor warhead (see Figure 6.3). Ramos-Guzmán et al. used MM and QM/MM (B3LYP-D3/6-31G*-ff14SB) MD simulations to investigate the reaction between M^{Pro} and N3 [158]. Based on these simulations and their previous simulations with a peptide substrate [155], they suggested possible modifications to improve the non-covalent affinity (by restoring interactions that the P2' group of the peptide substrate forms with M^{Pro}), as well as lower the barrier to covalent bond formation (by focusing on facilitating Cys145⁻/His41H⁺ ion pair formation). Awoonor-Williams and Abu-Saleh [159] compared the same N3 inhibitor with an alternative α -ketoamide inhibitor, using (MM) alchemical binding free energy calculations (for the non-covalent complex) and ONIOM QM/MM optimizations for covalent complex formation (using M06-2X/def2-TZVP for the QM region). This indicated that for N3, the covalent adduct formation is not very efficient, consistent with Ref. [158]. (Notably, this study modeled the Cys145/His41 catalytic dyad as neutral rather than ionic; the protonation states of these two residues can be affected by the pH or presence of a substrate/inhibitor [163, 164].) Ramos-Guzmán et al. further used their MM and QM/MM MD simulation protocols, combined with alchemical binding free energy calculations, to study the interaction between M^{Pro} and PF-00835231 [165]. This is a hydroxymethylketone derivative drug candidate (formed upon phosphate hydrolysis of lufotrelvir) that inhibits M^{Pro} with low nM affinity and suitable pharmacokinetics [166]. MM (ff14SB) MD simulations indicated that the P1' hydroxymethyl group mimics the serine residue interactions present in natural substrates, but also forms additional interactions with Gly143 and Asn142. Moreover, the B3LYP-D3/6-31+G*-ff14SB free energy profile obtained using the adaptive string method indicated that this group also participates in the rate-limiting step of the reaction (with an activation barrier consistent with experiment [167]): nucleophilic attack of the Cys145 thiolate on the carbonyl carbon of the inhibitor is concerted with proton transfer to the carbonyl oxygen from His41

via the P1' hydroxyl. Based on the analysis of the simulations, the authors suggested that the addition of a chloromethyl moiety to the P1' hydroxymethyl group should lower the reaction barrier for the formation of the covalent complex. This was then confirmed by simulation: thermodynamic integration (at the MM level) indicated that the modification does not affect the non-covalent binding free energy, whereas the reaction barrier calculated by their QM/MM MD simulations was significantly reduced, primarily due to the expected stabilization of the ionic dyad pair. This work thus indicates how QM/MM simulations (combined with MM MD) can suggest possible routes to further improve M^{Pro} covalent inhibitors.

The use of combining QM/MM simulations directly with the design of covalent SARS-CoV-2 M^{Pro} inhibitors was first demonstrated by Arafet et al. [160], who employed their QM/MM approach based on semi-empirical AM1/ff03 umbrella sampling corrected to the M06-2X/6-31+G(d,p) level (as used previously for investigating a peptide substrate [157], see above). First, the covalent reaction between M^{Pro} and the N3 inhibitor was characterized, starting from the covalent complex, indicating an exergonic reaction (with the covalent complex ~ 18 kcal/mol more stable than the initial non-covalent complex). Then, based on these and previous simulations alongside medicinal chemistry experience, two inhibitors were designed: **B1**, where the warhead of **N3** was retained, and **B2**, where both the recognition portion (with glutamine at P1, consistent with the M^{Pro} substrate specificity) and the warhead (now a nitroalkene) were changed (Figure 6.3). The subsequently calculated QM/MM free energy profiles, in which the catalytic dyad residues and the **P1'**, **P1**, and **P2** fragments of the ligands were modeled in the

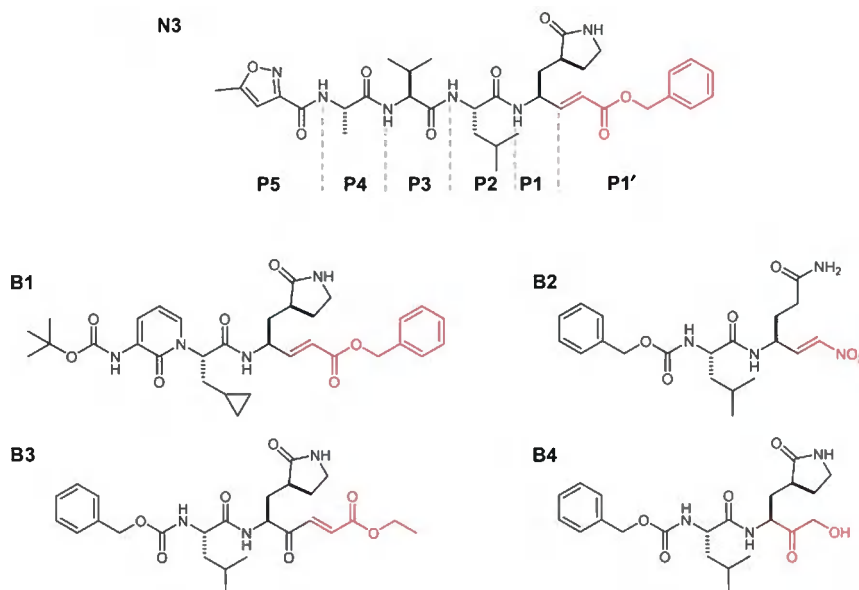


Figure 6.3 Chemical structures of peptidyl SARS-CoV-2 main protease (M^{Pro}) inhibitor **N3** (with the different fragments indicated by gray dashed lines) and proposed derivatives (**B1**, **B2**, **B3**, **B4**) [160, 161]. Warheads are highlighted in orange.

QM region, indicated the same mechanism as for N3. The barrier toward covalent inhibition was essentially the same between N3 and B1 and somewhat lower for B2. Further, the B1 compound resulted in the most stable Cys145⁻/His41H⁺ ion pair. The main difference between the two, however, was that B1 inhibition was predicted to be clearly irreversible (covalent complex ~28 kcal/mol more stable than the non-covalent complex) and B2 reversible (~11 kcal/mol difference). Overall, the QM/MM calculations indicate that interactions between the recognition portion and M^{pro} affect the energetics of the formation of the covalent complex, as these determine the orientation of the inhibitor in the active site.

Marti et al. [161] then further built on this work to design and evaluate two further peptidyl inhibitor compounds, B3 and B4 (Figure 6.3). The recognition portion was selected based on the M^{pro} interactions from their previous QM/MM simulations [160, 161], using the P1 moiety of B1 and the P2 and P3 moieties of B2. The warheads differ: B3 has an ethyl oxo-enoate warhead and B4 has a hydroxymethyl ketone warhead (the same as the PF-00835231 inhibitor). Starting from models of the non-covalent complex, free energy profiles for each inhibitor were obtained using umbrella sampling at the M06-2X/6-31+G(d,p)-ff03 level along a path collective variable combining key distances. After confirming the mechanism, the reaction was divided into two steps: formation of the Cys145⁻/His41H⁺ ion pair and the subsequent covalent complex formation by nucleophilic attack of Cys145 on the inhibitor and proton transfer from His41 to the (former) warhead. The resulting free energy profile indicates that the barrier and reaction energy for ion pair formation are essentially identical, with the ion pair ~8 kcal/mol higher in energy than the initial neutral catalytic dyad (Figure 6.4a). The rate-limiting covalent bond formation step (TS2 in Figure 6.4a) is influenced by the type of warhead present. For B3, nucleophilic attack by Cys145 and proton transfer from His41 to C α were predicted to take place concertedly (Figure 6.4b). For B4, the reactive thiolate already approaches the carbonyl carbon in the Cys145⁻/His41H⁺ ion pair state, and then completes the nucleophilic attack with concomitant protonation of the carbonyl oxygen by His41 through the B4 warhead hydroxyl moiety (Figure 6.4c). Although the energetics of the rate-limiting step are similar, B3 is indicated as a more promising lead for M^{pro} inhibitor design, as it is somewhat more reactive than B4 (activation barrier of 13.5 vs. 15.2 kcal/mol, Figure 6.4a) and leads to a ~2 kcal/mol more stable covalent complex. The mechanism for B4 covalent complex formation (alongside a relatively low barrier), however, indicates that modulating the pK_a of the warhead hydroxyl group can potentially lead to increased potency.

As well as the detailed QM/MM studies highlighted above, others have also developed QM/MM-based approaches to help evaluate and design M^{pro} inhibitors. Mondal and Warshel [168] first studied a reversible α -ketoamide inhibitor using the EVB approach for reaction simulation (with parameters to a reference reaction calculated at the B3LYP/6-31+G** level) together with a protein dipole Langevin dipole method they previously developed (PDLD/S-LRA/ β) for the non-covalent binding energy. They noted that, in addition to the electrophilicity of the warhead, the last step of the mechanism (protonation of the covalent complex) can be tuned to control the level of exothermicity, resulting in either reversible or

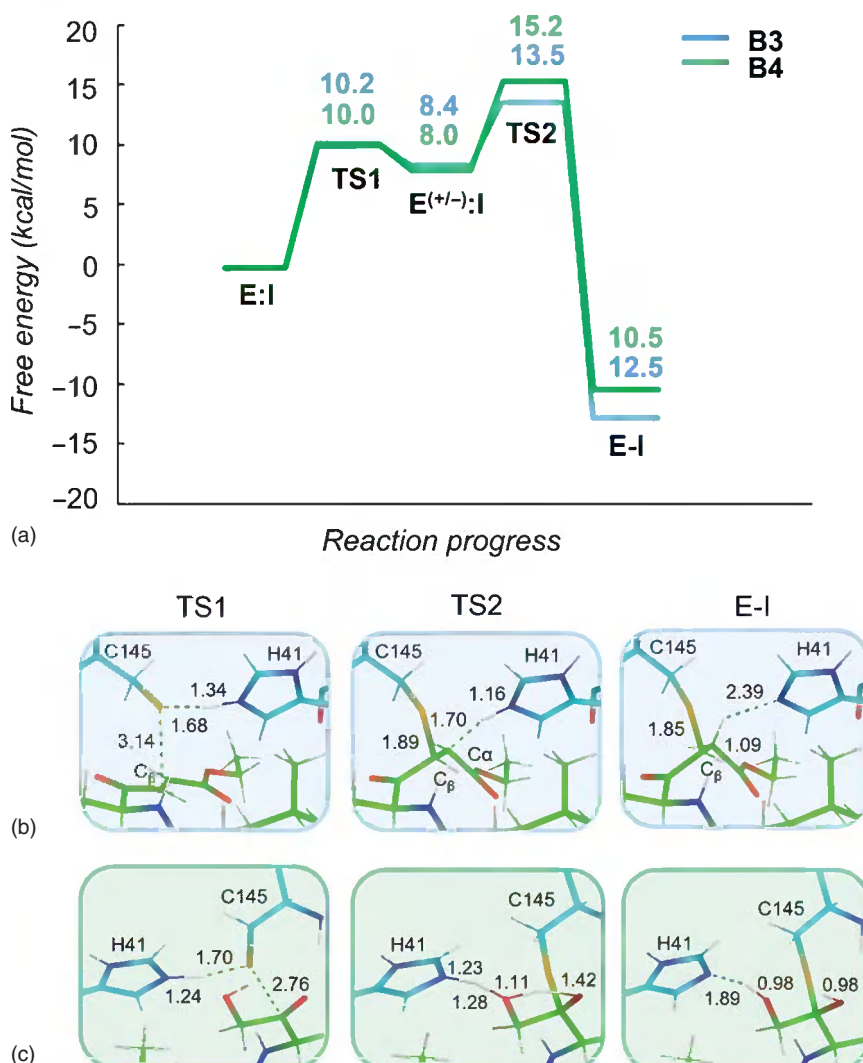


Figure 6.4 QM/MM (M06-2X/6-31+G(d,p)-ff03) reaction simulations for the designed B3 and B4 SARS-CoV-2 M^{Pro} inhibitors. (a) Free energy profiles of B3 (blue line) and B4 (green line) inhibitors. (b) Optimized structures of key states in the inhibition process of M^{Pro} by B3 (blue) and (c) B4 (green). Carbon atoms of the inhibitor are shown in green while those of the catalytic residues Cys145 and His41 are in cyan. Important hydrogen bonding interactions are indicated as green dashed lines, and key distances are given in Å. Adapted from Ref. [161].

irreversible covalent inhibition. The same group then developed an approach using a thermodynamic cycle with PDL/D/S-LRA/ β calculations for covalent inhibitor binding free energy calculations, avoiding more time-consuming QM/MM or EVB reaction simulations [169]. Calculations for covalent inhibitors against M^{Pro} and the 20S proteasome showed excellent agreement with experimental results, indicating

that the method is effective for inhibitors with different warheads (such as aldehyde and α -ketoamide) and can be applied to both reversible and irreversible inhibitors. Chan and co-workers [170] investigated a series of M^{pro} natural substrates and covalent inhibitors generated by the COVID Moonshot project [171] using a range of biomolecular simulation methods. QM/MM umbrella sampling simulations following the proton transfer between Cys145 and His41 proved to be useful in determining that the preferred state of the catalytic dyad was neutral. Notably, both extensive MM MD simulations [172] and QM calculations [173] have also highlighted the dependence of inhibitor binding on M^{pro} His protonation and tautomer preferences.

The simulation studies discussed in this section indicate that both the catalytic and inhibition mechanisms of M^{pro} depend on several factors that can influence the formation and stability of the covalent complex: pK_a of active site residues, solvent accessibility, induced fit effects, and the nature of the substrate/inhibitor [170]. However, for both substrates and inhibitors, the rate-limiting step is indicated to be covalent bond formation. Different studies can reach different conclusions for the detailed mechanistic pathway, even for the same inhibitor (e.g. N3); this can be partly due to the QM level, sampling method, and QM region used [158, 160, 161]. For instance, neglecting key residues involved in the stabilization of the oxyanion can lead to higher activation energies [156]. Nevertheless, QM/MM studies have shown to be able to aid the design of potent and selective inhibitors as lead compounds against SARS-CoV-2 M^{pro}.

6.4 Conclusions and Outlook

The practical application of QM/MM simulations of biomolecular systems is becoming increasingly routine, aided by the availability of accessible QM/MM software, as well as increases in computational power. The ability of QM methods to provide accurate descriptions of small molecules and their interactions in target binding sites is attractive, and methodological developments make their use increasingly feasible in the context of drug design [174], with QM/MM offering a tractable approach to include the wider biomolecular environment. Further, the determination of reaction pathways of pharmaceutically relevant enzymes with QM/MM methods can provide valuable insights: atomistic understanding of the interactions of transition states and reaction intermediates can suggest new scaffolds or modifications of existing compounds, leading to the design of both noncovalent and covalent inhibitors as starting points for drug discovery campaigns. For covalent inhibitors, QM/MM methods can be directly used to obtain insights into their reactivity, mechanism, and stability, providing priorities for synthesis and suggestions for further development. Many such applications now exist, as highlighted in this chapter.

A key trend, observed in studies across all target families discussed here, is to combine QM/MM reaction simulations with other simulation and modeling techniques. This allows for computationally tractable extensive (MM) conformational sampling of reactive complexes and stable intermediates, as well as the determination of

relative or absolute non-covalent binding free energies, prior to modeling chemical reactions with QM/MM. For example, for kinases and cysteine proteases, the pK_a of the key active site cysteine can first be determined using constant pH MD simulations or indeed machine learning algorithms [175]. Then, MM MD simulations, perhaps combined with alchemical free energy methods (thermodynamic integration or free-energy perturbation), can provide quantitative information on the non-covalent step of the binding process. For the covalent step, QM/MM reaction simulations would be used. The resulting complete characterization can highlight key aspects and interactions to focus on in further rounds of design, tuning potency, and selectivity. In the near future, we expect that such combinations of methods, including QM/MM reaction simulations, will be applied more routinely to aid the design of covalent drugs, likely alongside computational methods capable of high-throughput screening, such as covalent docking [176] (see Chapter 25).

It should be noted that QM/MM reaction simulations to accurately estimate differences in energy barriers are not yet straightforward to apply. However, the first demonstrations of the use of QM/MM as a practical computational assay for covalent drugs have now been reported [138, 177]. For efficiency, semi-empirical QM methods are typically used to make screening of tens of compounds (or target variants) tractable. Such methods ideally still require benchmarking and/or extensive testing against experimental data for each new chemical application to ensure reliability. A potential future solution for this could be the use of machine learning molecular potentials, which promise the accuracy of high-level QM at speeds faster than semi-empirical QM [178], so that accurate assays become feasible for larger amounts of compounds and/or more conformational sampling. For such machine learning potentials to be used in simulations in a biomolecular context, however, further methodological developments are still required, such as adequate solutions for electrostatic embedding in an ML/MM setting. Such developments, combined with improvements in automation, can make QM/MM reaction simulation a practical tool in the coming years, extending the existing experimental/computational toolbox [89] for covalent drug design.

References

- 1 Morzan, U.N., de Armino, D.J.A., Foglia, N.O. et al. (2018). Spectroscopy in complex environments from QM-MM simulations. *Chem Rev* 118 (7): 4071–4113.
- 2 Uddin, N., Choi, T.H., and Choi, C.H. (2013). Direct absolute pK_a predictions and proton transfer mechanisms of small molecules in aqueous solution by QM/MM-MD. *J Phys Chem B* 117 (20): 6269–6275.
- 3 Nelson, J.G., Peng, Y.X., Silverstein, D.W., and Swanson, J.M.J. (2014). Multi-scale reactive molecular dynamics for absolute $pK(a)$ predictions and amino acid deprotonation. *J Chem Theory Comput* 10 (7): 2729–2737.
- 4 Steinmann, C., Olsson, M.A., and Ryde, U. (2018). Relative ligand-binding free energies calculated from multiple short QM/MM MD simulations. *J Chem Theory Comput* 14 (6): 3228–3237.

- 5 van der Kamp, M.W. and Mulholland, A.J. (2013). Combined quantum mechanics/molecular mechanics (QM/MM) methods in computational enzymology. *Biochemistry* 52 (16): 2708–2728.
- 6 Senn, H.M. and Thiel, W. (2009). QM/MM methods for biomolecular systems. *Angew Chem Int Ed Engl* 48 (7): 1198–1229.
- 7 Warshel, A. and Levitt, M. (1976). Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J Mol Biol* 103 (2): 227–249.
- 8 Singh, U.C. and Kollman, P.A. (1986). A combined ab initio quantum mechanical and molecular mechanical method for carrying out simulations on complex molecular systems: applications to the $\text{CH}_3\text{Cl} + \text{Cl}^-$ exchange reaction and gas phase protonation of polyethers. *J Comput Chem* 7 (6): 718–730.
- 9 Field, M.J., Bash, P.A., and Karplus, M. (1990). A combined quantum mechanical and molecular mechanical potential for molecular dynamics simulations. *J Comput Chem* 11 (6): 700–733.
- 10 Bash, P.A., Field, M.J., Davenport, R.C. et al. (1991). Computer simulation and analysis of the reaction pathway of triosephosphate isomerase. *Biochemistry* 30 (24): 5826–5832.
- 11 Ranaghan, K.E. and Mulholland, A.J. (2017). Chapter 11 QM/MM methods for simulating enzyme reactions. In: *Simulating enzyme reactivity: computational methods in enzyme catalysis*, 375–403. The Royal Society of Chemistry.
- 12 Mulholland, A.J. (2005). Modelling enzyme reaction mechanisms, specificity and catalysis. *Drug Discov Today* 10 (20): 1393–1402.
- 13 Menikarachchi, L.C. and Gascon, J.A. (2010). QM/MM approaches in medicinal chemistry research. *Curr Top Med Chem* 10 (1): 46–54.
- 14 Lodola, A. and De Vivo, M. (2012). The increasing role of QM/MM in drug discovery. *Adv Protein Chem Struct Biol* 87: 337–362.
- 15 Barbault, F. and Maurel, F. (2015). Simulation with quantum mechanics/molecular mechanics for drug discovery. *Expert Opin Drug Discov* 10 (10): 1047–1057.
- 16 Kulkarni, P.U., Shah, H., and Vyas, V.K. (2022). Hybrid quantum mechanics/molecular mechanics (QM/MM) simulation: a tool for structure-based drug design and discovery. *Mini Rev Med Chem* 22 (8): 1096–1107.
- 17 Lodola, A., Callegari, D., Scalvini, L. et al. (2020). Design and SAR analysis of covalent inhibitors driven by hybrid QM/MM simulations. *Methods Mol Biol* 2114: 307–337.
- 18 Warshel, A. (1991). *Computer modeling of chemical reactions in enzymes and solutions*. New York: J. Wiley & Sons, Inc. ISBN: 0-47-1533955.
- 19 Kamerlin, S.C.L. and Warshel, A. (2010). The EVB as a quantitative tool for formulating simulations and analyzing biological and chemical reactions. *Faraday Discuss* 145: 71–106.
- 20 Loco, D., Lagardere, L., Caprasecca, S. et al. (2017). Hybrid QM/MM molecular dynamics with AMOEBA polarizable embedding. *J Chem Theory Comput* 13 (9): 4025–4033.

- 21 Brooks, B.R., Brooks, C.L. 3rd, Mackerell, A.D. Jr. et al. (2009). CHARMM: the biomolecular simulation program. *J Comput Chem* 30 (10): 1545–1614.
- 22 Thibault, J.C., Cheatham, T.E. 3rd, and Facelli, J.C. (2014). iBIOMES lite: summarizing biomolecular simulation data in limited settings. *J Chem Inf Model* 54 (6): 1810–1819.
- 23 Schrödinger Release 2021–4: QSite, Schrödinger, LLC, New York, NY (2021).
- 24 Murphy, R.B., Philipp, D.M., and Friesner, R.A. (2000). A mixed quantum mechanics/molecular mechanics (QM/MM) method for large-scale modeling of chemistry in protein environments. *J Comput Chem* 21 (16): 1442–1457.
- 25 Melo, M.C.R., Bernardi, R.C., Rudack, T. et al. (2018). NAMD goes quantum: an integrative suite for hybrid simulations. *Nat Methods* 15 (5): 351–354.
- 26 Kubar, T., Welke, K., and Groenhof, G. (2015). New QM/MM implementation of the DFTB3 method in the gromacs package. *J Comput Chem* 36 (26): 1978–1989.
- 27 Valiev, M., Yang, J., Adams, J.A. et al. (2007). Phosphorylation reaction in cAPK protein kinase-free energy quantum mechanical/molecular mechanics simulations. *J Phys Chem B* 111 (47): 13455–13464.
- 28 Neese, F., Wennmohs, F., Becker, U., and Riplinger, C. (2020). The ORCA quantum chemistry program package. *J Chem Phys* 152 (22): 224108.
- 29 Kuhne, T.D., Iannuzzi, M., Del Ben, M. et al. (2020). CP2K: an electronic structure and molecular dynamics software package – quickstep: efficient and accurate electronic structure calculations. *J Chem Phys* 152 (19): 194103.
- 30 Sherwood, P., Vries, A.H., Guest, M.F. et al. (2003). QUASI: a general purpose implementation of the QM/MM approach and its application to problems in catalysis. *J Mol Struct THEOCHEM* 632 (1–3): 1–28.
- 31 Lu, Y., Farrow, M.R., Fayon, P. et al. (2019). Open-source, Python-based redevelopment of the ChemShell multiscale QM/MM environment. *J Chem Theory Comput* 15 (2): 1317–1328.
- 32 Torras, J., Roberts, B.P., Seabra, G.M., and Trickey, S.B. (2015). PUPIL: a software integration system for multi-scale QM/MM-MD simulations and its application to biomolecular systems. *Adv Protein Chem Struct Biol* 100: 1–31.
- 33 Marti, S. (2021). QMCube (QM[3]): an all-purpose suite for multiscale QM/MM calculations. *J Comput Chem* 42 (6): 447–457.
- 34 Vreven, T., Byun, K.S., Komaromi, I. et al. (2006). Combining quantum mechanics methods with molecular mechanics methods in ONIOM. *J Chem Theory Comput* 2 (3): 815–826.
- 35 Ryde, U. (2016). QM/MM calculations on proteins. *Methods Enzymol* 577: 119–158.
- 36 Halder, S., Comitani, F., Saladino, G. et al. (2018). A multiscale simulation approach to modeling drug-protein binding kinetics. *J Chem Theory Comput* 14 (11): 6093–6101.
- 37 Cho, A.E., Guallar, V., Berne, B.J., and Friesner, R. (2005). Importance of accurate charges in molecular docking: quantum mechanical/molecular mechanical (QM/MM) approach. *J Comput Chem* 26 (9): 915–931.

- 38 Kim, M. and Cho, A.E. (2016). Incorporating QM and solvation into docking for applications to GPCR targets. *Phys Chem Chem Phys* 18 (40): 28281–28289.
- 39 Kurczab, R. (2017). The evaluation of QM/MM-driven molecular docking combined with MM/GBSA calculations as a halogen-bond scoring strategy. *Acta Crystallogr B Struct Sci Cryst Eng Mater* 73 (Pt 2): 188–194.
- 40 Chaskar, P., Zoete, V., and Rohrig, U.F. (2017). On-the-fly QM/MM docking with attracting cavities. *J Chem Inf Model* 57 (1): 73–84.
- 41 Burger, S.K., Thompson, D.C., and Ayers, P.W. (2011). Quantum mechanics/molecular mechanics strategies for docking pose refinement: distinguishing between binders and decoys in cytochrome C peroxidase. *J Chem Inf Model* 51 (1): 93–101.
- 42 Lee, T.S., Allen, B.K., Giese, T.J. et al. (2020). Alchemical binding free energy calculations in AMBER20: advances and best practices for drug discovery. *J Chem Inf Model* 60 (11): 5595–5623.
- 43 Hudson, P.S., Boresch, S., Rogers, D.M., and Woodcock, H.L. (2018). Accelerating QM/MM free energy computations via intramolecular force matching. *J Chem Theory Comput* 14 (12): 6327–6335.
- 44 Kearns, F.L., Warrensford, L., Boresch, S., and Woodcock, H.L. (2019). The good, the bad, and the ugly: “HiPen”, a new dataset for validating (S)QM/MM free energy simulations. *Molecules* 24 (4): 681.
- 45 Olsson, M.A. and Ryde, U. (2017). Comparison of QM/MM methods to obtain ligand-binding free energies. *J Chem Theory Comput* 13 (5): 2245–2253.
- 46 Giese, T.J. and York, D.M. (2019). Development of a robust indirect approach for MM → QM free energy calculations that combines force-matched reference potential and Bennett’s acceptance ratio methods. *J Chem Theory Comput* 15 (10): 5543–5562.
- 47 Rathore, R.S., Sumakanth, M., Reddy, M.S. et al. (2013). Advances in binding free energies calculations: QM/MM-based free energy perturbation method for drug design. *Curr Pharm Des* 19 (26): 4674–4686.
- 48 Genheden, S. and Ryde, U. (2015). The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin Drug Discov* 10 (5): 449–461.
- 49 Pu, C., Yan, G., Shi, J., and Li, R. (2017). Assessing the performance of docking scoring function, FEP, MM-GBSA, and QM/MM-GBSA approaches on a series of PLK1 inhibitors. *MedChemComm* 8 (7): 1452–1458.
- 50 Anisimov, V.M. and Cavasotto, C.N. (2011). Quantum mechanical binding free energy calculation for phosphopeptide inhibitors of the Lck SH2 domain. *J Comput Chem* 32 (10): 2254–2263.
- 51 Anisimov, V.M., Ziemys, A., Kizhake, S. et al. (2011). Computational and experimental studies of the interaction between phospho-peptides and the C-terminal domain of BRCA1. *J Comput Aided Mol Des* 25 (11): 1071–1084.
- 52 Pecina, A., Meier, R., Fanfrik, J. et al. (2016). The SQM/COSMO filter: reliable native pose identification based on the quantum-mechanical description of protein-ligand interactions and implicit COSMO solvation. *Chem Commun (Camb)* 52 (16): 3312–3315.

- 53 Pecina, A., Eyrilmez, S.M., Kopruluoglu, C. et al. (2020). SQM/COSMO scoring function: reliable quantum-mechanical tool for sampling and ranking in structure-based drug design. *ChemPlusChem* 85 (11): 2362–2371.
- 54 Glide, S. (2021). *Schrödinger release 2021–4: QM-polarized ligand docking protocol*. New York, NY: LLC.
- 55 Begum, J., Skamnaki, V.T., Moffatt, C. et al. (2015). An evaluation of indirubin analogues as phosphorylase kinase inhibitors. *J Mol Graph Model* 61: 231–242.
- 56 Wichapong, K., Rohe, A., Platzer, C. et al. (2014). Application of docking and QM/MM-GBSA rescoring to screen for novel Myt1 kinase inhibitors. *J Chem Inf Model* 54 (3): 881–893.
- 57 Kiss, M., Szabo, E., Bocskai, B. et al. (2021). Nanomolar inhibition of human OGA by 2-acetamido-2-deoxy-d-glucono-1,5-lactone semicarbazone derivatives. *Eur J Med Chem* 223: 113649.
- 58 Chetter, B.A., Kyriakis, E., Barr, D. et al. (2020). Synthetic flavonoid derivatives targeting the glycogen phosphorylase inhibitor site: QM/MM-PBSA motivated synthesis of substituted 5,7-dihydroxyflavones, crystallography, in vitro kinetics and ex-vivo cellular experiments reveal novel potent inhibitors. *Bioorg Chem* 102: 104003.
- 59 Jing, Z., Liu, C., Cheng, S.Y. et al. (2019). Polarizable force fields for biomolecular simulations: recent advances and applications. *Annu Rev Biophys* 48: 371–394.
- 60 Walker, B., Liu, C., Wait, E., and Ren, P. (2022). Automation of AMOEBA polarizable force field for small molecules: Poltype 2. *J Comput Chem* 43 (23): 1530–1542.
- 61 Rupakheti, C.R., MacKerell, A.D. Jr., and Roux, B. (2021). Global optimization of the Lennard-Jones parameters for the drude polarizable force field. *J Chem Theory Comput* 17 (11): 7085–7095.
- 62 Amezcua, M., El Khoury, L., and Mobley, D.L. (2021). SAMPL7 host-guest challenge overview: assessing the reliability of polarizable and non-polarizable methods for binding free energy calculations. *J Comput Aided Mol Des* 35 (1): 1–35.
- 63 Crespo, A., Rodriguez-Granillo, A., and Lim, V.T. (2017). Quantum-mechanics methodologies in drug discovery: applications of docking and scoring in lead optimization. *Curr Top Med Chem* 17 (23): 2663–2680.
- 64 Liu, H., Lu, Z., Cisneros, G.A., and Yang, W. (2004). Parallel iterative reaction path optimization in ab initio quantum mechanical/molecular mechanical modeling of enzyme reactions. *J Chem Phys* 121 (2): 697–706.
- 65 Cisneros, G.A., Liu, H., Lu, Z., and Yang, W. (2005). Reaction path determination for quantum mechanical/molecular mechanical modeling of enzyme reactions by combining first order and second order “chain-of-replicas” methods. *J Chem Phys* 122 (11): 114502.
- 66 Woodcock, H.L., Hodošček, M., Sherwood, P. et al. (2003). Exploring the quantum mechanical/molecular mechanical replica path method: a pathway optimization of the chorismate to prephenate Claisen rearrangement catalyzed by chorismate mutase. *Theoret Chem Acc* 109 (3): 140–148.

- 67 Ryde, U. (2017). How many conformations need to be sampled to obtain converged QM/MM energies? The curse of exponential averaging. *J Chem Theory Comput* 13 (11): 5745–5752.
- 68 Guimaraes, C.R., Udier-Blagovic, M., Tubert-Brohman, I., and Jorgensen, W.L. (2005). Effects of Arg90 neutralization on the enzyme-catalyzed rearrangement of Chorismate to prephenate. *J Chem Theory Comput* 1 (4): 617–625.
- 69 Rosta, E., Klahn, M., and Warshel, A. (2006). Towards accurate ab initio QM/MM calculations of free-energy profiles of enzymatic reactions. *J Phys Chem B* 110 (6): 2934–2941.
- 70 Hu, H., Lu, Z.Y., and Yang, W.T. (2007). QM/MM minimum free-energy path: methodology and application to triosephosphate isomerase. *J Chem Theory Comput* 3 (2): 390–406.
- 71 Kumar, S., Bouzida, D., Swendsen, R.H. et al. (1992). The weighted histogram analysis method for free-energy calculations on biomolecules. 1. The method. *J Comput Chem* 13 (8): 1011–1021.
- 72 Kästner, J. (2012). Umbrella integration with higher-order correction terms. *J Chem Phys* 136 (23): 234102.
- 73 Rosta, E. and Hummer, G. (2015). Free energies from dynamic weighted histogram analysis using unbiased Markov state model. *J Chem Theory Comput* 11 (1): 276–285.
- 74 Stelzl, L.S., Kells, A., Rosta, E., and Hummer, G. (2017). Dynamic histogram analysis to determine free energies and rates from biased simulations. *J Chem Theory Comput* 13 (12): 6328–6342.
- 75 Vanden-Eijnden, E. and Venturoli, M. (2009). Revisiting the finite temperature string method for the calculation of reaction tubes and free energies. *J Chem Phys* 130 (19): 194103.
- 76 Rosta, E., Nowotny, M., Yang, W., and Hummer, G. (2011). Catalytic mechanism of RNA backbone cleavage by ribonuclease H from quantum mechanics/molecular mechanics simulations. *J Am Chem Soc* 133 (23): 8934–8941.
- 77 Zinovjev, K. and Tunon, I. (2017). Adaptive finite temperature string method in collective variables. *J Phys Chem A* 121 (51): 9764–9772.
- 78 Zinovjev, K., Ruiz-Pernia, J.J., and Tunon, I. (2013). Toward an automatic determination of enzymatic reaction mechanisms and their activation free energies. *J Chem Theory Comput* 9 (8): 3740–3749.
- 79 Park, S., Khalili-Araghi, F., Tajkhorshid, E., and Schulten, K. (2003). Free energy calculation from steered molecular dynamics simulations using Jarzynski's equality. *J Chem Phys* 119 (6): 3559–3566.
- 80 Laio, A. and Parrinello, M. (2002). Escaping free-energy minima. *Proc Natl Acad Sci U S A* 99 (20): 12562–12566.
- 81 Bolnykh, V., Olsen, J.M.H., Meloni, S. et al. (2020). MiMiC: multiscale modeling in computational chemistry. *Front Mol Biosci* 7: 45.
- 82 Garcia-Viloca, M., Gao, J., Karplus, M., and Truhlar, D.G. (2004). How enzymes work: analysis by modern rate theory and computer simulations. *Science* 303 (5655): 186–195.

- 83 Serapian, S.A. and van der Kamp, M.W. (2019). Unpicking the cause of stereoselectivity in actinorhodin ketoreductase variants with atomistic simulations. *ACS Catal* 9 (3): 2381–2394.
- 84 Mlynsky, V., Banas, P., Sponer, J. et al. (2014). Comparison of ab initio, DFT, and semiempirical QM/MM approaches for description of catalytic mechanism of hairpin ribozyme. *J Chem Theory Comput* 10 (4): 1608–1622.
- 85 Claeysens, F., Harvey, J.N., Manby, F.R. et al. (2006). High-accuracy computation of reaction barriers in enzymes. *Angew Chem Int Ed* 45 (41): 6856–6859.
- 86 Bauer, R.A. (2015). Covalent inhibitors in drug discovery: from accidental discoveries to avoided liabilities and designed therapies. *Drug Discov Today* 20 (9): 1061–1073.
- 87 Smith, G.F. (2011). Designing drugs to avoid toxicity. *Prog Med Chem* 50: 1–47.
- 88 Sutanto, F., Konstantinidou, M., and Domling, A. (2020). Covalent inhibitors: a rational approach to drug discovery. *RSC Med Chem* 11 (8): 876–884.
- 89 Boike, L., Henning, N.J., and Nomura, D.K. (2022). Advances in covalent drug discovery. *Nat Rev Drug Discov* 1–18.
- 90 Pottier, C., Fresnais, M., Gilon, M. et al. (2020). Tyrosine kinase inhibitors in cancer: breakthrough and challenges of targeted therapy. *Cancer* 12 (3): 731.
- 91 Seshacharyulu, P., Ponnusamy, M.P., Haridas, D. et al. (2012). Targeting the EGFR signaling pathway in cancer therapy. *Expert Opin Ther Targets* 16: 15–31.
- 92 Bethune, G.C., Bethune, D.C., Ridgway, N.D., and Xu, Z. (2010). Epidermal growth factor receptor (EGFR) in lung cancer: an overview and update. *J Thorac Dis* 2 (1): 48–51.
- 93 Molina, J.R., Yang, P., Cassivi, S.D. et al. (2008). Non-small cell lung cancer: epidemiology, risk factors, treatment, and survivorship. *Mayo Clin Proc* 352 (8): 584–594.
- 94 Kobayashi, S.S., Boggon, T.J., Dayaram, T. et al. (2005). EGFR mutation and resistance of non-small-cell lung cancer to gefitinib. *N Engl J Med* 352 (8): 786–792.
- 95 Morgillo, F., Della Corte, C.M., Fasano, M., and Ciardiello, F. (2016). Mechanisms of resistance to EGFR-targeted drugs: lung cancer. *ESMO Open* 1 (3): e000060.
- 96 Yu, H.A. and Riely, G. (2013). Second-generation epidermal growth factor receptor tyrosine kinase inhibitors in lung cancers. *J Natl Compr Canc Netw* 11 (2): 161–169.
- 97 Schwartz, P.A., Kuzmič, P., Solowiej, J.E. et al. (2013). Covalent EGFR inhibitor analysis reveals importance of reversible interactions to potency and mechanisms of drug resistance. *Proc Natl Acad Sci* 111 (1): 173–178.
- 98 Hossam, M., Lasheen, D.S., and Abouzid, K.A.M. (2016). Covalent EGFR inhibitors: binding mechanisms, synthetic approaches, and clinical profiles. *Arch Pharm* 349 (8): 573–593.
- 99 Capoferri, L., Lodola, A., Rivara, S., and Mor, M. (2015). Quantum mechanics/molecular mechanics modeling of covalent addition between EGFR-cysteine 797 and N-(4-anilinoquinazolin-6-yl) acrylamide. *J Chem Inf Model* 55 (3): 589–599.

- 100 Blair, J.A., Rauh, D., Kung, C. et al. (2007). Structure-guided development of affinity probes for tyrosine kinases using chemical genetics. *Nat Chem Biol* 3 (4): 229–238.
- 101 Carmi, C., Galvani, E., Vacondio, F. et al. (2012). Irreversible inhibition of epidermal growth factor receptor activity by 3-aminopropanamides. *J Med Chem* 55 (5): 2251–2264.
- 102 Lence, E., van der Kamp, M.W., González-Bello, C., and Mulholland, A.J. (2018). QM/MM simulations identify the determinants of catalytic activity differences between type II dehydroquinase enzymes. *Org Biomol Chem* 16 (24): 4443–4455.
- 103 Yao, J., Guo, H.-B., Chaiprasongsuk, M. et al. (2015). Substrate-assisted catalysis in the reaction catalyzed by salicylic acid binding protein 2 (SABP2), a potential mechanism of substrate discrimination for some promiscuous enzymes. *Biochemistry* 54 (34): 5366–5375.
- 104 Demapan, D., Kussmann, J., Ochsenfeld, C., and Cui, Q. (2022). Factors that determine the variation of equilibrium and kinetic properties of QM/MM enzyme simulations: QM region, conformation, and boundary condition. *J Chem Theory Comput* 18 (4): 2530–2542.
- 105 Callegari, D., Ranaghan, K.E., Woods, C.J. et al. (2018). L718Q mutant EGFR escapes covalent inhibition by stabilizing a non-reactive conformation of the lung cancer drug osimertinib. *Chem Sci* 9 (10): 2740–2749.
- 106 Gao, X., Le, X., and Costa, D.B. (2016). The safety and efficacy of osimertinib for the treatment of EGFR T790M mutation positive non-small-cell lung cancer. *Expert Rev Anticancer Ther* 16 (4): 383–390.
- 107 He, J., Huang, Z., Han, L. et al. (2021). Mechanisms and management of 3rd-generation EGFR-TKI resistance in advanced non-small cell lung cancer (review). *Int J Oncol* 59 (5).
- 108 Bersanelli, M., Minari, R., Bordi, P. et al. (2016). L718Q mutation as new mechanism of acquired resistance to AZD9291 in EGFR-mutated NSCLC. *J Thorac Oncol* 11 (10): e121–e123.
- 109 Woods, C.J., Malaisree, M., Hannongbua, S., and Mulholland, A.J. (2011). A water-swap reaction coordinate for the calculation of absolute protein-ligand binding free energies. *J Chem Phys* 134 (5): 054114.
- 110 Castelli, R., Bozza, N., Cavazzoni, A. et al. (2019). Balancing reactivity and antitumor activity: heteroarylthioacetamide derivatives as potent and time-dependent inhibitors of EGFR. *Eur J Med Chem* 162: 507–524.
- 111 Weber, A.N.R., Bittner, Z.A., Liu, X. et al. (2017). Bruton's tyrosine kinase: an emerging key player in innate immunity. *Front Immunol* 8.
- 112 Wang, Q., Pechersky, Y., Sagawa, S. et al. (2019). Structural mechanism for Bruton's tyrosine kinase activation at the cell membrane. *Proc Natl Acad Sci U S A* 116 (19): 9390–9399.
- 113 López-Herrera, G., Vargas-Hernández, A., Gonzalez-Serrano, M.E. et al. (2014). Bruton's tyrosine kinase—an integral protein of B cell development that also has an essential role in the innate immune system. *J Leukoc Biol* 95 (2): 243–250.

- 114 Crofford, L.J., Nyhoff, L.E., Sheehan, J.H., and Kendall, P.L. (2016). The role of Bruton's tyrosine kinase in autoimmunity and implications for therapy. *Expert Rev Clin Immunol* 12 (7): 763–773.
- 115 Kil, L.P., de Bruijn, M.J.W., van Nimwegen, M. et al. (2012). Btk levels set the threshold for B-cell activation and negative selection of autoreactive B cells in mice. *Blood* 119 (16): 3744–3756.
- 116 Wen, T., Wang, J., Shi, Y. et al. (2021). Inhibitors targeting Bruton's tyrosine kinase in cancers: drug development advances. *Leukemia* 35 (2): 312–332.
- 117 Gayko, U., Fung, M.-C., Clow, F. et al. (2015). Development of the Bruton's tyrosine kinase inhibitor ibrutinib for B cell malignancies. *Ann N Y Acad Sci* 1358: 82–94.
- 118 Voice, A., Tresadern, G., Twidale, R.M. et al. (2021). Mechanism of covalent binding of ibrutinib to Bruton's tyrosine kinase revealed by QM/MM calculations. *Chem Sci* 12 (15): 5511–5516.
- 119 Kaptein, A., de Bruin, G., Emmelot-van Hoek, M. et al. (2019). Potency and selectivity of BTK inhibitors in clinical development for B-cell malignancies. *Clin Lymphoma Myeloma Leuk* 132: 1871.
- 120 Voice, A., Tresadern, G., Hv, V., and Mulholland, A.J. (2019). Limitations of ligand-only approaches for predicting the reactivity of covalent inhibitors. *J Chem Inf Model* 59 (10): 4220–4227.
- 121 Awoonor-Williams, E. and Rowley, C.N. (2021). Modeling the binding and conformational energetics of a targeted covalent inhibitor to Bruton's tyrosine kinase. *J Chem Inf Model* 61 (10): 5234–5242.
- 122 Murray, C.J.L., Ikuta, K.S., Sharara, F. et al. (2022). Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet (London, England)* 399 (10325): 629–655.
- 123 O'Neill, J. (2016). Tackling Drug-Resistant Infections Globally: Final Report and Recommendations. Government of the United Kingdom.
- 124 Hermann, J.C., Ridder, L., Hölftje, H.-D., and Mulholland, A.J. (2006). Molecular mechanisms of antibiotic resistance: QM/MM modelling of deacylation in a class a beta-lactamase. *Org Biomol Chem* 4 (2): 206–210.
- 125 Hermann, J.C., Ridder, L., Mulholland, A.J., and Holtje, H.D. (2003). Identification of Glu166 as the general base in the acylation reaction of class A beta-lactamases through QM/MM modeling. *J Am Chem Soc* 125 (32): 9590–9591.
- 126 Hermann, J.C., Hensen, C., Ridder, L. et al. (2005). Mechanisms of antibiotic resistance: QM/MM modeling of the acylation reaction of a class A beta-lactamase with benzylpenicillin. *J Am Chem Soc* 127 (12): 4454–4465.
- 127 Meroueh, S.O., Fisher, J.F., Schlegel, H.B., and Mobashery, S. (2005). Ab initio QM/MM study of class A beta-lactamase acylation: dual participation of Glu166 and Lys73 in a concerted base promotion of Ser70. *J Am Chem Soc* 127 (44): 15397–15407.
- 128 Hermann, J.C., Pradon, J., Harvey, J.N., and Mulholland, A.J. (2009). High level QM/MM modeling of the formation of the tetrahedral intermediate in the

- acylation of wild type and K73A mutant TEM-1 class A beta-lactamase. *J Phys Chem A* 113 (43): 11984–11994.
- 129 Choi, H., Paton, R.S., Park, H., and Schofield, C.J. (2016). Investigations on recyclisation and hydrolysis in avibactam mediated serine β -lactamase inhibition. *Org Biomol Chem* 14 (17): 4116–4128.
 - 130 Das, C.K. and Nair, N.N. (2020). Elucidating the molecular basis of avibactam-mediated inhibition of class A beta-lactamases. *Chemistry* 26 (43): 9639–9651.
 - 131 Lizana, I., Uribe, E.A., and Delgado, E.J. (2021). A theoretical approach for the acylation/deacylation mechanisms of avibactam in the reversible inhibition of KPC-2. *J Comput Aided Mol Des* 35 (9): 943–952.
 - 132 Tripathi, R.C. and Nair, N.N. (2013). Mechanism of acyl-enzyme complex formation from the Henry-Michaelis complex of class C β -lactamases with β -lactam antibiotics. *J Am Chem Soc* 135 (39): 14679–14690.
 - 133 Gherman, B.F., Goldberg, S.D., Cornish, V.W., and Friesner, R.A. (2004). Mixed quantum mechanical/molecular mechanical (QM/MM) study of the deacylation reaction in a penicillin binding protein (PBP) versus in a class C beta-lactamase. *J Am Chem Soc* 126 (24): 7652–7664.
 - 134 Tripathi, R.C. and Nair, N.N. (2016). Deacylation mechanism and kinetics of acyl-enzyme complex of class C β -lactamase and cephalothin. *J Phys Chem B* 120 (10): 2681–2690.
 - 135 Sgrignani, J., Grazioso, G., and De Amici, M. (2016). Insight into the mechanism of hydrolysis of meropenem by OXA-23 serine- β -lactamase gained by quantum mechanics/molecular mechanics calculations. *Biochemistry* 55 (36): 5191–5200.
 - 136 Swarén, P., Maveyraud, L., Raquet, X. et al. (1998). X-ray analysis of the NMC-A β -lactamase at 1.64-Å resolution, a class A carbapenemase with broad substrate specificity. *J Biol Chem* 273 (41): 26714–26721.
 - 137 Chudyk, E.I., Limb, M.A.L., Jones, C.E.S. et al. (2014). QM/MM simulations as an assay for carbapenemase activity in class A β -lactamases. *Chem Commun* 50 (94): 14736–14739.
 - 138 Hirvonen, V.H.A., Hammond, K., Chudyk, E.I. et al. (2019). An efficient computational assay for β -lactam antibiotic breakdown by class A β -lactamases. *J Chem Inf Model* 59 (8): 3365–3369.
 - 139 Chudyk, E.I., Beer, M., Limb, M.A.L. et al. (2022). QM/MM simulations reveal the determinants of carbapenemase activity in class A β -lactamases. *ACS Infect Dis* 8 (8): 1521–1532.
 - 140 Fritz, R.A., Alzate-Morales, J.H., Spencer, J. et al. (2018). Multiscale simulations of clavulanate inhibition identify the reactive complex in class A β -lactamases and predict the efficiency of inhibition. *Biochemistry* 57 (26): 3560–3563.
 - 141 Song, Z. and Tao, P.-C. (2022). Graph-learning guided mechanistic insights into imipenem hydrolysis in GES carbapenemases. *Electron Struct* 4 (3).
 - 142 Charnas, R.L. and Knowles, J.R. (1981). Inhibition of the RTEM beta-lactamase from *Escherichia coli*. Interaction of enzyme with derivatives of olivanic acid. *Biochemistry* 20 (10): 2732–2737.

- 143** Easton, C.J. and Knowles, J.R. (1982). Inhibition of the RTEM beta-lactamase from *Escherichia coli*. Interaction of the enzyme with derivatives of olivanic acid. *Biochemistry* 21 (12): 2857–2862.
- 144** Poirel, L., Potron, A., and Nordmann, P. (2012). OXA-48-like carbapenemases: the phantom menace. *J Antimicrob Chemother* 67 (7): 1597–1606.
- 145** Hirvonen, V.H.A., Spencer, J., and van der Kamp, M.W. (2021). Antimicrobial resistance conferred by OXA-48 β -lactamases: towards a detailed mechanistic understanding. *Antimicrob Agents Chemother* 65 (6): e00184–e00121.
- 146** Hirvonen, V.H.A., Weizmann, T.M., Mulholland, A.J. et al. (2022). Multi-scale simulations identify origins of differential carbapenem hydrolysis by the OXA-48 β -lactamase. *ACS Catal* 12 (8): 4534–4544.
- 147** Hirvonen, V.H.A., Mulholland, A.J., Spencer, J., and van der Kamp, M.W. (2020). Small changes in hydration determine cephalosporinase activity of OXA-48 β -lactamases. *ACS Catal* 10 (11): 6188–6196.
- 148** Huang, C., Wang, Y.-m., Li, X.-w. et al. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet (London, England)* 395 (10223): 497–506.
- 149** Li, Q., Guan, X.-h., Wu, P. et al. (2020). Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. *N Engl J Med* 382: 1199–1207.
- 150** WHO (2022). *COVID-19 dashboard*. Geneva: World Health Organization [updated 2022 Oct; cited 2022 Oct 20]. Available from: <https://covid19.who.int>.
- 151** Cevik, M., Grubaugh, N.D., Iwasaki, A., and Openshaw, P. (2021). COVID-19 vaccines: keeping pace with SARS-CoV-2 variants. *Cell* 184 (20): 5077–5081.
- 152** Mahase, E. (2021). Covid-19: what new variants are emerging and how are they being investigated? *BMJ* 372: n158.
- 153** Ullrich, S. and Nitsche, C. (2020). The SARS-CoV-2 main protease as drug target. *Bioorg Med Chem Lett* 30 (17): 127377.
- 154** Solowiej, J., Thomson, J.A., Ryan, K. et al. (2008). Steady-state and pre-steady-state kinetic evaluation of severe acute respiratory syndrome coronavirus (SARS-CoV) 3CLpro cysteine protease: development of an ion-pair model for catalysis. *Biochemistry* 47 (8): 2617–2630.
- 155** Ramos-Guzman, C.A., Ruiz-Pernia, J.J., and Tunon, I. (2020). Unraveling the SARS-CoV-2 main protease mechanism using multiscale methods. *ACS Catal* 10: 12544–12554.
- 156** Fernandes, H.S., Sousa, S.F., and Cerqueira, N. (2022). New insights into the catalytic mechanism of the SARS-CoV-2 main protease: an ONIOM QM/MM approach. *Mol Divers* 26 (3): 1373–1381.
- 157** Swiderek, K. and Moliner, V. (2020). Revealing the molecular mechanisms of proteolysis of SARS-CoV-2 M(pro) by QM/MM computational methods. *Chem Sci* 11 (39): 10626–10630.
- 158** Ramos-Guzman, C.A., Ruiz-Pernia, J.J., and Tunon, I. (2021). A microscopic description of SARS-CoV-2 main protease inhibition with Michael acceptors. Strategies for improving inhibitor design. *Chem Sci* 12 (10): 3489–3496.

- 159 Awoonor-Williams, E. and Abu-Saleh, A.A.A. (2021). Covalent and non-covalent binding free energy calculations for peptidomimetic inhibitors of SARS-CoV-2 main protease. *Phys Chem Chem Phys* 23 (11): 6746–6757.
- 160 Arafet, K., Serrano-Aparicio, N., Lodola, A. et al. (2020). Mechanism of inhibition of SARS-CoV-2 M(pro) by N3 peptidyl Michael acceptor explained by QM/MM simulations and design of new derivatives with tunable chemical reactivity. *Chem Sci* 12 (4): 1433–1444.
- 161 Marti, S., Arafet, K., Lodola, A. et al. (2022). Impact of warhead modulations on the covalent inhibition of SARS-CoV-2 M(pro) explored by QM/MM simulations. *ACS Catal* 12 (1): 698–708.
- 162 Jin, Z., Du, X., Xu, Y. et al. (2020). Structure of M(pro) from SARS-CoV-2 and discovery of its inhibitors. *Nature* 582 (7811): 289–293.
- 163 Zanetti-Polzi, L., Smith, M.D., Chipot, C. et al. (2021). Tuning proton transfer thermodynamics in SARS-CoV-2 main protease: implications for catalysis and inhibitor design. *J Phys Chem Lett* 12 (17): 4195–4202.
- 164 Kneller, D.W., Phillips, G., Weiss, K.L. et al. (2020). Unusual zwitterionic catalytic site of SARS-CoV-2 main protease revealed by neutron crystallography. *J Biol Chem* 295 (50): 17365–17373.
- 165 Ramos-Guzman, C.A., Ruiz-Pernia, J.J., and Tunon, I. (2021). Inhibition mechanism of SARS-CoV-2 main protease with ketone-based inhibitors unveiled by multiscale simulations: insights for improved designs. *Angew Chem Int Ed Engl* 60 (49): 25933–25941.
- 166 Hoffman, R.L., Kania, R.S., Brothers, M.A. et al. (2020). Discovery of ketone-based covalent inhibitors of coronavirus 3CL proteases for the potential therapeutic treatment of COVID-19. *J Med Chem* 63 (21): 12725–12747.
- 167 Ma, C., Sacco, M.D., Hurst, B. et al. (2020). Boceprevir, GC-376, and calpain inhibitors II, XII inhibit SARS-CoV-2 viral replication by targeting the viral main protease. *Cell Res* 30 (8): 678–692.
- 168 Mondal, D. and Warshel, A. (2020). Exploring the mechanism of covalent inhibition: simulating the binding free energy of alpha-ketoamide inhibitors of the main protease of SARS-CoV-2. *Biochemistry* 59 (48): 4601–4608.
- 169 Zhou, J., Saha, A., Huang, Z., and Warshel, A. (2022). Fast and effective prediction of the absolute binding free energies of covalent inhibitors of SARS-CoV-2 main protease and 20S proteasome. *J Am Chem Soc* 144 (17): 7568–7572.
- 170 Chan, H.T.H., Moesser, M.A., Walters, R.K. et al. (2021). Discovery of SARS-CoV-2 M^{Pro} peptide inhibitors from modelling substrate and ligand binding. *Chem Sci* 12 (41): 13686–13703.
- 171 Achdout, H., Aimon, A., Bar-David, E. et al. (2022). Open science discovery of oral non-covalent SARS-CoV-2 main protease inhibitor therapeutics. *bioRxiv*.
- 172 Pavlova, A., Lynch, D.L., Daidone, I. et al. (2021). Inhibitor binding influences the protonation states of histidines in SARS-CoV-2 main protease. *Chem Sci* 12 (4): 1513–1527.
- 173 Poater, A. (2020). Michael acceptors tuned by the pivotal aromaticity of histidine to block COVID-19 activity. *J Phys Chem Lett* 11 (15): 6262–6265.

- 174 Bryce, R.A. (2020). What next for quantum mechanics in structure-based drug discovery? 2114: 339–353.
- 175 Gokcan, H. and Isayev, O. (2022). Prediction of protein pK_a with representation learning. *Chem Sci* 13 (8): 2462–2474.
- 176 Schirmeister, T., Kesselring, J., Jung, S. et al. (2016). Quantum chemical-based protocol for the rational design of covalent inhibitors. *J Am Chem Soc* 138 (27): 8332–8335.
- 177 Galvani, F., Scalvini, L., Rivara, S. et al. (2022). Mechanistic modeling of mono-glyceride lipase covalent modification elucidates the role of leaving group expulsion and discriminates inhibitors with high and low potency. *J Chem Inf Model* 62 (11): 2771–2787.
- 178 Smith, J.S., Nebgen, B.T., Zubatyuk, R. et al. (2019). Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nat Commun* 10 (1): 2903.

7

Recent Advances in Practical Quantum Mechanics and Mixed-QM/MM-Driven X-Ray Crystallography and Cryogenic Electron Microscopy (Cryo-EM) and Their Impact on Structure-Based Drug Discovery

Oleg Borbulevych and Lance M. Westerhoff

QuantumBio Inc, 2790 West College Ave, State College, PA 16801, United States

7.1 Introduction

X-ray crystallography is a core experimental procedure used to determine the three-dimensional (3D) atomic structure of biomolecular systems and inform structure-based drug discovery (SBDD) and fragment-based drug discovery (FBDD) efforts in most pharmaceutical companies and laboratories. SBDD utilizes protein (or DNA or RNA) and protein–ligand and protein–protein 3D structures to provide insights in lead discovery and optimization campaigns. In these campaigns, bonded and nonbonded interactions are explored and optimized to find compounds that best fit (chemically) the protein binding pocket. Therefore, obtaining an accurate representation of these structures is critical to the successful execution of SBDD projects. Successful FBDD screening also depends on accurate structure, but with the added complexity that it is typically carried out by soaking protein crystals with a cocktail of up to 10 small molecule compounds prior to the data collection, structure solution, and ligand placement utilized in SBDD [1]. Given the size of fragment compounds, this cocktail presents a challenge in that the density can accommodate any number of fragments in any number of orientations. Therefore, determining the correct orientation(s) of a particular fragment within an electron density blob, as such density is often weak or partial in the fragment area, becomes more difficult. Furthermore, in both SBDD and FBDD, ligands containing “flippable” functional groups, such as an amide group, are particularly susceptible to placement uncertainties since light elements (e.g. N and O) are often not readily distinguishable in macromolecular crystallography. Given the quality of the crystal structures is essential for the success of high-throughput screening, docking, and scoring (e.g. rank ordering) of potential drug candidates, this uncertainty can impact the entire drug discovery process.

Due to recent advances in data collection, processing, structure solution, and refinement automation, X-ray crystallography has become a routine method.

Despite that, the majority of crystal structures are still determined at modest or low resolutions, which generally leads to significant uncertainties in atomic coordinates and other structural errors [2, 3]. It has been argued that those structural errors adversely impact ligand binding affinity predictions [2], which are critical to SBDD/FBDD applications. A significant drawback of traditional macromolecular refinement stems from the fact that conventional stereochemical restraints – which are used almost exclusively for the refinement process – are rudimentary in nature and do not account for nonbonded interactions such as electrostatics, polarization, hydrogen bonds, dispersion, and charge transfer [4–6]. Moreover, conventional refinement methods rely entirely on a detailed, *ex situ* description of the molecular geometry for each ligand or cofactor in the model as captured in a Crystallographic Information File (CIF). Unfortunately, the creation of *accurate* CIFs is a nontrivial task, and this process often leads to bound ligand structures with less than desirable quality [5] due to an incomplete *a priori* understanding of *in situ* bound bond lengths and angles and a lack of intermolecular interactions in conventional refinement functionals [7].

One way to improve X-ray models is to utilize quantum mechanics (QM) during the crystallographic refinement; however, traditionally, the size of virtually all biological systems prohibited a straightforward application of the QM methods. Nevertheless, in 2002, with the aid of the program COMQUM-X [8], the first mixed-quantum mechanics/molecular mechanics (QM/MM) X-ray refinement was conducted using a small QM portion of the system (around 25 heavy atoms). Since then, several examples of the QM-refined structures against X-ray data have been reported [9–17], emphasizing the ligand geometry improvement and protonation state determination [18]. In 2014, QuantumBio Inc. – building on the previous work of the Merz laboratory [13, 15–17] – introduced a new, much more automated QM refinement technique that works by replacing the conventional stereochemical restraints of the ligand(s), cofactor(s), active sites(s), residue(s), or even the entire protein–ligand complex with accurate quantum-based energy functionals in “real-time” during the refinement [19, 20] as computed by the linear-scaling QM semiempirical quantum mechanics (SE-QM) method [20–22]. Prior to this work, it was demonstrated that such QM linear scaling calculations can capture the critical interactions between a target and its ligand(s), such as hydrogen bonds, electrostatics, polarization, charge transfer, and metal coordination [23–27], and because this QM refinement protocol explicitly skips any information provided by CIF, the method gives rise to better, more accurate *in situ* ligand and active site geometries. This earlier work gave rise to an even more performant, QM/MM methodology based on the ONIOM formalism [28] as implemented in DivCon to treat most any macromolecular structure using a single functional [29]. It is this primary work that has led to routine, high-throughput QM/MM X-ray refinement (and more recently cryo-EM refinement), which has a direct impact on the models used in SBDD, and this impact will be discussed in detail in this publication.

7.2 Feasibility of Routine and Fast QM-Driven X-Ray Refinement

One of the first practical approaches to incorporating QM/MM functional into X-ray refinement was the program COMQUM-X [8], implemented to integrate a QM/MM algorithm with the crystallographic software crystallography and NMR system (CNS) [30]. With that method, a ligand and approximately 25 atoms around it were treated at the *ab initio* Becke-Perdew86/6-31G* or B3LYP/6-31G* QM level of theory, and the rest of the residues were computed with molecular mechanics (MM) with the AMBER force field. During these refinements, the bulk of the structure was fixed to reduce computational costs. Merz's group then made a significant advance in the field by implementing the divide-and-conquer (D&C), linear scaling, and SE-QM methods previously described [21, 22, 31–33]. D&C SE-QM utilizes an approximate solution of the Schrödinger equation that can be written using the Hartee–Fock–Roothaan formalism as

$$\mathbf{FC} = \mathbf{CE} \quad (7.1)$$

where $\mathbf{F}$ is the Fock matrix, $\mathbf{C}$ is the matrix of molecular orbital (MO) coefficients, and $\mathbf{E}$ is the eigenvalue energy matrix. D&C SE-QM divides the protein–ligand complex into subsystems generally corresponding to the amino acid residues in the protein, and Eq. (7.1) is solved for each subsystem. Therefore, matrix diagonalization – the most expensive part of the QM calculation – is performed on each subsystem (along with a buffer region) instead of the entire complex, leading to significant savings in CPU time and memory use. Obtained MOs and density matrixes for the subsystems are then combined to yield a solution for the whole system. As a result, the calculation's memory and CPU time requirements scale $\sim$ linearly or $\sim O(n)$, where n is the number of atoms in the system. This is contrasted to traditional QM methods in which the memory requirements and CPU time exhibit $O(n^3)$ scaling. Thus, when this linear scaling formalism is joined with the already fast semiempirical level of theory used in DivCon, D&C makes routine QM calculations – including all-atom model optimization/refinement – possible on very large biological systems. In this early work, this method was applied to X-ray crystallography via integration with the CNS [30] package [13], in which all atoms in the structure were treated using the AM1 Hamiltonian [34]. But even with the linear scaling, applying the all-atom SE-QM in the X-ray refinement regime is more computationally expensive than conventional refinement methods. Furthermore, SE-QM when applied to protein systems can cause systematic deviations from the standard geometry in the backbone [13, 19]. Finally, the lack of d-orbital support in AM1 led to limits in compatibility with metal-containing complexes. Therefore, the next step in the evolution of QM refinement was to combine linear scaling SE-QM using the more modern PM6 Hamiltonian [35, 36] for the ligand(s), cofactor(s), metals, active site(s), and chosen residue regions, with MM using the AMBER ff14sb [37] force field as implemented in DivCon for the remainder of the

macromolecular system [29]. Finally, instead of CNS, which has been largely superseded in SBDD organizations, the DivCon module or plugin was integrated first with PHENIX [4] and then with BUSTER [38] to deploy the method on more modern platforms [19, 29].

To increase the accessibility of QM and QM/MM refinement for the community and support high-throughput crystallographic refinement, QuantumBio [39] went beyond this core development and implemented a user-friendly, **fully automated** molecular perception and preparation protocol that supports almost any protein/DNA/RNA/ligand structure. This development addresses a long-standing need to perform QM, MM, and QM/MM calculations quickly and easily with much fewer convergence problems or setup issues. Using the following protocol, models are determined and refined, which are not only chemically correct but chemically complete as well (with likely protonation states, residue rotamer states, and so on):

- Fast structure protonation, including optimization of the hydrogen network and flip states and pH effects implemented based on [40].
- Automated molecular perception [41] and formal charge determination of the entire system based on graph theory algorithms, including any unknown species, e.g. ligands, cofactors, metal coordination, nonstandard amino acids, and truncated residues.
- Automatic assignment of MM types for the entire system, including ligands, etc., based on molecular perception, and hence corresponding MM parameters for any AMBER forcefield chosen.
- Automatic residue-based selection of any number of QM regions extended by a given radius from any center, such as ligands, etc.
- Automatic link-atom (proton) addition for any internally “broken” bonds at the QM:MM interface.

Finally, to address traditional convergence problems in macromolecular QM calculations, this new DivCon uses several modern QM convergence optimization algorithms combined with Extended Hückle theory [42].

7.3 Metrics to Measure Improvement

To gage the performance of the QM-driven refinement in comparison to conventional approaches, several metrics reviewed below in detail have been used [15, 19, 29, 43–45].

7.3.1 Ligand Strain Energy

Ligand strain – which shows how much strain the ligand must accept to bind with the target protein – is an industry-standard method to define the quality of refined ligand structural models [15, 19, 45–47]. We calculate [15, 48] the local ligand strain energy or E_{Strain} as the difference between the energy of the isolated ligand

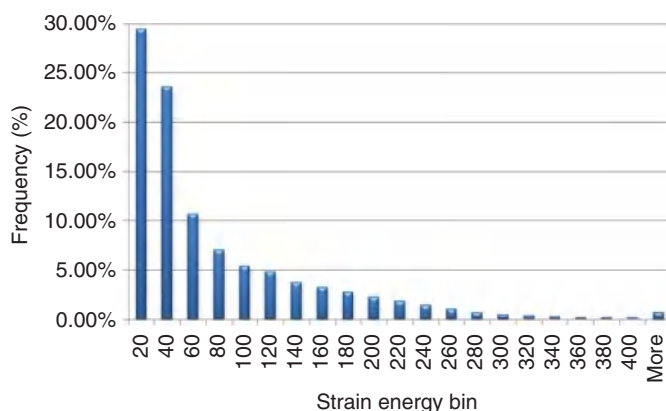


Figure 7.1 The distribution of strain energy values for 134 345 ligand poses calculated at the PM6 level.

conformation (optimized) and the protein-bound ligand conformation according to Eq. (7.2),

$$E_{\text{Strain}} = E_{\text{SinglePoint}} - E_{\text{Optimized}} \quad (7.2)$$

where $E_{\text{SinglePoint}}$ is the single-point energy computed for the ligand X-ray geometry, and $E_{\text{Optimized}}$ is the energy of the optimized ligand that corresponds to the local minimum. In 2012, we explored the strain energy distribution of over 134 345 ligand poses deposited in PDB [49]. As shown in (Figure 7.1), about 55% of all ligand poses belong in a 0–40 kcal/mol bin, ~25% of poses have strain energy above 100 kcal/mol, and the balance falls into three bins between 40 and 100 kcal/mol.

7.3.2 ZDD of Difference Density

In crystallography, the difference density ($\Delta\rho$) reveals the disagreement between the experimental data and the refined model. The positive (+) and negative (–) difference density peaks indicate not only missing or incorrectly placed ligands, fragments, and water molecules, but they also show more subtle details such as incorrectness in geometry or conformations of a molecule or positions of individual atoms. To quantitatively characterize the amount of the difference density around any residue in the crystal structure, a new quality indicator – the real-space Z score of difference density (ZDD) – was proposed by Tickle [43]. The benefit of this indicator is that it measures the *accuracy* of the model. In contrast, the often-used real space correlation coefficient (RSCC) correlates with *both* the accuracy *and* precision of the model, making RSCC less useful for measuring model accuracy (vs. experimental density).

A detailed derivation of ZDD can be found elsewhere [43, 50], but briefly, the Z score of difference density values at the point $\mathbf{r}$ is expressed as follows,

$$Z(\Delta\rho(\mathbf{r})) = \frac{\Delta\rho(\mathbf{r})}{\sigma(\Delta\rho(\mathbf{r}))} \quad (7.3)$$

where $\sigma(\Delta\rho(\mathbf{r}))$ is the standard deviation of the difference density, corresponds to the random error of the model, and is pure **precision**. In contrast, the Z score of the difference density measures the residual, nonrandom error and is pure **accuracy**. Positive and negative grid density points are analyzed separately, leading to two values, ZDD– and ZDD+. The maximum absolute number of those two values gives a final ZDD score (7.4).

$$ZDD = \max(\text{abs}(ZDD-), ZDD+) \quad (7.4)$$

7.3.3 Overall Crystallographic Structure Quality Metrics: MolProbity Score and Clashscore

MolProbity is a macromolecular model validation tool, which uses multiple quality criteria [44] to calculate a MolProbity score (MPScore). This logarithm-based score combines three key component metrics: Clashscore, Ramachandran plot statistics, and rotamer outliers [51]. The lower the MPScore, the better the quality of the model. Out of these three metrics, the Clashscore is a useful metric in itself and is the number of clashes per 1000 atoms. Clashes are determined by the construction of a nonbonded atom contact surface around each atom using the rolling probe algorithm [52]. A clash occurs when the nonbonded surface around one atom overlaps with the surface of another atom by more than 0.4 Å [3]. Overall, a crystal structure with problematic geometry results in many clashes and a high Clashscore value [44].

7.4 QM Region Refinement

Our first approach to the integration of SE-QM with the PHENIX [4] suite (PHENIX/DivCon) is called *Region-QM* refinement [19]. In this algorithm, the refined protein structure is divided into three regions: the main or core region(s), the buffer region(s), and the stereochemistry restraint region(s) (Figure 7.2). The core region(s) contains one or more ligands of interest as well as the selection of the target residues or other species such as water molecules, metal ions, cofactors, and so on within the given radius (e.g. 5 Å) from any ligand atom. The buffer region(s) are a second set of selection residues beyond each core region, which are the residues located at a second given distance (e.g. 3 Å) from any atom of the core region. Finally, the balance of the protein is treated as a pure stereochemical restraint region. The core regions, if there are more than one in the structure, do not need to be contiguous (and neither do buffer regions). The entire core and buffer regions are computed at the QM level of theory, but only QM gradients of the core region are employed in the refinement. Thus, each buffer region chemically insulates its core region to limit errors that may occur in the gradients due to capping or other artifacts in the surrounding chemical environment. Finally, for the remainder of the protein (outside of the core region), the atomic gradients are calculated using the standard stereochemistry restraint functional as implemented in the chosen X-ray crystallography platform (PHENIX or BUSTER). Mathematically, the QM

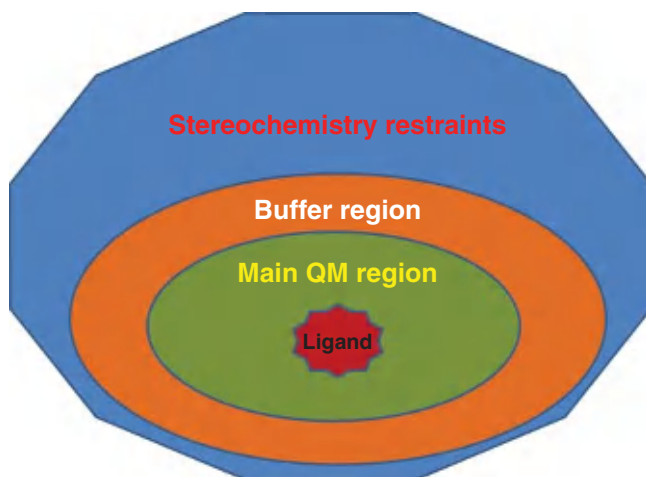


Figure 7.2 Schematic view of the QM region refinement concept.

and stereochemistry gradients on each atom with coordinates $\mathbf{x}$ are combined in the crystallography platform according to the equation,

$$(\nabla \mathbf{x}_i)_{total} = \kappa \times \Omega_{xray} \times (\nabla \mathbf{x}_i)_{xray} + \mathbf{w}_i \times (\nabla \mathbf{x}_i)_{QM} + (1 - \mathbf{w}_i) \times (\nabla \mathbf{x}_i)_{geom} \quad (7.5)$$

where the weight $\mathbf{w}$ is set to 1 for the core QM region(s) and 0 for the rest of the atoms, including the buffer region. Ω_{xray} is a variable weight determined using an automatic procedure in PHENIX or a fixed weight in BUSTER, and κ is the additional scale factor implemented in PHENIX. It is notable that the *full* QM refinement can be performed by setting the weight $\mathbf{w}$ of 1 for all atoms in the whole system.

Prior to this first effort, it was shown that the local chemistry of the ligand within the binding pocket could be improved with the integration of the QM methods into the X-ray refinement on the example of several crystal structures [9–13]. The Region-QM refinement approach is consistent, and we systematically demonstrated significant improvement of E_{strain} of 50, quasi-randomly chosen protein–ligand structures from the PDB. In particular, the average ligand strain energy for the set of 50 structures calculated for the deposited coordinates is 83.50 ± 9.03 kcal/mol, and the minimum and maximum values are 6.88 and 283.35 kcal/mol, respectively, or a range of 276.47 kcal/mol. After Region-QM refinement, significant improvement was observed in the E_{strain} throughout the set: the average strain energy of the re-refined set of structures is 24.60 ± 3.67 kcal/mol, or 3.5 times smaller than that of the deposited structures (Table 7.1). To validate these QM E_{strain} energies, we compared them with those calculated *ab initio* with the HF/6–311 + G** basis set. The change in the strain energies based on the *ab initio* calculations is less pronounced than that obtained with the SE-QM Hamiltonian (Table 7.1). It was expected, as the *ab initio* method was not used directly in the X-ray refinement. Also, the HF level of theory, despite the large basis set, does not consider electronic correlation, while it is partially incorporated into SE-QM methods such as AM1 and PM6. However, despite those factors, in all cases studied, SE-QM X-ray crystallographic refinement

Table 7.1 Average ligand strain energies over 50 crystal structures refined using region-QM refinement method.

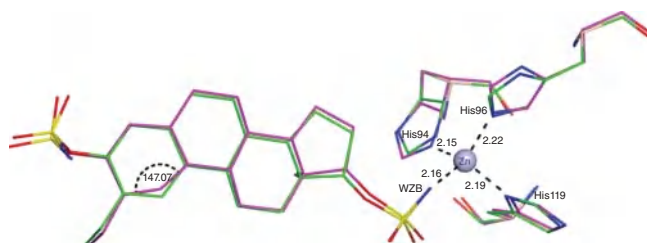
	Deposited PDB	QM refined	Improvement, fold
StrainEnergy, AM1	83.50 ± 9.03	24.60 ± 3.67	3.4
StrainEnergy, HF/6-311G**	93.39 ± 9.00	53.76 ± 4.61	1.5

Source: Adapted from Borbulevych et al. [19].

lead to significantly improved ab initio-calculated ligand strains. It confirms the robustness of the SE-QM methods for X-ray refinement [19].

As an example, PDB 2X7T at 2.8 Å resolution [53] is the structure of the enzyme carbonic anhydrase inhibited by the ligand WZB in an active site that includes tetrahedral zinc coordinated by the three histidine residues. The WZB is bound to the zinc via an amino group to complete the coordination sphere of the metal (Figure 7.3). In conventional X-ray crystallographic refinement protocols, structures involving ligands coordinated with metals usually require tedious work to create library files to account for these interactions. For instance, in the deposited structure 2X7T, all coordination distances involving Zn are in the range of 2.14–2.25 Å, which is longer than the average length of 2.00(2) Å for $\text{Zn} \cdots \text{N}$ and $\text{Zn} \cdots \text{O}$ coordination bonds [54]. Such discrepancies result in a distortion of the coordination sphere of the metal (Figure 7.3), which also affects the ligand geometry. In particular, the phenyl ring of the ligand is heavily distorted, including a bond angle of 147 degree and significant deviations of bond lengths from the average $\text{C}_{\text{ar}}-\text{C}_{\text{ar}}$ bond length of 1.398 Å [55]. These anomalies contribute to a high ligand E_{strain} of 96.71 kcal/mol for the deposited conformation. The QM region refinement was performed with none of these geometry assumptions concerning the coordination, yielding a model in which the coordination distances with the zinc are in the range of 1.98–2.02 Å. When these coordination distances are compared to the literature, we find them to be very close to the average values mentioned above [54]. Furthermore, the aforementioned ligand distortions evaporate, and the ligand E_{strain} drops to 14.5 kcal/mol indicating a significant improvement in the geometry of the bound ligand WZB.

We argued [19] that the main source of observed ligand E_{strain} improvement in the structures refined using the region-QM protocol is the elimination of errors

**Figure 7.3** Superimposition of the residues in the coordination sphere of zinc in the structure 2X7T from the region-QM (green) refinements and the original PDB (magenta).

in the ligand geometry. On the one hand, because it completely disregards any bond angle/length parameters provided by the CIF, QM X-ray crystallographic refinement *automatically* resolves the “garbage in/garbage out” problem [5], which results from the inaccurate or imprecise ligand descriptions found in these standard ligand libraries. On the other hand, the QM potential also influences the geometry of the ligand through electrostatic, polarization, and charge transfer interactions observed in situ that are not available in the rudimentary conventional restraints (especially those built from ex situ states).

7.5 ONIOM Refinement

Despite the success of the region-QM X-ray crystallographic refinement, this approach has two fundamental drawbacks. First, most protein atoms are still treated using simple conventional restraints. Second, nonbonded and electrostatic interactions between the QM region and the rest of the macromolecular system (outside of the core + buffer regions) are not considered. To address these weaknesses, we developed a more holistic approach to conduct QM refinement based on the all-atom QM/MM scheme [29]. Subtractive QM/MM or ONIOM [28], allows for the straightforward computation of the system’s energy (7.6),

$$E_{\text{ONIOM}}^{\text{QM/MM}} = E_{\text{region}}^{\text{QM}} + E_{\text{all}}^{\text{MM}} - E_{\text{region}}^{\text{MM}} \quad (7.6)$$

where the $E_{\text{all}}^{\text{MM}}$ term is the MM energy calculated for the entire system, the $E_{\text{region}}^{\text{MM}}$ term is the MM energy for the QM region, and $E_{\text{region}}^{\text{QM}}$ is the energy of the QM region computed with the chosen SE-QM Hamiltonian. QM/MM gradients in the subtractive scheme are calculated as follows,

$$\nabla \mathbf{x}_{\text{ONIOM}}^{\text{QM/MM}} = \nabla \mathbf{x}_{\text{region}}^{\text{QM}} + \nabla \mathbf{x}_{\text{all}}^{\text{MM}} - \nabla \mathbf{x}_{\text{region}}^{\text{MM}} \quad (7.7)$$

in which the gradients of the QM region(s) include terms from both the QM and the MM functionals, and the electrostatics and van der Waals interactions between the QM and MM regions are explicitly included in the energy and gradient calculations. Generally, the ONIOM approach leads to faster and more convergent calculations vs. other QM/MM methods (like additive QM/MM), and the approach readily supports models with multiple QM regions such as those with multiple active site/ligand centers commonly found in crystal structures. Using the ONIOM formalism for the whole system, X-ray refinement with PHENIX and BUSTER is expressed as follows,

$$(\nabla \mathbf{x}_i)_{\text{total}} = \kappa \times \Omega_{\text{Xray}} \times (\nabla \mathbf{x}_i)_{\text{Xray}} + \nabla \mathbf{x}_{\text{ONIOM}}^{\text{QM/MM}} \quad (7.8)$$

where $\nabla \mathbf{x}_{\text{ONIOM}}^{\text{QM/MM}}$ corresponds to the ONIOM gradients determined using expression (7.7). Therefore, any ligand(s) and surrounding binding pocket(s) is (are) defined as the QM region(s), and the remainder of the protein–ligand model is designated as the MM region(s) and characterized using the AMBER force field such as amberff14sb [37]. When this calculation is performed in the PHENIX platform, the Ω_{xray} term is a variable weight determined using an automatic procedure in the platform [56], and κ is the additional scale factor implemented in the platform [57]. Within

this ONIOM X-ray crystallographic protocol, *all* stereochemical restraint gradients (including those for the ligand(s), the waters, the cofactor(s), any metals, and all of the protein/DNA/RNA residues) are replaced with high-quality QM/MM gradients.

Overall, we consider ONIOM particularly well suited to fast, routine, high-throughput, and user-friendly QM/MM-based crystallographic refinement, as demonstrated in [29]. In that work, we show that ONIOM refinement exhibits superior performance as judged by four different metrics, including strain, ZDD, Mol-Probity, and Clashscore (Figure 7.4), when validated against the 80 structures and 141 discrete ligand poses of the Astex Diverse Set [58]. In this validation, three different types of X-ray refinements were performed and compared: ONIOM QM/MM, region-QM, and conventional PHENIX [29]. The strain energy distributions are similar for the 2 QM-driven refinements in which the average strain energies calculated over 141 ligands equal 9.95 ± 3.77 kcal/mol for ONIOM and 10.49 ± 4.52 kcal/mol for region-QM refinements. Furthermore, the ligand strain histograms (Figure 7.4a) for both QM refinements have peaks around 3.0 kcal/mol, accounting for approximately 75% of the ligand poses in the set. This congruence makes sense when one considers that in both cases, the same SE-QM Hamiltonian was brought to bear on approximately the same residues in each case. Therefore, the difference between the two values is likely due to the impact of the more complete Hamiltonian in ONIOM vs. the presence of stereochemical restraints in the region-QM method. Contrasting with the QM results, the conventional refinement yields strain energy data that are mostly evenly distributed across a broad range from 10 to 40 kcal/mol of which ~30% of data are in the last bin of 50+ kcal/mol. As a result, the average ligand strain energy after the Conventional refinement of the Astex set is 35.64 ± 9.35 kcal/mol, or about 3.5-fold higher than in the QM-based refinements.

The histogram for ZDD (Figure 7.4b) shows a broad peak at 1.4 units, a similar feature of all three distributions. Nevertheless, the number of ligands that fall into the bin range from 0 to 1.2 ZDD units is higher for ONIOM and region-QM refined structures than that of conventionally refined ones. Thus, the average ZDD for the ligands in ONIOM-refined structures (2.3 ± 0.8) is slightly lower (better) than that after the conventional refinements (2.9 ± 1.1). As expected, the region-QM refinement results (2.6 ± 0.9) fall in the middle. Interestingly, ligand strain and ZDD arrays are uncorrelated, as concluded from the Pearson correlation coefficient between those two metrics being close to zero for all refinement methods.

Where ONIOM refinement shows significant impact vs. both region-QM and conventional refinement is the improvement of the overall structure quality measured using Clashscore. Across the 80 refined Astex structures, the average Clashscores for conventional (4.83 ± 1.2 units) and region-QM (5.54 ± 1.6 units) refined models are similar. In contrast, the average Clashscore for the ONIOM structures is 1.10 ± 0.41 units, thus exhibiting a dramatic improvement of 4.5–5.0-fold. Furthermore, the Clashscore histogram (Figure 7.4c) reveals a sharp peak located at the 0.5 unit mark, which comprises 90% of the ONIOM models compared to the peak for both conventional and region-QM model data at around 3.5 units. About 50% of those data in the conventional and region-QM histograms are found in the tail of the respective peaks and distributed in histogram bins 4.5+ units, and above

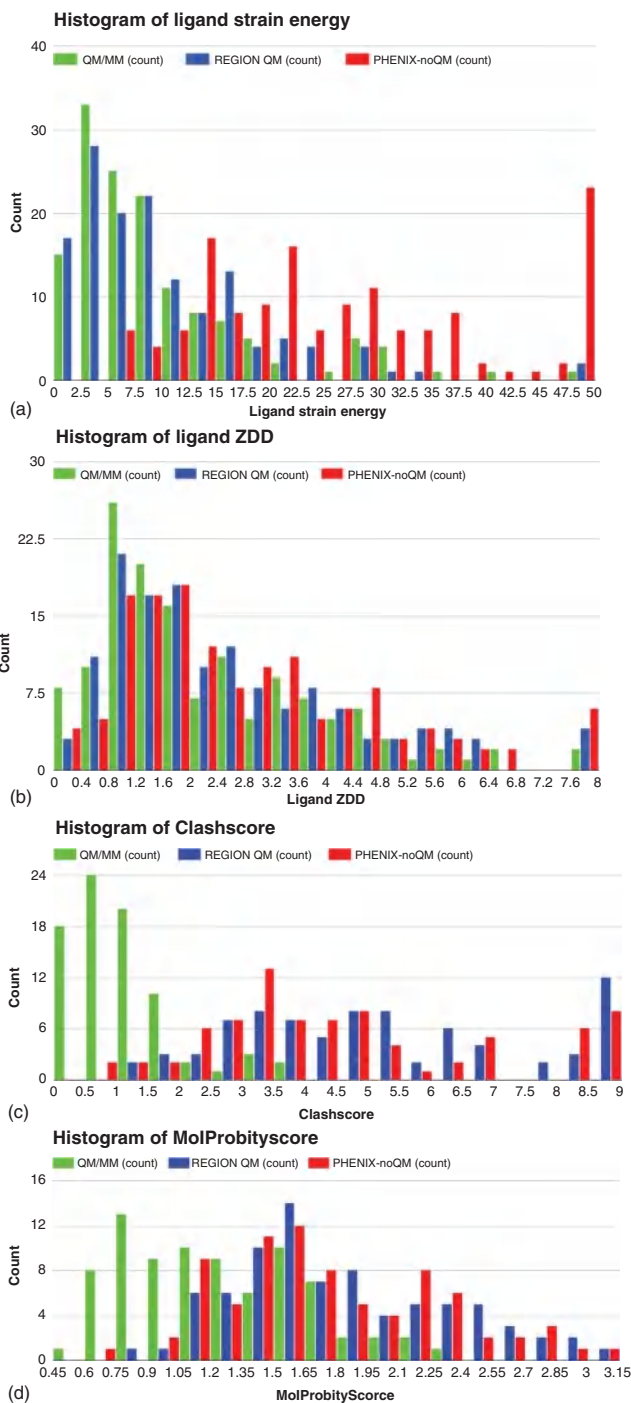


Figure 7.4 Histogram of ligand strain energy distributions (a), ligand ZDD distributions (b), MolProbity Clashscore distributions (c), and MolProbity score distributions (d) for 80 Astex structures refined with QM/MM (ONIOM), region-QM, and conventional methods. Histograms (a) and (b) include data for 141 ligand instances.

while ONIOM data are not even observed in that range. The similarity between region-QM and conventional refinements result indicates that the significant improvement in Clashscore for the bulk of the protein structure arises from using the QM/MM Hamiltonian on the entire structure. Notably, a similar improvement upon ONIOM refinement is observed for MPIScore (Figure 7.4d).

7.6 XModeScore: Distinguish Protomers, Tautomers, Flip States, and Docked Ligand Poses

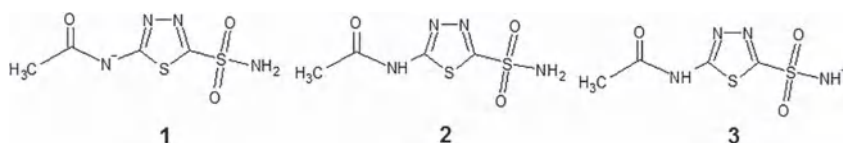
Given that X-rays are scattered by electrons, this leads to a fundamental limitation of X-ray crystallography in which the experiment is unable to (easily) directly detect the positions of hydrogen atoms. H atoms not only have just one electron, but their electron cloud is shifted toward the heavy atoms to which they are bound. Thus, the hydrogen atom has the weakest scattering power for X-rays among all elements [59]. Hence, with the rare exception of ultrahigh-resolution X-ray data, it is generally impossible to experimentally determine the protonation or tautomeric state of both the ligand and the surrounding active site in macromolecular crystallography at the resolutions often used in SBDD. Thus, a possible application of the QM methods to facilitate the determination of the protonation states of molecules or protein residues (e.g. GLU, ASP, or HIS) in the crystal structure and this has attracted attention in the past [18, 60]. Previous work has also shown that distinguishing among possible ligand tautomers is often pivotal in steering new drug discovery campaigns in the right direction [61].

Although X-ray crystallography does not permit direct observation of hydrogen atoms, QM/MM does not suffer from this problem, and therefore, using the QM/MM functional, it should be possible to detect the **influence** of hydrogen atoms on the heavy atoms (carbon, nitrogen, oxygen) to which they are bound and compare that influence to the experimental density to determine the protomer/tautomer state in the crystal. This idea is behind the XModeScore method developed by QuantumBio [50]. Using automatic enumeration of all likely protonation/tautomer ligand (and active site) states, followed by the QM/MM refinement of each protein–ligand state, this method is able to determine which protonation/tautomer form is most closely in agreement with the experimental X-ray data. The impact of the different protonation configurations is measured using two indicators of the ligand: E_{strain} and ZDD. In the XModeScore scoring procedure, both indicators are scaled according to the Z-score formula and then combined to produce the overall score of the i -tautomer form,

$$XModeScore_i = - \left\{ \frac{ZDD_i - \mu_{ZDD}}{\sigma_{ZDD}} + \frac{E_{strain_i} - \mu_{E_{strain}}}{\sigma_{E_{strain}}} \right\} \quad (7.9)$$

where μ is the mean value and σ is the standard deviation of the corresponding array of data (ZDD or E_{strain}). For example, the E_{strain} array contains E_{strain} values for all protomers/tautomers included in the calculations. $E_{strain,i}$ and ZDD_i are corresponding values of the i -tautomer. The highest $XModeScore_i$ corresponds to the protomeric/tautomeric form “ i ” that best fits both E_{strain} and ZDD criteria.

XModeScore was challenged using a Human carbonic anhydrase II (HCA II) structure bound to a high-affinity inhibitor [62, 63] acetazolamide (AZM). HCA II carbonate hydration/dehydration are involved in numerous metabolic processes, including CO₂ transport and pH regulation, and AZM was approved as a drug known as “Diamox” [64, 65]. AZM binds to the Zn atom within the active site of the enzyme via the nitrogen atom of the sulfonamide group, which completes its tetrahedral coordination by making coordination bonds with nitrogen atoms of His94, His96, and His119. In this configuration, as depicted in Scheme 7.1, AZM can exist in the three tautomeric/protonation forms. However, even high-resolution X-ray diffraction studies failed to determine which state of AZM exists in the crystal [66, 67]. It was only when the community obtained a neutron diffraction model of this enzyme [67] that it was proven that AZM exists in form **3**, and thus binds to zinc via the negatively charged sulfonamide SO₂NH group in the crystal form.



Scheme 7.1 Potential binding states of the compound AZM.

XModeScore results [50] based on the region-QM refinements of the three considered forms of AZM using the X-ray data from PDB 3HS4 reveal that form **3** is the superior form and is the best (lowest) in both E_{strain} and ZDD scoring components (Table 7.2, Figure 7.5). This finding is entirely consistent with the neutron diffraction results [67] regarding the protonation state of AZM in the crystal phase. Further examination of XModeScore results revealed that the ZDD of form **3** is twice as low as that of the other two forms, suggesting that protomer **3**, with the negatively charged N1 atom coordinated with zinc, is much more consistent with the experimental X-ray data than are the other two tautomers with the amino group at this position. Furthermore, the difference density maps of tautomers **1** and **2**, obtained after the QM refinement (Figure 7.5a,b), show prominent negative/positive difference density peaks around the nitrogen atom N1, which also support this conclusion. Notably, a series of refinements using incremental truncation of the original high-resolution data set 3HS4 demonstrates that XModeScore remains robust and predictive up to at least 3.0 Å resolution (Table 7.2).

7.7 Impact of the QM-Driven Refinement on Protein–Ligand Affinity Prediction

Given structural errors in crystal structures negatively impact ligand binding affinity predictions [2] and QM/MM crystallographic refinement improves the quality of protein–ligand structures, we have demonstrated [68] the impact of QM/MM refinement (and XModeScore) on protein–ligand binding affinity prediction. In

Table 7.2 XModeScore results for three forms of ligand AZM in PDB 3HS4 at different resolutions.

Pose	SE	RSCC	ZDD	XModeScore
Structure-3HS4				
3	5.55	0.989	12.8	2.72
2	8.89	0.978	24.9	−0.74
1	10.8	0.975	27.2	−1.98
Resolution 1.6 Å				
3	6.01	0.987	7.87	2.72
2	8.71	0.98	14.3	−0.70
1	9.75	0.978	16.8	−2.02
Resolution 2.0 Å				
3	5.58	0.989	6.56	2.68
2	8.74	0.982	12.3	−1.24
1	7.86	0.975	15.6	−1.45
Resolution 2.2 Å				
3	5.77	0.989	6.17	2.77
2	7.73	0.981	10.8	−1.31
1	8.35	0.984	10	−1.47
Resolution 2.5 Å				
3	5.4	0.989	7.65	2.47
2	8.2	0.986	8.62	−0.04
1	11.1	0.984	9.48	−2.43
Resolution 2.8 Å				
3	5.45	0.984	9.67	2.8
2	8.25	0.984	10.2	−1.39
1	8.74	0.982	10.1	−1.41

this work, QM/MM X-ray crystallographic refinement was applied to the set of structures from the community structure activity resource (CSAR) dataset released in 2012 [69]. This well-curated CSAR set is available with experimental binding affinities and is intended to be used as a benchmark for developing and testing docking and scoring functions. The CSAR set consists of the following subsets: cyclin-dependent kinase 2 (CDK2) with 15 ligands, checkpoint kinase 1 (CHK1) with 16 ligands, mitogen-activated protein kinase 1 (ERK2) with 12 ligands, and urokinase-type plasminogen activator (uPA) with 7 ligands.

As a baseline, we explored the impact of QM/MM refinement alone by measuring the R^2 correlation coefficients between experimental binding affinities and predicted

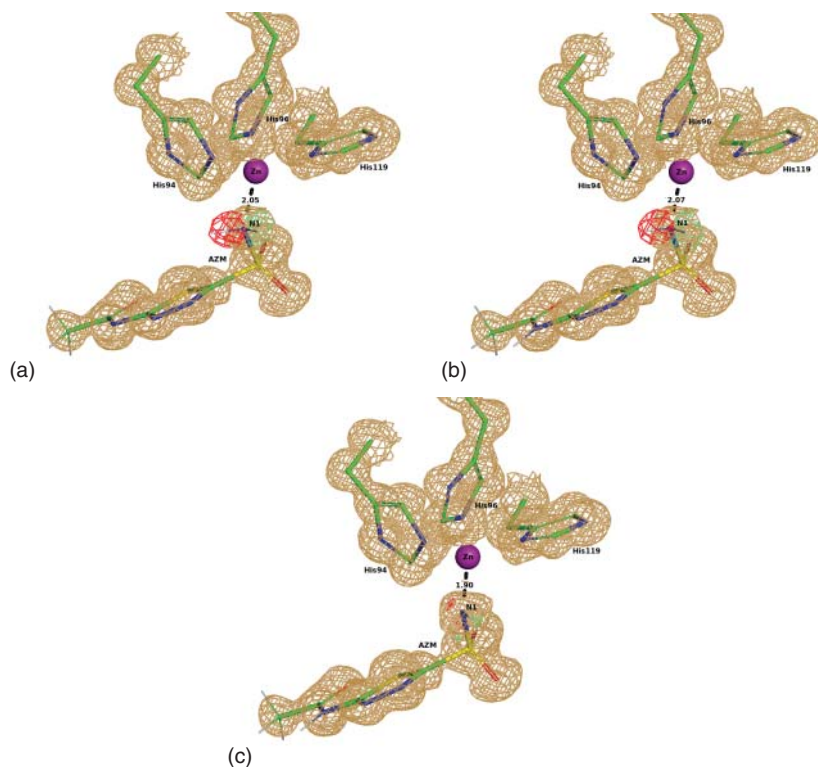


Figure 7.5 The coordination sphere of Zn in the catalytic center of HCA II with bound AZM molecules at three alternative binding modes **1** (a), **2** (b), and **3** (c) after the QM refinement of the PDB structure 3HS4. The difference density around the key nitrogen atom N1 of the sulfonamide group of AZM is contoured at 3.5σ level.

binding affinities as computed with a conventional, mainstream GBVI/WSA score function [70] available in the molecular operating environment (MOE) from Chemical Computing Group ULC. These results are depicted in Figure 7.6. QM/MM refinement improves the correlation within each subset, and the most significant magnitude of the transformation is observed for the CDK2 group, which comprises 15 structures [68]. Notably, the predicted affinities based on conventionally refined structures of this subset exhibit virtually no correlation to the experimental affinity with R^2 of 0.25 (Figures 7.6 and 7.7a), while QM/MM refined structures yield GBVI/WSA scores which correlate with the experimental binding affinity with R^2 of 0.60. Such a dramatic improvement is attributable to a much better fit of the crystal structures to the observed density in the QM/MM X-ray refinement, as these results show a twofold improvement of the average ZDD for the 15 CDK2 structures after the QM/MM refinement compared to the conventionally refined structure [68]. For example, consider the structure 4FKS, the worst outlier on the correlation graph based on the conventional refinement (Figure 7.7a). As a result of the QM/MM refinement ZDD around the ligand dramatically decreased from 16.63 to 3.86, mainly due to a different orientation of the ligand's benzyl moiety, as

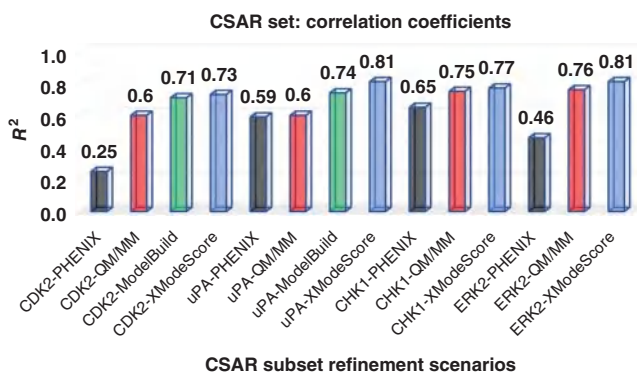


Figure 7.6 Correlation coefficients R^2 between experimental and predicted binding affinities for CSAR sets for various refinement scenarios: PHENIX (black), QM/MM (red), model-built QM/MM (green), and QM/MM with XModeScore chosen tautomers (blue).

seen in Figure 7.8. Interestingly, such an improvement model gives rise to a shift of the predicted GBVI/WSA score from -5.70 to -7.50 kcal/mol, bringing the new predicted value almost precisely on the trendline.

7.7.1 Impact of Structure Inspection and Modification

Even the most advanced refinements will ultimately only reach the nearest local minimum. Therefore, any structural changes and improvements resulting from the X-ray refinement are relatively limited in scope and achieved within a limited radius of convergence. Thus, the refinement cannot fix significant structural defects such as misplaced side chains or functional groups, including incorrect flip states, missing or extra water, etc. Visual inspection of the electron density maps usually allows one to easily spot problematic regions by the presence of prominent positive and negative peaks of the difference density, and manual model building is then required to fix such regions. Notably, these structure errors can even occur in well-curated crystal structures and could significantly impact the predicted binding scores.

For example, the correlation analysis of the uPA subset [68] revealed that the structure 4FU9 is one of the worst outliers, and the visual inspection of the ligand binding pocket in the structure revealed several problems, as shown in Figure 7.9a. First, the small peak of the positive difference density around the atom N18 indicates the presence of an alternative protonation state of the ligand 675 in the crystal. Second, there is a prominent peak of the negative electron density around the water molecule (WAT526) in the vicinity of the ligand, suggesting that this water is to be removed from the model. Third, the succinate molecule, SIN304, which is a crystallization buffer compound, was initially added to the model with an occupancy of 0.5. But a large blob of positive electron density observed around the molecule, suggests full occupancy of SIN304. With these changes made, the new QM/MM refinement leads to a better difference density distribution in the binding pocket (Figure 7.9b), and the GBVI/WSA score for the ligand 675 decreases from -7.11 to -6.40 kcal/mol. This shift leads to a considerable improvement in correlation for the

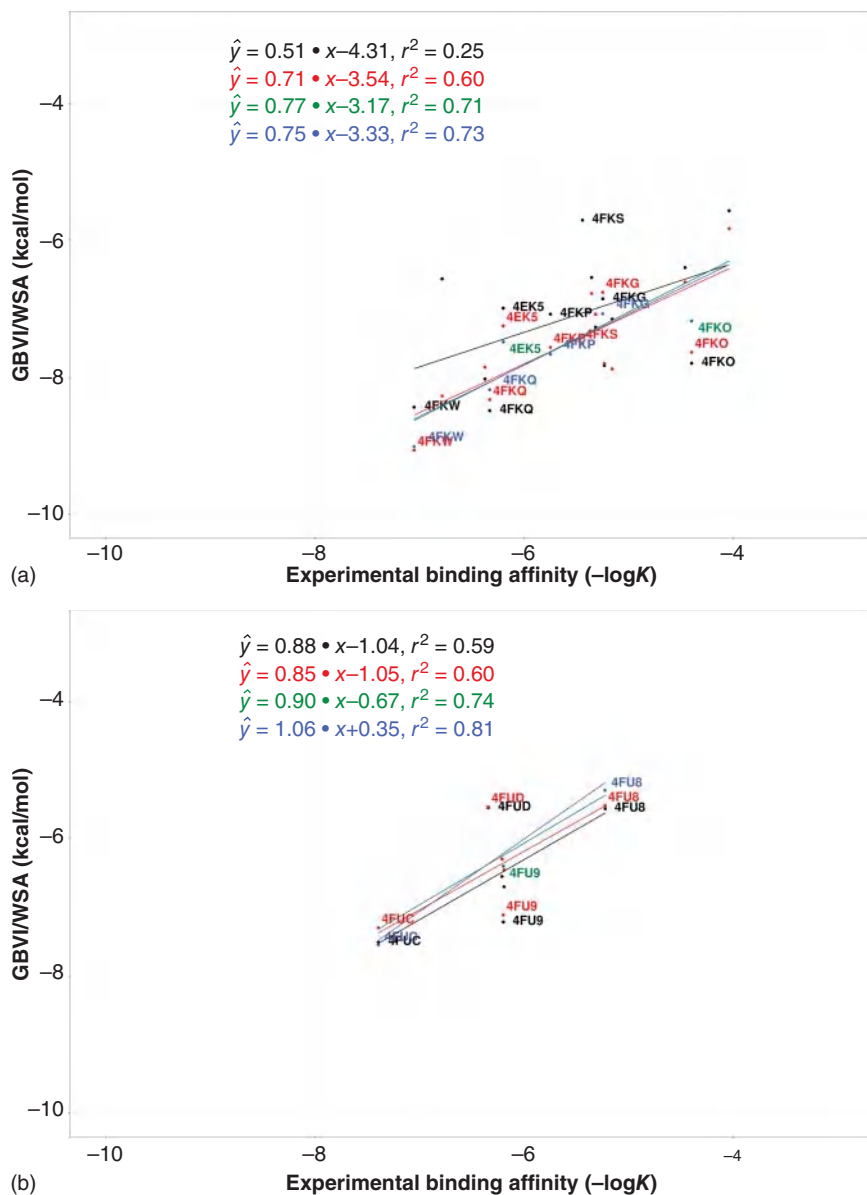


Figure 7.7 The regression lines of the correlation between experimental affinity ($-\log K$) and computationally predicted GBVI/WSA scores for the 15 protein–ligand CDK2 complexes (a) and the 7 protein–ligand uPA complexes (b) for PHENIX structures (black), QM/MM structures (red), model-built QM/MM structures (green), and QM/MM refined structures with XModeScore chosen tautomers (blue).

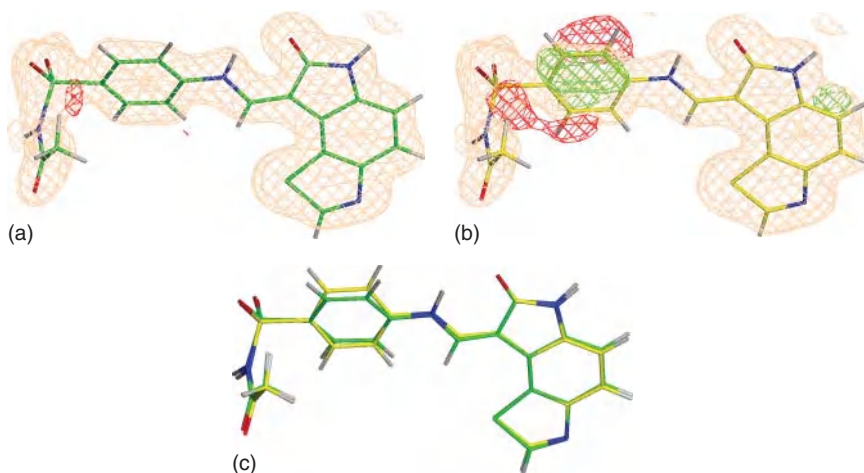


Figure 7.8 The σ_A -weighted mFo-DFc difference electron density map drawn at 3σ level around the ligand (ligand ID 46K) in the PDB structure 4FKS was refined with QM/MM (green) (a) and conventional (yellow) (b), as well as the superimposition of the two structures (c). The σ_A -weighted 2mFo-DFc electron density map is contoured at 1σ .

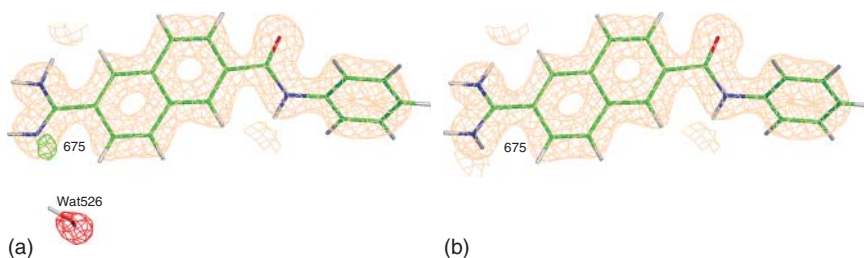


Figure 7.9 Positive (green) and negative (red) peaks of the σ_A -weighted mFo-DFc difference electron density map around the ligand (ligand ID 675) and Wat526 in the binding pocket of the protein target uPA in the PDB structure 4FU9 refined with QM/MM before (a) and after (b) the manual fit. The σ_A -weighted 2mFo-DFc electron density map is contoured at 1σ .

uPA set in which increases from R^2 0.6 to 0.74. A similar improvement in correlation was also achieved for the CDK2 subset by manual model correction (removing unjustified waters and choosing alternative side-chain positions of certain residues) of the worst outlier structures [68].

7.7.2 Impact of Selecting Protomer States: Implications of XModeScore on SBDD

As discussed above, the correct selection of the tautomeric/protonation state is often critical to the success of SBDD campaigns. In this work, we demonstrated that the choice of the correct ligand state significantly impacts the GBVI/WSA score and therefore the binding affinity of the method [68]. The application of

XModeScore to the CSAR set demonstrated that while often the default ligand protonation states chosen using default protonation are correct, alternative ligand protonation states are found to be correct in a significant plurality of the cases explored. Those structures are distributed across the subsets as follows: CDK2–4 structures, uPA – 3 structures, CHK1 2 structures, and one structure is from ERK2. Notably, including those alternative forms in the QM/MM refinement improves – to various degrees – the overall correlation between predicted and experimental affinities for all subsets, as shown in Figure 7.6.

When exploring the uPA subset, it was mentioned above that the protonation state of the ligand 675 in the structure 4FU9 was adjusted based on the study of the electron density map (Figure 7.9a). The XModeScore procedure also confirms that this ligand form – with a fully protonated amino(imino)methyl group – is the most favorable state (Figure 7.9b). The same alternative state with the fully protonated amino(imino)methyl group was also established by XModeScore for another ligand 2UP (PDB 4FU8) from the uPA subset, and using the alternative protomer in the QM/MM refinement resulted in the change of GBVI/WSA score from -5.52 to -5.29 for 4FU8. It was also found that in structure 4FUC that the default protonation state of the ligand 239 with the charged NH_3^+ group has a *worse* XModeScore score than that of the protomer the NH_2 group. Despite weaker H-bond interactions between the ammonia group and neighboring Asp50, the new-QM/MM refinement with NH_2 state reveals much better agreement with the experimental density as proven by a smaller ZDD value (1.97 units) compared to its magnitude of 4.42 units in the original QM/MM structure [68]. Overall results of incorporation of the correct protonation states according to XModeScore lead to a significant improvement in correlation for the uPA subset in which the R^2 increases from 0.74 to 0.81.

7.8 Conclusion

X-ray crystallography has become an integral tool in SBDD, and it provides the primary data used in new method development in CADD (including docking/scoring/sampling algorithm innovation, force field parameterization, and even training for artificial intelligence/machine learning (AI/ML)-based methods like AlphaFold [71]). While traditionally, computational chemists and medicinal chemists will begin with the X-ray structure, add protons, and complete an optimization process prior to using the model, QM/MM refinement is able to strike the right balance and provide insights into SBDD while still staying true to the experimental data. And given that QM/MM is built on QM for critical areas of the structure, it can account for interactions that are not properly represented in conventional, CIF-based methods like hydrogen bonds, electrostatics, charge transfer, polarization, and even metal coordination. We have been able to deploy these methods and make them available for routine use by also implementing mature molecular perception and preparation protocols, tautomer/protomer (and flip-state and chirality) enumeration methods, and density-based statistical analyses. QM/MM refinement allows us to not only better understand what target:ligand interactions

a particular ligand satisfies (and perhaps more importantly, what interactions a candidate molecule does not), but with XModeScore, these methods also give us a more accurate picture of more subtle effects due to protonation state, binding mode, rotamer position, and bound-isomer character.

Going forward, cryogenic electron microscopy (Cryo-EM) is another approach to structure solution, which has several strengths vs. X-ray crystallography [72–74]. For example, cryo-EM does not require protein–ligand crystallization, and therefore it is applicable to target classes that poorly crystallize (such as certain globular proteins, many membrane-bound proteins, and so on). This method has opened opportunities for SBDD for additional campaigns, but cryo-EM does suffer from several limitations. These include mixed resolution characteristics in which different parts of the structure (e.g. pocket vs. ligand vs. protein-bulk) may resolve at vastly different resolutions [72, 73], EM is often unable to obtain atomic-scale resolutions required for SBDD, greater disorder and heterogeneity since the sample is frozen instead of crystallized, and so on. Because QM and MM functionals include more information than stereochemical restraints alone, they are applicable to the mixed or lower quality data available in cryo-EM: often in X-ray, for example, as shown in Table 7.2, we obtain similar results regardless of resolution.

To explore the question of whether or not ZDD will also work with cryo-EM data and an estimated difference map [75], we implemented a proof-of-concept cryo-EM real-space refinement protocol using DivCon combined with the opensource Clipper v. 2.1.20201109 library [76] for real-space map and density gradient calculation. During this proof-of-concept study, DivCon was joined with `phenix.real_space_diff_map` tool within the PHENIX package [4, 75] to estimate a real-space difference density map between both the initial input model and map downloaded from the PDB and the final model and map resulting from 50 steps of real-space QM/MM cryo-EM refinement. These estimated maps were supplied to our built-in ZDD calculator, and as hoped, ZDD does indeed improve upon refinement. In fact, as shown in Table 7.3, we observe consistent improvement in each case in all considered metrics. We, therefore, expect that, given the improvement in both strain and ZDD that XModeScore will be applicable to Cryo-EM

Table 7.3 Cryo-EM RS-DivCon refinement preliminary results.

Cryo-EM PDBid	Res (Å)	Ligand	Strain _{Initial}	Strain _{Final}	ZDD _{Initial}	ZDD _{Final}	MP _{Initial}	MP _{Final}
7efc	1.7	BTN	39.65	15.02	7.89	6.27	1.49	0.84
7jsy	1.8	I3C	38.75	14.09	7.4	5.47	1.3	0.7
7sq6	2.3	AQV	32.71	11.97	17.57	13.38	1.4	1.17
7vdf	2.6	BMA	13.22	7.64	17.71	15.61	1.65	1.25
7p02	2.9	CLR	24.73	11.18	11.96	11.05	1.8	1.25

ZDD = Z score of the *estimated* difference density calculated with `phenix.real_space_diff_map`.

Strain = all-atom ligand strain (E_{strain}) of the Ligand calculated at the QM level of theory.

MP = MolProbity score calculated with the `phenix.molprobity` tool.

Initial = published structure/final = QM(ligand+3.0 Å pocket)/MM real-space refined structure.

refinement as we have already shown it to be with X-ray crystallographic refinement. Going forward, we have implemented a built-in real space difference map estimator in DivCon, and we will continue to explore the impact of local (pocket:ligand) resolution, bound-ligand disorder, and ensemble heterogeneity in XModeScore in order to provide actionable intelligence in cryo-EM-enabled SBDD campaigns.

QM/MM refinement – both with X-ray and cryo-EM data – is a logical next step in routine structure solution, and as these methods have become faster and more highly convergent, they have shown themselves to be extremely useful for SBDD and FBDD efforts in the pharmaceutical space both for the determination of better, more accurate models and for the characterization of states – like protonation states – that would be difficult if not impossible to characterize without access to the more accurate models. As we move forward, routine QM/MM X-ray and cryo-EM refinement will no doubt become as indispensable for structural biology teams as QM, QM/MM, and MM sampling and characterization methods have already become for computer-aided drug design campaigns.

Acknowledgments

The authors wish to acknowledge the support of our clients and users, who have provided valuable feedback. We also thank the continued support of the PHENIX Consortium, in particular Drs. Nigel Moriarty, Pavel Afonine, and Paul Adams, for maintaining the application programming interface (API) “hooks” to our software within PHENIX. Likewise, we thank Global Phasing Limited, in particular Drs. Clemens Vornrhein and Gerard Bricogne, for supporting our development of analogous hooks with BUSTER. We would also like to thank Chemical Computing Group (in particular Alain Deschenes, Chris Williams, Paul Labute, and the entire CCG support team) for their continued support with MOE best practices and with the scientific vector language (SVL). Finally, we thank the National Institutes of Health (NIH) through SBIRs #R44GM121162 and #R44GM134781 for funding the research and development effort. The DivCon plugin to PHENIX and BUSTER along with XModeScore are provided by QuantumBio Inc. and they are available at the following: https://www.quantumbioinc.com/products/software_licensing.

References

- 1 Chilingaryan, Z., Yin, Z., and Oakley, A.J. (2012). Fragment-based screening by protein crystallography: successes and pitfalls. *Int J Mol Sci* 13: 12857.
- 2 Davis, A.M., Teague, S.J., and Kleywegt, G.J. (2003). Application and limitations of X-ray crystallographic data in structure-based ligand and drug design. *Angew Chem Int Ed* 42: 2718–2736.
- 3 Davis, I.W., Leaver-Fay, A., Chen, V.B. et al. (2007). MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res* 35: W375–W383.

- 4 Adams, P.D., Afonine, P.V., Bunkoczi, G. et al. (2010). PHENIX: a comprehensive python-based system for macromolecular structure solution. *Acta Cryst Sect D* 66: 213–221.
- 5 Kleywegt, G.J. (2007). Crystallographic refinement of ligand complexes. *Acta Cryst Sect D* 63: 94–100.
- 6 Kleywegt, G.J., Henrick, K., Dodson, E.J., and van Aalten, D.M.F. (2003). Pound-wise but penny-foolish: how well do micromolecules fare in macromolecular refinement? *Structure* 11: 1051–1059.
- 7 Read, R.J., Adams, P.D., Arendall, W.B. III, et al. (2011). A new generation of crystallographic validation tools for the protein data bank. *Structure* 19: 1395–1412.
- 8 Ryde, U., Olsen, L., and Nilsson, K. (2002). Quantum chemical geometry optimizations in proteins using crystallographic raw data. *J Comput Chem* 23: 1058–1070.
- 9 Caldararu, O., Manzoni, F., Oksanen, E. et al. (2020). Refinement of protein structures using a combination of quantum-mechanical calculations with neutron and X-ray crystallographic data. Corrigendum. *Acta Crystallogr D Struct Biol* 76: 85–86.
- 10 Nilsson, K. and Ryde, U. (2004). Protonation status of metal-bound ligands can be determined by quantum refinement. *J Inorg Biochem* 98: 1539–1546.
- 11 Rulisek, L. and Ryde, U. (2006). Structure of reduced and oxidized manganese superoxide dismutase: a combined computational and experimental approach. *J Phys Chem B* 110: 11511–11518.
- 12 Ryde, U. and Nilsson, K. (2003). Quantum chemistry can locally improve protein crystal structures. *J Am Chem Soc* 125: 14232–14233.
- 13 Yu, N., Yennawar, H.P., and Merz, K.M. (2005). Refinement of protein crystal structures using energy restraints derived from linear-scaling quantum mechanics. *Acta Cryst Sect D* 61: 322–332.
- 14 Bergmann, J., Oksanen, E., and Ryde, U. (2022). Combining crystallography with quantum mechanics. *Curr Opin Struct Biol* 72: 18–26.
- 15 Fu, Z., Li, X., and Merz, K.M. (2011). Accurate assessment of the strain energy in a protein-bound drug using QM/MM X-ray refinement and converged quantum chemistry. *J Comput Chem* 32: 2587–2597.
- 16 Li, X., Hayik, S.A., and Merz, K.M. (2010). QM/MM X-ray refinement of zinc metalloenzymes. *J Inorg Biochem* 104: 512–522.
- 17 Yu, N., Li, X., Cui, G. et al. (2006). Critical assessment of quantum mechanics based energy restraints in protein crystal structure refinement. *Protein Sci* 15: 2773–2784.
- 18 Yu, N., Hayik, S.A., Wang, B. et al. (2006). Assigning the protonation states of the key aspartates in β -secretase using QM/MM X-ray structure refinement. *J Chem Theory Comput* 2: 1057–1069.
- 19 Borbulevych, O.Y., Plumley, J.A., Martin, R.I. et al. (2014). Accurate macromolecular crystallographic refinement: incorporation of the linear scaling, semiempirical quantum-mechanics program DivCon into the PHENIX refinement package. *Acta Cryst Sect D* 70: 1233–1247.

- 20 QuantumBio Inc. (2015) LibQB. (Inc, Q., ed., 6.0, www.quantumbioinc.com (Ed.), QuantumBio Inc.
- 21 Dixon, S.L. and Merz, K.M. (1996). Semiempirical molecular orbital calculations with linear system size scaling. *J Chem Phys* 104: 6643–6649.
- 22 Dixon, S.L. and Merz, K.M. (1997). Fast, accurate semiempirical molecular orbital calculations for macromolecules. *J Chem Phys* 107: 879–893.
- 23 Diller, D.J., Humblet, C., Zhang, X.H., and Westerhoff, L.M. (2010). Computational alanine scanning with linear scaling semiempirical quantum mechanical methods. *Proteins* 78: 2329–2337.
- 24 Raha, K., Peters, M.B., Wang, B. et al. (2007). The role of quantum mechanics in structure-based drug design. *Drug Discov Today* 12: 725–731.
- 25 Raha, K., van der Vaart, A.J., Riley, K.E. et al. (2005). Pairwise decomposition of residue interaction energies using semiempirical quantum mechanical methods in studies of protein-ligand interaction. *J Am Chem Soc* 127: 6583–6594.
- 26 van der Vaart, A. and Merz, K.M. (1999). Divide and conquer interaction energy decomposition. *J Phys Chem A* 103: 3321–3329.
- 27 Zhang, X.H., Gibbs, A.C., Reynolds, C.H. et al. (2010). Quantum mechanical pairwise decomposition analysis of protein kinase B inhibitors: validating a new tool for guiding drug design. *J Chem Inf Model* 50: 651–661.
- 28 Vreven, T., Morokuma, K., Farkas, Ö. et al. (2003). Geometry optimization with QM/MM, ONIOM, and other combined methods. I. Microiterations and constraints. *J Comput Chem* 24: 760–769.
- 29 Borbulevych, O., Martin, R.I., and Westerhoff, L.M. (2018). High-throughput quantum-mechanics/molecular-mechanics (ONIOM) macromolecular crystallographic refinement with PHENIX/DivCon: the impact of mixed Hamiltonian methods on ligand and protein structure. *Acta Cryst Sect D* 74: 1063–1077.
- 30 Brünger, A.T., Adams, P.D., Clore, G.M. et al. (1998). Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr D Biol Crystallogr* 54: 905–921.
- 31 Van der Vaart, A., Gogonea, V., Dixon, S.L., and Merz, K.M. (2000). Linear scaling molecular orbital calculations of biological systems using the semiempirical divide and conquer method. *J Comput Chem* 21: 1494–1504.
- 32 Van der Vaart, A., Suarez, D., and Merz, K.M. (2000). Critical assessment of the performance of the semiempirical divide and conquer method for single point calculations and geometry optimizations of large chemical systems. *J Chem Phys* 113: 10512–10523.
- 33 Wang, B., Westerhoff, L.M., and Merz, K.M. (2007). A critical assessment of the performance of protein-ligand scoring functions based on NMR chemical shift perturbations. *J Med Chem* 50: 5128–5134.
- 34 Dewar, M.J.S., Zoebisch, E.G., Healy, E.F., and Stewart, J.J.P. (1985). The development and use of quantum-mechanical molecular-models. 76. Am1 – a new general-purpose quantum-mechanical molecular-model. *J Am Chem Soc* 107: 3902–3909.
- 35 Rezac, J., Fanfrlik, J., Salahub, D., and Hobza, P. (2009). Semiempirical quantum chemical PM6 method augmented by dispersion and H-bonding correction

- terms reliably describes various types of noncovalent complexes. *J Chem Theory Comput* 5: 1749–1760.
- 36 Stewart, J.J.P. (2009). Application of the PM6 method to modeling proteins. *J Mol Model* 15: 765–805.
 - 37 Case, D.A., Babin, V., Berryman, J.T. et al. (2014). *AMBER 14*. San Francisco: University of California.
 - 38 Bricogne, G., Blanc, E., Brandl, M. et al. (2017). *BUSTER*. Cambridge, United Kingdom: Global Phasing Ltd.
 - 39 QuantumBio. (2022) LibQB. (Inc, Q. ed., 7.0, www.quantumbioinc.com (Ed.), QuantumBio Inc.
 - 40 Bietz, S., Urbaczek, S., Schulz, B., and Rarey, M. (2014). Protoss: a holistic approach to predict tautomers and protonation states in protein-ligand complexes. *J Chem* 6: 12.
 - 41 Labute, P. (2005). On the perception of molecules from 3D atomic coordinates. *J Chem Inf Model* 45: 215–221.
 - 42 Mukhopadhyay, A.K. and Mukherjee, N.G. (1981). Self-consistent methods in Hückel and extended Hückel theories. *Int J Quant Chem* 19: 515–519.
 - 43 Tickle, I. (2012). Statistical quality indicators for electron-density maps. *Acta Cryst Sect D* 68: 454–467.
 - 44 Chen, V.B., Arendall, W.B., Headd, J.J. et al. (2010). MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Cryst Sect D* 66: 12–21.
 - 45 Janowski, P.A., Moriarty, N.W., Kelley, B.P. et al. (2016). Improved ligand geometries in crystallographic refinement using AFITT in PHENIX. *Acta Cryst Sect D* 72: 1062–1072.
 - 46 Mobley, D.L. and Dill, K.A. (2009). Binding of small-molecule ligands to proteins: “what you see”; is not always “what you get”. *Structure* 17: 489–498.
 - 47 Perola, E. and Charifson, P.S. (2004). Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding. *J Med Chem* 47: 2499–2510.
 - 48 Fu, Z., Li, X., and Merz, K.M. (2012). Conformational analysis of free and bound retinoic acid. *J Chem Theory Comput* 8: 1436–1448.
 - 49 Borbulevych, O.Y., Plumley, J.A., and Westerhoff, L.M. (2012). Systematic study of the ligand strain energy derived from the quantum mechanics crystallographic refinement using the linear scaling program DivCon integrated into the PHENIX package. *Abstr Pap Am Chem Soc* 478.
 - 50 Borbulevych, O., Martin, R.I., Tickle, I.J., and Westerhoff, L.M. (2016). XMod-eScore: a novel method for accurate protonation/tautomer-state determination using quantum-mechanically driven macromolecular X-ray crystallographic refinement. *Acta Cryst Sect D* 72: 586–598.
 - 51 MacCallum, J.L., Hua, L., Schnieders, M.J. et al. (2009). Assessment of the protein-structure refinement category in CASP8. *Proteins* 77: 66–80.
 - 52 Word, J.M., Lovell, S.C., LaBean, T.H. et al. (1999). Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. *J Mol Biol* 285: 1711–1733.

- 53 Cozier, G.E., Leese, M.P., Lloyd, M.D. et al. (2010). Structures of human carbonic anhydrase II/inhibitor complexes reveal a second binding site for steroidal and nonsteroidal inhibitors. *Biochemistry* 49: 3464–3476.
- 54 Harding, M.M. (1999). The geometry of metal-ligand interactions relevant to proteins. *Acta Cryst Sect D* 55: 1432–1443.
- 55 Allen, F.H., Kennard, O., Watson, D.G. et al. (1987). Tables of bond lengths determined by X-ray and neutron-diffraction. 1. Bond lengths in organic-compounds. *J Chem Soc Perkin Trans 2*: S1–S19.
- 56 Adams, P.D., Pannu, N.S., Read, R.J., and Brunger, A.T. (1997). Cross-validated maximum likelihood enhances crystallographic simulated annealing refinement. *Proc Natl Acad Sci U S A* 94: 5018–5023.
- 57 Afonine, P.V., Grosse-Kunstleve, R.W., Echols, N. et al. (2012). Towards automated crystallographic structure refinement with phenix.refine. *Acta Cryst Sect D* 68: 352–367.
- 58 Hartshorn, M.J., Verdonk, M.L., Chessari, G. et al. (2007). Diverse, high-quality test set for the validation of protein-ligand docking performance. *J Med Chem* 50: 726–741.
- 59 Rupp, B. (2009). *Biomolecular crystallography: principles, practice, and application to structural biology*. Garland Science.
- 60 Ryde, U. and Nilsson, K. (2003). Quantum refinement—a combination of quantum chemistry and protein crystallography. *J Mol Struct THEOCHEM* 632: 259–275.
- 61 Martin, Y.C. (2009). Let's not forget tautomers. *J Comput Aided Mol Des* 23: 693–704.
- 62 USP-DI (1995). *United States pharmacopeia*, 15th, 659. Rockville, MD: The United States Pharmacopeial Convention Inc.
- 63 Moldow, B., Sander, B., Larsen, M., and Lund-Andersen, H. (1999). Effects of acetazolamide on passive and active transport of fluorescein across the normal BRB. *Invest Ophthalmol Vis Sci* 40: 1770–1775.
- 64 Krishnamurthy, V.M., Kaufman, G.K., Urbach, A.R. et al. (2008). Carbonic anhydrase as a model for biophysical and physical-organic studies of proteins and protein-ligand binding. *Chem Rev* 108: 946–1051.
- 65 Merz, K.M. and Banci, L. (1997). Binding of bicarbonate to human carbonic anhydrase II: a continuum of binding states. *J Am Chem Soc* 119: 863–871.
- 66 Sippel, K.H., Robbins, A.H., Domsic, J. et al. (2009). High-resolution structure of human carbonic anhydrase II complexed with acetazolamide reveals insights into inhibitor drug design. *Acta Cryst Sect F* 65: 992–995.
- 67 Fisher, S.Z., Aggarwal, M., Kovalevsky, A.Y. et al. (2012). Neutron diffraction of acetazolamide-bound human carbonic anhydrase II reveals atomic details of drug binding. *J Am Chem Soc* 134: 14726–14729.
- 68 Borbulevych, O.Y., Martin, R.I., and Westerhoff, L.M. (2021). The critical role of QM/MM X-ray refinement and accurate tautomer/protomer determination in structure-based drug design. *J Comput Aided Mol Des* 35: 433–451.

- 69 Dunbar, J.B. Jr., Smith, R.D., Damm-Ganamet, K.L. et al. (2013). CSAR data set release 2012: ligands, affinities, complexes, and docking decoys. *J Chem Inf Model* 53: 1842–1852.
- 70 Corbeil, C.R., Williams, C.I., and Labute, P. (2012). Variability in docking success rates due to dataset preparation. *J Comput Aided Mol Des* 26: 775–786.
- 71 Jumper, J., Evans, R., Pritzel, A. et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596: 583–589.
- 72 Wang, H.W. and Wang, J.W. (2017). How cryo-electron microscopy and X-ray crystallography complement each other. *Protein Sci* 26: 32–39.
- 73 Merino, F. and Raunser, S. (2017). Electron cryo-microscopy as a tool for structure-based drug development. *Angew Chem Int Ed* 56: 2846–2860.
- 74 Shoemaker, S.C. and Ando, N. (2018). X-rays in the cryo-electron microscopy era: structural biology’s dynamic future. *Biochemistry* 57: 277–285.
- 75 Afonine, P.V., Klaholz, B.P., Moriarty, N.W. et al. (2018). New tools for the analysis and validation of cryo-EM maps and atomic models. *Acta Crystallogr D Struct Biol* 74: 814–840.
- 76 McNicholas, S., Croll, T., Burnley, T. et al. (2018). Automating tasks in protein structure determination with the clipper python module. *Protein Sci* 27: 207–216.

8

Quantum-Chemical Analyses of Interactions for Biochemical Applications

Dmitri G. Fedorov

Research Center for Computational Design of Advanced Functional Materials (CD-FMat), National Institute of Advanced Industrial Science and Technology (AIST), Central 2, Umezono 1-1-1, Tsukuba 305-8568, Japan

8.1 Introduction

Interactions [1] in molecular systems determine their behavior. Building reliable models of interactions is the pivotal task of theoretical physics and chemistry. Quantum-mechanical (QM) methods are suitable for building these models because they are capable of describing all kinds of interactions in chemistry, including those that are difficult to obtain in faster models such as force fields.

The application of QM methods to biomolecules is challenging, given the high computational cost of these simulations. Biomolecules are not only large, but they also have the additional complexity of flexibility, which needs to be taken into account by sampling the conformational space at a given temperature and computing free energy. For biochemical processes, it may be necessary to describe a solvent, and other constituents of solutions, such as counterions.

Driven by the progress in method development and computational hardware, QM methods are increasingly used in computational drug discovery [2, 3]. There are a variety of fragment-based methods [4, 5], which not only reduce the computational cost but also deliver the properties of molecular parts, which can be useful for fragment-based drug discovery (FBDD) [6]. Interaction energies are used in structure- and interaction-based drug design (SIBDD) [7] for the rational enhancement of binding affinities.

In practice, some compromises have to be made in the application of QM methods to biochemical studies. Atomic structures are often refined in molecular mechanics (MM) simulations, although there are fast parametrized QM methods [8], such as density-functional tight-binding (DFTB) [9] that are adequate for full geometry optimizations of biochemical systems. These parametrized methods can also be used for molecular dynamics (MD) simulations, taking into account temperature effects.

Dividing biochemical systems into small parts (fragments) is very attractive because not only is the computational cost drastically reduced, but detailed properties of fragments can also be obtained. Among them, interactions between

fragments are very useful, providing quantitative information on the role of residues, or individual functional groups, for molecular recognition in protein-ligand, protein-DNA, and protein-protein binding. Likewise, enzymes can be treated gaining valuable insight into the contributions of residues in lowering a reaction barrier.

The energy decomposition analysis (EDA) [10, 11] has been an important conceptual starting point for the fragment molecular orbital (FMO) method [12–15]. FMO-based analyses suitable for biochemical studies are presented here, with a brief methodological description and a review of their applications.

8.2 Introduction to FMO

In FMO, a molecular system is divided into fragments. There are well-defined patterns for fragmenting all major types of biochemical systems, which are accomplished automatically using graphical user interfaces, such as Facio [16]. Polypeptides (proteins) can be automatically divided into amino acid residues (Figure 8.1), polynucleotides (DNA and RNA) into nucleotides, polysaccharides (cellulose, heparin, etc.) into saccharides, and explicit solvents into individual molecules.

It is possible either to calculate whole proteins [17] or truncate [18] them, keeping only residues within a suitable threshold (~ 5 Å) from the ligand [19], although it was argued [20] that for charged ligands the threshold has to be larger (by ~ 1 Å when optimizing geometry).

A pattern for fragmentation is chosen to balance two factors: fragments should be as close as possible to commonly used biochemical units, and the error of fragmentation (the difference between FMO energies and the values without fragmentation) should be as small as possible. In FMO, fragments are similar but not identical to the commonly used biochemical units. Namely, for proteins, a residue fragment in FMO has the same chemical composition as a conventional residue, but the border between two fragments is shifted so that the fragment for residue i includes the carbonyl of residue $i-1$ rather than its own carbonyl (Figure 8.1). For polynucleotides, fragments have the same composition as conventional nucleotides, but phosphate

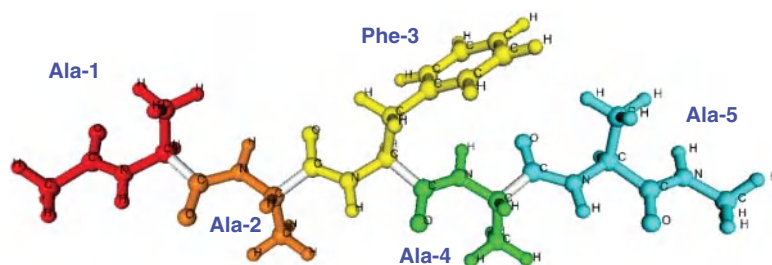


Figure 8.1 Automatic fragmentation of the AAFAA polypeptide into five residue fragments, whose names by convention include a dash to distinguish them from conventional residues. Terminal fragments include caps.

groups are likewise assigned to adjacent fragments. This is done to avoid having a fragment boundary at C—N (peptide) and P—O (nucleotide) bonds, which have a strong delocalized character, ill-suited for a QM fragmentation.

After N fragments are defined, FMO calculations can be conducted for a chosen QM level: wave function, basis set, and solvent model. First, individual fragments (monomers) are calculated in the electrostatic (ES) embedding, followed by calculations of pairs (dimers), and, optionally, triples (trimers). Combining these results, one obtains the total energy E . In a three-body expansion (FMO3), the energy is

$$E = \sum_I^N E'_I + \sum_{I>J}^N \Delta E_{IJ} + \sum_{I>J>K}^N \Delta E_{IJK} \quad (8.1)$$

where E'_I is the internal energy of polarized fragment I , ΔE_{IJ} is the pair interaction energy (PIE) between fragments I and J , and ΔE_{IJK} is the coupling of pair interactions in trimer IJK . The most commonly used method is FMO2, in which the last term in Eq. (8.1) is omitted.

Polarization energies [21] can be obtained by computing fragments with and without the electrostatic embedding. In most FMO analyses, polarization is contained implicitly without an explicit separation of a polarization contribution.

It is possible to compute analytic gradients of E , optimize geometry [22], and perform MD simulations using FMO [23]. Molecular structures can be refined using the frozen domain FMO [20, 24] and FMO-DFTB [25] methods. The latter approach can be used for FMO/MD simulations [26]. Partial geometry optimizations with density functional theory (DFT) and full optimizations with DFTB can be conducted with FMO for realistic atomic models containing thousands of atoms. Infrared (IR) and Raman spectra of proteins can be computed [27, 28].

Solvent can be treated both as explicit molecules or implicitly as a continuum in the polarizable continuum model (PCM) [29] and the solvent model density (SMD) method [30]. Analyses with explicit solvents are complicated by the conformational aspect, whereas implicit continuum models are easy to use. Because biochemical processes typically involve charged species in solution, solvent effects are of paramount importance.

Periodic boundary conditions (PBC) can be used for FMO-DFTB [31], making it possible to compute liquids and solutions [32], molecular crystals (e.g. of ice [33] and proteins [34]), and solid state of inorganic materials [35]. Some analyses can be combined with PBC, as described below.

QM calculations with FMO-DFTB can be conducted for molecular systems containing more than 1 million atoms [36, 37], whereas *ab initio* methods (second-order Møller–Plesset perturbation theory, MP2) for thousands of atoms can be routinely done [17].

Analyses described in this chapter can be performed with FMO implemented [38–40] in GAMESS [41]. Molecular electrostatic potential (MEP) [42], taking into account polarization and charge transfer, can be computed using FMO to guide ligand docking. MEP can be used to visualize electrostatic complementarity, for example, in protein–protein complexes [43].

8.3 Pair Energy Decomposition Analysis (PIEDA)

Pair interaction energies ΔE_{IJ} in Eq. (8.1) quantify the strength of interactions, often taken as a measure of binding, although binding and interaction energies are not the same. Binding energies are defined with respect to the reference state of isolated, noninteracting species, whereas interaction energies in FMO are defined with respect to polarized fragments [40].

One can argue that for molecules, the isolated state might be a better reference, whereas for covalently bound fragments (as residues in a protein), the naturally embedded polarized state is more suitable.

For example, a PIE between two polarized water molecules in a water droplet can be taken as a measure of the hydrogen bond energy. In a vacuum, a pair interaction may be slightly more attractive than the corresponding binding energy because it includes the stabilization part of the polarization [21]. In solution, interaction and binding energies differ more, as explained below.

PIEs between two fragments connected by a covalent bond reflect the effects of separating electrons in the bond detached atom (BDA) on the interfragment border between two fragments. Such PIEs are quite large, and they are seldom used in analyses, although it is possible to split the artificial component using BDA corrections [44].

PIEs can be decomposed into components for the purpose of gaining deeper physicochemical insight. A decomposition has to be designed for each QM method and solvent model, yielding quantitative information about the role of different physical interactions as described next.

8.3.1 Formulation of PIEDA

The pair interaction energy decomposition analysis (PIEDA) [44–46] is based on ideas from EDA [10]. The original PIEDA was formulated for FMO2 [44], later generalized to FMO3 [47]. These decompositions can be thought of as FMO*n*-EDA combined with EDA, so they are referred to as FMO*n*-EDA (PIEDA is FMO2-EDA). FMO1-EDA is used for defining polarization energies.

PIEDA can be done for DFTB, Hartree-Fock (HF), DFT, second-order Møller-Plesset perturbation theory (MP2), and coupled cluster, e.g. CCSD(T). The components in PIEDA differ depending on the QM method.

Taking as an example PIEDA for MP2, a commonly used method in biochemical applications, the decomposition of a PIE is

$$\Delta E_{IJ} = \Delta E_{IJ}^{ES} + \Delta E_{IJ}^{EX} + \Delta E_{IJ}^{CT+MIX} + \Delta E_{IJ}^{DI+RC} + \Delta E_{IJ}^{solv} \quad (8.2)$$

with the following components: electrostatic (ES), exchange-repulsion (EX), charge transfer (CT) and mix coupling terms (CT + MIX), dispersion (DI) and remainder correlation (RC), and solvent screening (solv). The convention is to use capital and small letters for solute and solvent terms, respectively.

The ES term is the Coulomb interaction between two polarized fragments, including the stabilization component of the polarization. ES describes the interaction

between charge distributions (electron density clouds and point charges of protons in the nuclei) in a vacuum. This interaction can be very strong for charged fragments. In solution, there is a solvent screening solv term, which typically reduces the Coulomb (ES) interaction. The solv term is computed as

$$\Delta E_{IJ}^{\text{solv}} = \Delta E_{IJ}^{\text{es}} + \Delta E_{IJ}^{\text{non-es}} \quad (8.3)$$

where es and non-es are the solute-solvent electrostatic and non-electrostatic screening interactions, respectively.

The es term is defined in continuum models, PCM or SMD, whereas the non-es term is present in PCM only. The es term is computed as the interaction of the solute charge distribution with induced solvent charges of the solvent. There are two models for the es term: local [45] and partial [48]. They differ in the definition of solvent charges. In the local model, solvent charges induced by the combined potential of all fragments are divided among fragments geometrically. In the partial model, solvent charges are induced by the partial potential of individual fragments. The charge quenching effect [45] (the cancelation of the solvent charges due to the partial potentials of oppositely charged fragments) is responsible for a large underestimation of the solvent screening in the local model. So the partial model, which has some extra cost, is the preferred way of defining solvent screening.

It is useful to add mutually canceling ES and es terms, producing the solute-solute electrostatic interaction screened in solution (ES + es). The ES and ES + es terms are long-ranged interactions, slowly decaying with interfragment separation. If two fragments *I* and *J* are sufficiently separated, the interaction energy can be computed as

$$\Delta E_{IJ} \approx \Delta E_{IJ}^{\text{ES}} + \Delta E_{IJ}^{\text{solv}} \quad (8.4)$$

which reduces the cost of FMO calculations very considerably.

The exchange-repulsion (EX) interaction arises due to the Pauli exclusion principle, describing the repulsion of two fermions (electrons). It corresponds to the repulsive part of the Lennard-Jones potential. In QM methods, the EX term is rigorously computed based on the wave function. Without this repulsion, two ions of the opposite charge would stick to each other. EX is a short-ranged interaction, which arises whenever two fragments are strongly attracted to each other. Thus, EX is an inevitable companion of a strong attraction. However, EX may be substantial without a strong attraction due to a steric repulsion in a poorly optimized structure. If a large EX term is found without other attractive terms, it is an indication of a need to refine the structure, although it may be inevitable that for one pair of fragments to be strongly attracted, another pair may be forced into repulsion.

The RC + DI term is the contribution of the electron correlation, some part of which is the dispersion (DI) interaction, and the rest is called the RC. For DFT with empirical dispersion, the RC and DI terms are separable. The DI term corresponds to the attractive term of the Lennard-Jones potential describing the van-der-Waals interactions, as pertinent to hydrophobic contacts in biochemical systems. The RC term describes non-dispersive interaction due to the electron correlation [49].

Basis set superposition error (BSSE) can be corrected using the auxiliary polarization (AP) method [50] or HF-3c [51], with a basis set (BS) term $\Delta E_{IJ}^{\text{BS}}$ added to

Eq. (8.2). HF-3c is Hartree-Fock with three corrections (3c): empirical dispersion, short- and long-ranged BSSE corrections. For PIEDA/AP, each interaction term (see Eq. (8.2)) can be BSSE-corrected [40].

In DFTB, the RC and EX terms are combined in the so-called 0-order term (0-order refers to the Taylor expansion of the electron density in DFTB), and due to the parametrization, the two components cannot be separated.

In FMO3/EDA, a three-body interaction ΔE_{IJK} in Eq. (8.1) can be decomposed for MP2 as

$$\Delta E_{IJK} = \Delta E_{IJK}^{EX} + \Delta E_{IJK}^{CT+MIX} + \Delta E_{IJK}^{RC+DI} + \Delta E_{IJK}^{solv} \quad (8.5)$$

Comparing Eqs. (8.2) and (8.5), it can be seen that an ES term is absent in the latter. It is because ES is purely two-body without three-body corrections. The same applies to DI and BS in empirical models (HF-3c).

Sometimes, FMO/EDA3 terms are compressed [52, 53] into three-body corrected effective two-body terms. This can be done for the total PIE and for individual components. The contracted PIE is defined as,

$$\Delta \tilde{E}_{IJ} = \Delta E_{IJ} + \frac{1}{3} \sum_{K \neq I, J}^N \Delta E_{IJK} \quad (8.6)$$

The total interaction energy (TIE) can be computed for the binding of a ligand I to a protein via summing over residues J in the protein as

$$\Delta E_I = \sum_{J \in \text{protein}} \Delta E_{IJ} \quad (8.7)$$

For drug design, it may be useful to split a large ligand into several fragments. Then, for each ligand fragment I , its fragment efficiency [54] can be defined using the number of heavy (non-hydrogen) atoms N_I as

$$\Delta \bar{E}_I = \frac{\Delta E_I}{N_I} \quad (8.8)$$

By comparing $\Delta \bar{E}_I$ for different ligand fragments I , a decision can be made to replace or remove those parts that contribute little, guiding drug design.

Repulsions can be excluded from analyses by defining $\Delta \tilde{E}_I$ as ΔE_I minus the EX interaction. It can be useful if the structure optimization is not done with the same QM method as the analysis.

For a clear labeling of residue contributions, it was suggested [55] to define the fraction of a component A to binding,

$$f_I^A = \frac{\Delta E_I^A}{\Delta \tilde{E}_I} \quad (8.9)$$

where A can be ES, DI, or CT + MIX (in the original scheme, the solvent screening was not considered). In other words, for each residue I , its ligand binding is represented by a composite “color” assigned in an RGB-like scheme, with three primary colors (A = electrostatic, dispersion, or charge transfer) mixed with fractions f_I^A . The fractions are normalized to 1 for each I ,

$$\sum_A f_I^A = 1 \quad (8.10)$$

Using the fractions, signature ratios were defined [55],

$$1 : \frac{f_I^{DI}}{f_I^{ES}} : \frac{f_I^{CT+MIX}}{f_I^{ES}} \quad (8.11)$$

with which the physical type of the contribution of residue *I* can be easily identified. Using the fractions in Eq. (8.11), it was shown [55] that dispersion and electrostatics contribute comparably to a wide range of representative GPCR proteins. As a further simplification, fractions can be grouped into two, rather than three, categories: electrostatic (ES plus CT + MIX) and dispersion (DI) [56, 57].

PIEs can be used as descriptors in structure–activity relationships (SAR) [58], in particular, in quantitative SAR (QSAR) [59]. There is an ongoing effort to use PIE [60] and TIE [61] in combination with machine learning, to predict free energies of binding based on FMO.

8.3.2 Applications of PIEs and PIEDA

For cross-validation, PIEs were compared for two different QM methods, and TIEs were compared to the experiment, as summarized in Table 8.1. There are two aspects of the validation: verify that QM results have a good correlation to the experiment (of course, not perfect, because some effects such as entropy are not accounted for in QM calculations) and test if very fast QM methods like DFTB or HF-3c can be reliably used for a high-throughput drug design pipelines.

Questions that can be answered by considering the signatures of PIE components include: should CH...O bonds be classified as weak hydrogen bonds or do they have a different physical nature? Answer: the latter [67]. What is the nature of CH- π bonding that is frequently found in biochemical systems [68]? A singular value decomposition can be applied to PIEs [69] for a clear identification of the principal factors in protein–ligand binding. Some representative applications of PIEs are summarized in Table 8.2.

8.3.3 Example of PIEDA

To show the kind of data that can be obtained, PIEDA is applied to a protein–ligand complex of Trp-cage (PDB: 1L2Y) with deprotonated *p*-phenolic acid (an anion) [20] (Figure 8.2). The structures of the complex and isolated proteins and ligands were optimized at the level of FMO2-DFTB3/D3(BJ)/C-PCM<1> using 3ob parameters [81]. The protein was divided into 1 residue per fragment, and the ligand was a separate fragment.

PIEDA calculations for the complex were done at the level of MP2/PCM/cc-pVDZ. PIEs for fragments are shown in Figure 8.3. According to this plot, there are strong residue–ligand interactions involving residue fragments Gln-5, Pro-18, and Trp-6. The ligand interacts mainly with the two terminal sections of the protein, as can be seen in Figure 8.2.

Individual components describing the nature of interactions are shown in Figure 8.4. For Gln-5, a strong attraction is found in the CT + MIX, RC + DI, and

Table 8.1 Validation of PIEs and TIEs to other computational and experimental results.

Methods	System (PDB ID)	Correlation	Reference
MP2 vs experiment	OX ₂ orexin (4S0V)	TIE vs pEC ₅₀ ($R^2 = 0.872$).	[62]
MP2 vs experiment	OX ₂ orexin receptor (4S0V) β_2 adrenergic receptor (3SN6) κ opioid receptor (4DJH) P2Y ₁₂ receptor (4NTJ)	TIE vs pK _i ($R^2 = 0.748$), pEC ₅₀ ($R^2 = 0.729$), pK _e ($R^2 = 0.576$), and pIC ₅₀ ($R^2 = 0.763$). ^{a)}	[55]
DFTB vs experiment	β_2 adrenergic receptor (3SN6) κ opioid receptor (4DJH) P2Y ₁₂ receptor (4NTJ)	TIE vs pK _i ($R^2 = 0.783$), pEC ₅₀ ($R^2 = 0.662$), and pIC ₅₀ ($R^2 = 0.812$). ^{a)}	[63]
MP2 vs experiment	Cyclin-dependent kinase-2 inhibitor (4FKL, etc.)	TIE vs ΔG ($R^2 = 0.99$),	[64]
MP2 vs experiment	Estrogen receptor β (7XVY, etc.)	TIE vs ΔH ($R^2 = 0.870$)	[65]
DFTB vs MP2	β_2 adrenergic receptor (3SN6) κ opioid receptor (4DJH) P2Y ₁₂ receptor (4NTJ)	TIEs ($R^2 = 0.943$, $R^2 = 0.913$, and $R^2 = 0.959$). ^{a)}	[63]
DFTB vs MP2	Trp-cage (1L2Y)	PIEs ($R^2 = 0.990$ and 0.988). ^{b)}	[66]
HF-3c vs MP2	Trp-cage (1L2Y)	PIEs ($R^2 = 0.999$ and 0.983). ^{b)}	[51]

a) Each of these R^2 values is for multiple ligands bound to a single protein.

b) Each of these R^2 values is for multiple residues bound to a single ligand.

ES + solv terms, compensated partially by repulsion in the EX term, as is usually the case. RC + DI has substantial contributions, mainly because of the phenyl ring, and the hydrophobic RC + DI term is roughly on par with the screened electrostatics ES + solv, although the latter is somewhat larger on average (the ligand has a charge of -1).

8.4 Partition Analysis (PA)

Pair interaction energies between fragments are very useful, but there are three problems with them: (1) fragments differ (albeit slightly) from conventional units (residues or nucleotides), (2) fragment pairs with a covalent boundary between the two fragments have a large artificial interaction, and (3) it is not feasible to get functional group contributions. All of these problems are solved by the use of

Table 8.2 Representative applications of PIEs.

Level	Task	References
ROMP2	Prediction of the hydrogen abstraction site for radical damage in lipids.	[70]
MP2	Hit-to-lead phase in drug design using SAR-by-FMO (CDK2 receptor).	[39, 54]
MP2	Signatures of interaction types in GPCR proteins.	[55, 57]
MP2	Lead optimization for kinase inhibition.	[19]
MP2	Discovery of a preclinical candidate for the treatment of type 2 diabetes.	[71]
MP2	Formation of an amorphous formulation in pharmaceuticals.	[72]
MP2	Protein-residue topological network models of protein–protein complexes.	[73]
MP2	Dissipative particle dynamics using PIE-based parametrization.	[74]
MP2	Hot spot identification in protein–protein interfaces in relation to site-directed mutagenesis.	[75]
MP2	Protonation and pK_a in polynucleotides	[76]
RI-MP2 and DFTB	Virtual screening for metalloprotease.	[77]
RI-MP2	Virtual screening for SARS-CoV-2	[78]
RI-MP2	Combination of a molecular interaction field (MIF) with PIEs for ligand docking.	[79]
DFTB	Machine learning for predicting binding affinities to the SARS-CoV-2 spike glycoprotein.	[61]
DFTB	Ensemble docking drug discovery pipeline for COVID-19.	[80]

segments in the partition analysis (PA) [82] for electronic energies and the partition analysis of vibrational energies (PAVEs) [83].

What are segments? Segments are sets of atoms, like fragments. The most fundamental difference is that QM calculations are done for fragments, but not for segments. The QM results of fragments are post-processed (repartitioned) into segments, somewhat analogously to computing atomic charges for a converged wave function. Capital (I, J) and small (i, j) indices are used for fragments and segments, respectively.

There is no limit to the definition of segments. They can be conventional residues, functional groups, or even individual atoms. There is no accuracy loss due to the use of small segments because the partitioning of properties is exact.

FMO fragments are used for a fast computation of QM properties, which are at the end partitioned into segments in PA. It can be done for the electronic energy in DFTB or the vibrational energy in any QM method.

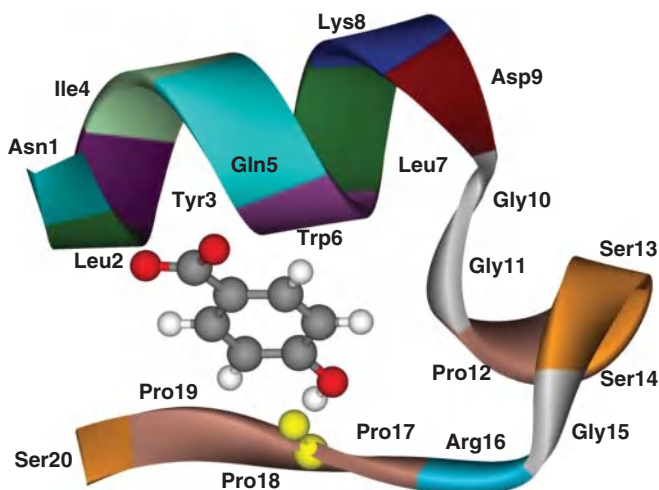


Figure 8.2 Protein–ligand complex for Trp-cage (1L2Y). The two yellow atoms are the carbonyl group of Pro-17 assigned to the Pro-18 fragment.

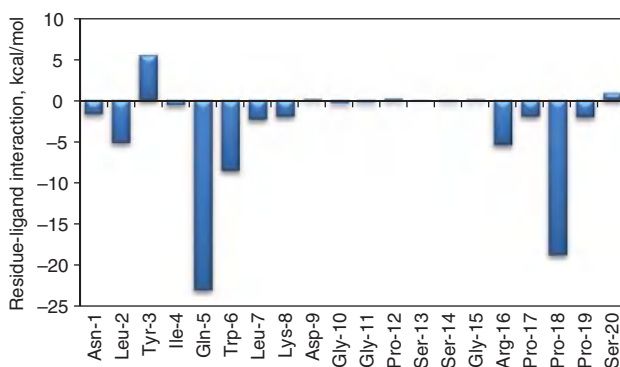


Figure 8.3 Total residue–ligand interactions ΔE_l (PIEDA) in the protein–ligand complex at the level of MP2/PCM/cc-pVDZ.

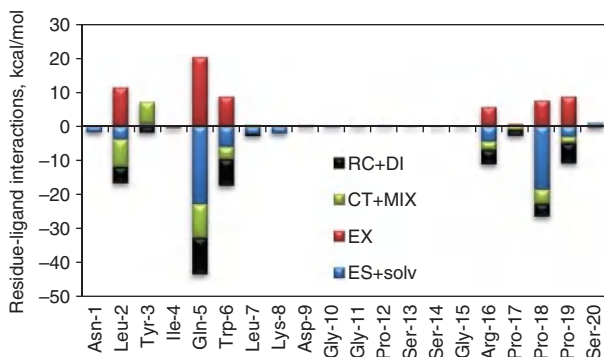


Figure 8.4 Components A of residue–ligand interactions ΔE_l^A (PIEDA) in the protein–ligand complex at the level of MP2/PCM/cc-pVDZ ($A = RC + DI$, $CT + MIX$, EX , and $ES + solv$).

Table 8.3 Fragments vs segments as units for analyses.

Property	Fragments	Segments
QM method	Any	DFTB(elec), any (vib) ^{a)}
Analyses	PIEDA	PA, PAVE
Bonds between units	Single	Any
Functional groups as units	No ^{b)}	Yes
Residues/nucleotides	Shifted	Conventional ^{c)}
Unit charges	Integer	Fractional
Invariance ^{d)}	No	Yes
Use without FMO	No	Yes ^{e)}

a) For electronic (elec) or vibrational (vib) energies.

b) Technically possible but not recommended due to artificial results.

c) Shifted residues can also be used, with segments defined identically to fragments.

d) Invariance of the total energies with respect to a different division into units.

e) Segments can be used in unfragmented calculations (technically accomplished as FMO with 1 fragment).

On a fragment boundary, one atom (BDA) is shared between two fragments [40], whereas no atom is shared between segments. Bonds of any order can be on a segment boundary, and PIEs for segment pairs with a covalent bond are in the same order as non-covalent PIEs.

Another major difference is that for segments, charge transfer is treated at a higher order than for fragments. It is accomplished by using FMO_n atomic charges for defining the charges of segments, which are in general fractional. A summary of the differences is shown in Table 8.3.

8.4.1 Formulation of PA

The energy of M segments in PA [82] is (compare to Eq. (8.1)),

$$E = \sum_{i=1}^M E'_i + \sum_{i>j} \Delta E_{ij} \quad (8.12)$$

where E'_i is the internal energy of segment i and the PIE for two segments is

$$\Delta E_{ij} = \Delta E_{ij}^{ES} + \Delta E_{ij}^{DI} + \Delta E_{ij}^{solv} \quad (8.13)$$

PA does not have trimer terms, even if PA is conducted for post-processing FMO3 results. It is because of the nature of DFTB interaction terms, which involve at most two particles (two atoms).

There are three components, electrostatic (ES), dispersion (DI) and solvent screening (solv). They have the same meaning (but, in general, different values) as fragments. 0-order (i.e. EX+RC) and CT-related terms are absent in Eq. (8.13), and these

interactions are incorporated into monomer values E'_i . As a result, ΔE_{ij} in PA has no repulsive term corresponding to EX in PIEDA.

PA may be used with DFTB only, possibly combined with PCM, SMD, or PBC. A PA calculation can use a PDB file, reading all important biochemical information, such as atomic and residue names. Conventional residues can be used exactly as defined in a PDB. Side chains can be automatically split from amino acid residues, whereas bases can be split from nucleotides. This produces two segments for each residue or nucleotide. Functional groups can be treated as separate segments by adjusting the residue ID in the PDB, which is used to index atoms in segments [40].

8.4.2 Applications and an Example of PA

PA/PBC was applied [34] to calculate the interactions between solvent molecules and proteins in protein crystals, where some superbinding water molecules were identified that formed 3 hydrogen bonds and 2 charge–dipole interactions with the protein. PA/PBC was used [33] to analyze a self-assembled monolayer of organic molecules on an ice surface, where the importance of dispersion for guest–guest binding in competition with guest–surface binding was shown. PA/PBC was applied [35] to analyze the catalytic activity of faujasite zeolite, where the importance of the charge delocalization and interactions of multiple zeolite segments in the transition state stabilization was shown.

As a demonstration of PA, it is applied to the protein–ligand complex used in Section 7.3.3, at the level of FMO2-DFTB3/3ob/D3(BJ)/PCM. In the Trp-cage protein, each conventional residue out of 20 is defined as a segment. The ligand is split into 3 segments: phenyl (Ph), hydroxyl (OH), and carboxylate (COO^-).

The interaction energies of residues with the three ligand segments are shown in Figure 8.5 (here, conventional residues are used for analysis, so that their names have no dash). The phenyl interacts with Gln5, Trp6, Pro17, and Pro19. The hydroxyl interacts with Trp6, Arg16, and Pro17. The carboxylate interacts with Ile5 and Leu2. Such quantitative information can be useful for FBDD, where segment efficiency $\Delta \bar{E}_i$ can be defined similarly to Eq. (8.8).

Essentially no repulsion is observed in PA (Figure 8.5), whereas some is seen in PIEDA (Figure 8.3). It is attributed to (a) the consistent use of the same method (DFTB) in the structure optimization and analysis (PA in Figure 8.5) and (b) the assignment of EX repulsion to monomer terms in PA without a corresponding PIE contribution.

In Figure 8.3, a strong interaction of the ligand is observed for Pro-18 fragment. However, according to Figure 8.5, Pro17 segment has a strong interaction. A close inspection reveals that it is because of the carbonyl of Pro17 that is assigned to Pro-18 in the fragment-based analysis. This carbonyl is shown in Figure 8.2.

Because solvent screening is taken into account, the binding of the anionic carboxylate segment is on par with the binding of neutral functional groups. Segment charges are fractional; for phenyl, carboxylate, and hydroxyl, the computed segment charges in the complex are 0.026, -0.868 , and -0.131 , respectively (a.u.). The sum of

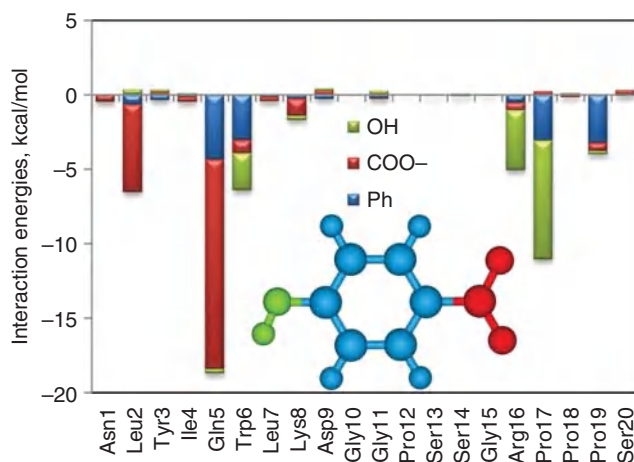


Figure 8.5 Interaction energies ΔE_{ij} (PA) of functional groups of the ligand with residues in the protein-ligand complex at the level of DFTB3/PCM.

them, -0.973 , reflects the protein-ligand charge transfer (the formal ligand charge is -1).

PA offers a unique feature of evaluating the energy of individual bonds between functional groups. In the application of PA to a protein-DNA complex [84], nucleotides in the DNA were divided into phosphates, ring remainders, and tiny segments made of 2–3 atoms, which constitute functional groups (such as amides or carbonyls). By doing this, the energies of individual hydrogen bonds and $\text{CH}\cdots\text{O}$ interactions were defined in PA, explaining the relative strength of the nucleotide-nucleotide binding in the natural pairs $\text{C}\cdots\text{G}$ and $\text{A}\cdots\text{T}$, as well as in a mutant pair $\text{G}\cdots\text{T}$.

8.5 Partition Analysis of Vibrational Energy (PAVE)

PAVE [83] is used for computing the vibrational enthalpy, entropy, and free energy of individual segments. It is accomplished by computing a Hessian matrix of second derivatives of E with respect to nuclear coordinates either analytically or semi-numerically (by a numerical differentiation of analytic gradients), evaluating partition functions, and hence thermodynamical quantities. PAVE is computed at a given temperature for a single minimum (it does not take into account conformational entropy).

PAVE is done for a manually chosen selection of active atoms by constructing a partial Hessian for it. PAVE can be used for PBC (semi-numerical Hessians only). The decomposition of vibrational energies is done based on the localization of the eigenvectors of the Hessian (the squared coefficients summed for each atom are taken as the weight of the contribution of that atom, and atomic values are summed into segments).

8.5.1 Formulation of PAVE

The vibrational zero-point energy (ZPE) is decomposed into segment contributions as

$$E^{ZPE} = \sum_{i=1}^M E_i^{ZPE} \quad (8.14)$$

The vibrational (vib) enthalpy is

$$H^{vib} = E^{ZPE} + H^T \quad (8.15)$$

where the temperature (T) dependent enthalpy contribution is

$$H^T = \sum_{i=1}^M H_i^T \quad (8.16)$$

The vibrational Gibbs free energy is

$$G^{vib} = H^{vib} - TS^{vib} \quad (8.17)$$

where the entropy is

$$S^{vib} = \sum_{i=1}^M S_i^{vib} \quad (8.18)$$

By combining terms, the free energy is decomposed as

$$G^{vib} = \sum_{i=1}^M G_i^{vib} = \sum_{i=1}^M (E_i^{ZPE} + H_i^T - TS_i^{vib}) \quad (8.19)$$

PAVE has no pairwise contributions. Instead, only monomer vibrational properties are defined (e.g. G_i^{vib}), which correspond to electronic energies E'_i in PA. Segment contributions to a binding or activation energy can be computed by subtracting appropriate values of G_i^{vib} (e.g. for transition states and reactants). PAVE results can be combined with PA, for a composite electronic+vibrational analysis.

Like PA, PAVE can be used for unfragmented calculations (technically done as FMO with 1 fragment), although typically PAVE is used with FMO. PAVE is a post-processing of Hessian results, like PA is a post-processing of DFTB energies.

PAVE is sensitive to the accuracy of the Hessian, especially for H^T and S contributions, which are determined by low-frequency vibrations. The most accurate gradient (Hessian) should be used for a minimum located with a tight threshold (10^{-5} Eh/bohr or smaller) in geometry optimizations.

8.5.2 Applications of PAVE

PAVE was applied to evaluate the enthalpy and entropy of the guest binding on an ice surface, with a good agreement between computed and experimental enthalpies but a less favorable comparison of entropies [33].

PAVE was applied [35] to study guest binding to two types of zeolite crystals and the full production cycle of *p*-xylene catalyzed by faujasite zeolite. In both of these studies (adsorption and solid-state catalysis), the PAVE engine was used to get the total values of enthalpy H , entropy S , and free energy G , but individual segment contributions were not discussed. A demonstrative example of an application of PAVE is given below.

8.6 Subsystem Analysis (SA)

The analyses described above, PIEDA and PA, can be conducted for a single structure, for example, for a protein-ligand complex. Such a calculation can reveal insight into the complex stability. Interaction and binding energies are inherently different, and subsystem analysis (SA) [85] can be used to analyze the binding energies, which are more relevant to many biochemical applications.

8.6.1 Formulation of SA

For an FMO-based analysis of binding, the starting point is the equation that describes the process (a complex formation or a chemical reaction). Taking as an example a protein (A) – ligand (B) complex formation,



its binding (bind) energy is simply

$$\Delta E^{bind} = E^{AB} - E^A - E^B \quad (8.21)$$

using the FMO energies E in Eq. (8.1) computed for A , B , and AB . One can take electronic energies only, or add vibrational contributions in Eq. (8.19). It is possible to consider the effects of a structure's deformation (optimizing each system separately) or, as an approximation, neglect the deformation effects, optimizing only the complex and using its geometry for computing A and B separately.

The binding energy can be studied with SA, which requires at least 3 separate calculations (of AB , A , and B), and the arithmetic burden of subtracting numbers lies with the user. SA can be performed with either fragments or segments.

The energy decomposition in SA can be written for fragments as

$$\Delta E^{bind} = \sum_{I \in A} \Delta E_I^{part} + \sum_{I \in B} \Delta E_I^{part} + \sum_{I \in A} \sum_{J \in B} \Delta E_{IJ} \quad (8.22)$$

where ΔE_I^{part} is the difference in the partial (part) energies E_I^{part} of fragment I in the complex and isolated states; ΔE_I^{part} describes deformation, polarization, and desolvation (it is possible to decouple these three contributions and define them separately [40]). For residues I and J , the partial energy of I in the protein is

$$E_I^{part} = E'_I + \frac{1}{2} \sum_{J \neq I} \Delta E_{IJ} \quad (8.23)$$

The reason for using partial energies is to reduce the complexity of data. For example, when a ligand binds to a protein, it can affect residue–residue interactions (the ΔE_{IJ} term in Eq. (8.23)) via polarization. Usually, these effects are small, and they can be conveniently compressed into more manageable partial energies of residues E_I^{part} in the protein, so that the decomposition in Eq. (8.22) is formulated using differential partial energies ΔE_I^{part} of residues ($I \in A$) and ligand ($I \in B$), and residue–ligand interactions ΔE_{IJ} . The number of terms in Eq. (8.22) is linear with respect to the number of residues N^{res} (if the ligand is not fragmented, there are $2N^{res} + 1$ terms). In contrast, the number of symmetric residue–residue interactions ΔE_{IJ} is quadratic $\sim N^{res2}/2$.

The difference in residue–residue interactions ΔE_{IJ} (differential $\Delta\Delta E_{IJ}$) contributes to the binding. The values of $\Delta\Delta E_{IJ}$ may be pertinent to rationalize allosteric regulation in proteins. Heat maps of ΔE_{IJ} in the complex and $\Delta\Delta E_{IJ}$ are shown in Figure 8.6 (fragment pairs connected by a covalent bond are excluded). Differential interactions reflect two effects: deformation (the structure is separately optimized for the bound and isolated states) and polarization of residue–residue interactions by the ligand. In other words, they explain how the ligand changes the protein. For absolute values in the complex, Lys-8 and Gln-5 are the happiest pair, and Pro-17 and Pro-12 are the unhappiest couple.

An isolated ligand (protein) is fully immersed in the solution, but in a complex, some solute–solvent interaction energy is lost (the desolvation penalty). Charged ligands can have a large attractive interaction energy ΔE_{IJ} with the protein and a large repulsive desolvation penalty in ΔE_I . The values of TIE (the last term in Eq. (8.22)) can be a large overestimate of the binding energy.

If there is just one ligand fragment J , then it is possible to simplify Eq. (8.22) as

$$\Delta E^{bind} = \sum_{I \in A, B} \Delta E_I^{bind} \quad (8.24)$$

where the binding energies ΔE_I^{bind} of fragments I include all relevant effects (deformation, polarization, desolvation, and interaction). For residue I and ligand J , $\Delta E_I^{bind} = \Delta E_I^{part} + \Delta E_{IJ}$ and $\Delta E_J^{bind} = \Delta E_J^{part}$. The values of ΔE_I^{bind} are better descriptors of binding than PIEs ΔE_{IJ} .

SA can be performed for segments. By combining PA, PAVE, and SA, the free energy of binding can be decomposed as [83]

$$\Delta G^{bind} = \Delta E^{bind} + \Delta G^{vib} = \sum_{i \in A, B} \Delta G_i^{bind} \quad (8.25)$$

where the free binding energy contribution of residue i is

$$\Delta G_i^{bind} = \Delta E_i^{part} + \Delta G_i^{vib} + \sum_{j \in \text{ligand}} \Delta E_{ij} \quad (8.26)$$

and the contribution of functional group i of ligand is

$$\Delta G_i^{bind} = \Delta E_i^{part} + \Delta G_i^{vib} \quad (8.27)$$

where ΔG_i^{vib} is obtained by subtracting G_i^{vib} for the complex and isolated species.

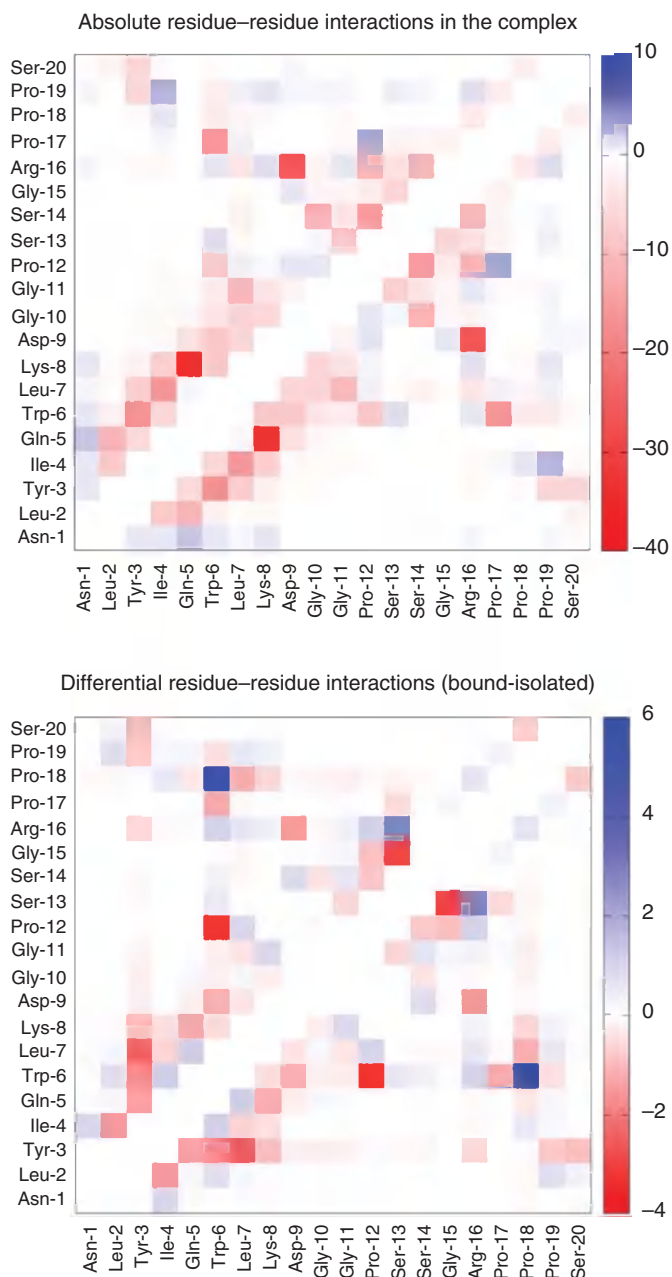


Figure 8.6 Residue–residue pair interactions ΔE_{ij} in the complex (top) and differential values of $\Delta\Delta E_{ij}$ (bound minus isolated protein, bottom) for unconnected dimers in the complex of Trp-cage with its ligand (MP2/PCM/cc-pVDZ), in kcal/mol.

8.6.2 Examples of SA and PAVE

The same system as in Section 7.4.3 is analyzed using a combination of SA, PA, and PAVE. The protein is divided into 20 residue segments, and the ligand into 3 functional group segments. Two subsystems are defined in SA, the protein and the ligand. The structures of AB , A , and B are separately optimized, so that deformation energies are implicitly included in each term. PAVE was conducted at 298 K. To obtain the vibrational contribution to the carboxylate binding, the atoms Leu2, Gln5, and COO^- are defined to be active (for them a partial Hessian is computed). The results are shown in Figure 8.7.

The binding energy ΔE^{bind} is obtained by subtracting three total energies in PA (Eq. (8.12)). The result is -24.7 kcal/mol. In contrast, TIE (the sum of all residue–ligand PIEs, the last term in Eq. (8.26)) is -54.3 kcal/mol, so that TIE is an overestimate of binding mainly due to desolvation effects. ΔG^{vib} is -1.5 kcal/mol. ΔG^{bind} is -26.2 kcal/mol.

The polarization, deformation, and desolvation are described by the partial energy, ΔE_i^{part} . Two ligand segments have a substantial value: phenyl, because of the desolvation loss of the dispersion interaction with solvent (the non-es term in Eq. (8.3)), and carboxylate, because some of its electrostatic interaction with solvent (the es term in Eq. (8.3)) is lost in the complex. The cavitation energy (included in E_i) also makes a contribution to binding, describing the entropy loss of the solvent. After all contributions are added, the following final binding analysis is obtained (Figure 8.8).

Ligand segments (Ph and COO^-) lose quite a bit of energy due to complexation (ΔG_i^{bind} is repulsive). On the other hand, there is a gain due to the protein–ligand binding, which overweighs the loss. Thus, in analyzing protein–ligand binding, the desolvation penalty (the main difference between interaction and binding energies) should be considered.

SA has been applied to biscarbene–Gold(I)/DNA G-quadruplex complex [86], stabilized in part by π -cation interactions involving by Au(I).

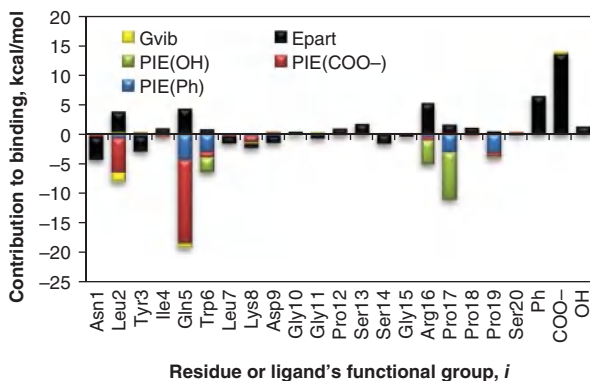
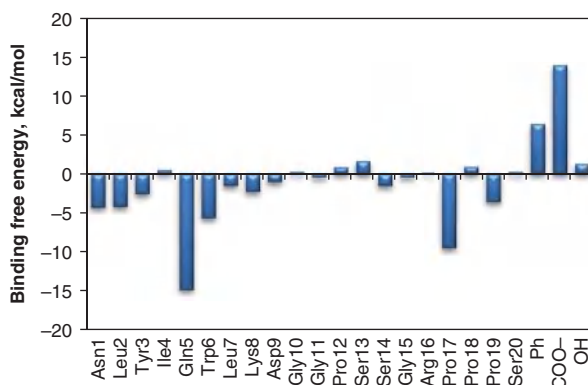


Figure 8.7 Contributions of partial energies (E_{part} , ΔE_i^{part}), vibrational free energies (G_{vib} , ΔG_i^{vib}), and pair interaction energies of residue i with functional group j of the ligand ($PIE(j)$, ΔE_{ij}) to the protein–ligand binding energy (DFTB3/PCM).

Figure 8.8 Free energies of binding ΔG_i^{bind} at 298 K for residues and functional groups in the ligand (DFTB3/PCM).



8.7 Fluctuation Analysis (FA)

Temperature and conformational flexibility are important for soft materials, such as proteins or sugars. In the fluctuation analysis (FA) [66], a conformational averaging in FMO/MD is combined with the many-body expansion in Eq. (8.1).

In their most common form, energy fluctuations are measured relative to a reference value E_0 , typically chosen to be the minimal energy in MD. FMO2/FA can be written as

$$\langle E \rangle = E_0 + \langle E - E_0 \rangle = \sum_{I=1}^N E_I^0 + \sum_{I>J}^N \Delta E_{IJ}^0 + \sum_{I=1}^N \langle \Delta E_I \rangle + \sum_{I>J}^N \langle \Delta \Delta E_{IJ} \rangle \quad (8.28)$$

where brackets indicate averaging over an MD trajectory. Thus, the total QM energy (the internal energy in the thermodynamical sense) is a sum of the reference energy and fluctuations from it, decomposed into monomer and dimer values as

$$\langle \Delta E_I \rangle = \langle E_I \rangle - E_I^0 \quad (8.29)$$

$$\langle \Delta \Delta E_{IJ} \rangle = \langle \Delta E_{IJ} \rangle - \Delta E_{IJ}^0 \quad (8.30)$$

The total kinetic (kin) energy, normalized in NVT or NPT simulations to a given temperature T , can be decomposed into fragment values E_I^{kin} . Using the number of atoms N_I^{at} , the effective temperature of each fragment I is

$$T_I = \frac{2\langle E_I^{kin} \rangle}{3N_I^{at}k_B} \quad (8.31)$$

where k_B is the Boltzmann's constant. Some fragments are hotter than others, although not necessarily hot spots.

FA was applied to a protein-ligand binding complex [85], where the weakening of interactions at room temperature was attributed to the kinetic pressure, forcing the molecular structure to stray away from the minimum due to thermal fluctuations. Dynamical aspects of protein-ligand binding can be very important for designing new drugs, in particular because they affect the drug residence time [87]. Averaging over multiple structures helps to increase the correlation of QM predictions to experimental binding energies [64].

8.8 Free Energy Decomposition Analysis (FEDA)

In the free energy decomposition analysis (FEDA) [32], FA is combined with constrained MD (umbrella sampling MD), used to obtain the potential of mean force (PMF). PMF is taken to be equal to the free energy.

So far, FEDA has been applied to chemical reactions, for which a reaction coordinate ζ can be designed a priori. In FMO, all reactants are usually assigned to one fragment. By doing a series of FMO/MD simulations for a set of values of ζ_0 , adding the constraining potential

$$U(\zeta) = \frac{k}{2}(\zeta - \zeta_0)^2 \quad (8.32)$$

a PMF $F(\zeta)$ is obtained. By plotting it, one can identify three values of the reaction coordinate, describing reactants ζ^A , transition state ζ^B , and products ζ^C .

To analyze the reaction barrier, the points ζ^A and ζ^B are important. Two more MD simulations are performed with ζ_0 equal to ζ^A and ζ^B , with a very large value of k , strongly constraining the system to be near the desirable points. The free energy of the reaction barrier is then decomposed as

$$\Delta F = F(\zeta^B) - F(\zeta^A) = \Delta E - T\Delta S \quad (8.33)$$

The QM energy ΔE is obtained using FA,

$$\Delta E = \langle E(\zeta^B) \rangle - \langle E(\zeta^A) \rangle = \sum_{I=1}^N \Delta E_I + \sum_{I>J}^N \Delta \Delta E_{IJ} \quad (8.34)$$

where monomer terms describe the polarization and deformation of fragments

$$\Delta E_I = \langle E_I(\zeta^B) \rangle - \langle E_I(\zeta^A) \rangle \quad (8.35)$$

and differential PIEs describe the change in PIEs due to the transition state formation,

$$\Delta \Delta E_{IJ} = \langle \Delta E_{IJ}(\zeta^B) \rangle - \langle \Delta E_{IJ}(\zeta^A) \rangle \quad (8.36)$$

FEDA is performed by first computing ΔF (from MD trajectories for a chosen set of ζ_0) using PMF, then ΔE is calculated from two more MD trajectories (for ζ^A and ζ^B), and, finally, the entropy ΔS is computed from Eq. (8.33).

FEDA was applied [32] to study S_N2 reactions in explicit solvent using FMO/PBC/MD, where the mechanism of the intrasolute and solute-solvent charge transfer during the reaction was elucidated.

8.9 Other Analyses of Chemical Reactions

Prior to the development of FEDA, there were other applications of FMO/MD to find reaction pathways for organic reactions in solution [88, 89].

FMO can be used to locate a transition state by computing the Hessian and using standard engines for a saddle point search. A reaction path can be mapped using

the intrinsic reaction coordinate (IRC) combined with FMO [90]. For enzymes, a full structure relaxation may be time-consuming, and a feasible approach is to use the frozen domain (FD) formulation of FMO [20] to optimize only a part of the system while treating the whole enzyme quantum-mechanically. There are several examples [90–92] of mapping a reaction path for enzymatic catalysis in this way using FMO for systems up to about 9000 atoms.

Alternatively, a reaction path can be mapped with QM/MM [93]. For some representative structures, PIEs can be computed, and the roles of different residues can be identified [94–96].

8.10 Conclusions

The FMO method is feasible for applications to biochemical systems on a realistic time scale. Both traditional methods like DFT and MP2, and parametrized approaches like DFTB can be used with FMO. The latter is suitable for full geometry optimizations and MD simulations, with a conformational sampling (although rather short by modern MM standards, on the order of 1 ns for fully QM MD simulations).

For a balanced description of binding, the contributions of desolvation, polarization, and deformation should be included, whereas interaction energies overestimate binding.

Free binding energies in PA + PAVE can be decomposed into segment values, although they include only electronic and vibrational contributions for a single minimum. There are important conformational entropic effects not considered in such analysis.

An important role of theoretical chemistry is to explain chemical phenomena and enhance our understanding of them, with the goal of being able to improve the materials, such as drugs or enzymes, by designing them *in silico*. FMO delivers very useful quantitative data, in particular pair interaction energies, which can pinpoint hot spots in proteins and other macromolecules.

A molecular design *in silico* is by no means easy, because of the complexity of molecular objects, including conformational sampling. However, pathways to solve typical problems in biochemistry using QM methods are set out.

Machine learning and artificial intelligence can be useful in building predictive models based on some training data, but it remains to be seen if such models can explain the physics of a given problem rather than simply make a black-box prediction (though useful it may be).

References

- 1 Phipps, M.J.S., Fox, T., Tautermann, C.S., and Skylaris, C.-K. (2015). Energy decomposition analysis approaches and their evaluation on prototypical protein–drug interaction patterns. *Chem. Soc. Rev.* 44: 3177–3211.

- 2 Merz, K.M. (2014). Using quantum mechanical approaches to study biological systems. *Acc. Chem. Res.* 47: 2804–2811.
- 3 Ryde, U. and Söderhjelm, P. (2016). Ligand-binding affinity estimates supported by quantum-mechanical methods. *Chem. Rev.* 116: 5520–5566.
- 4 Gordon, M.S., Pruitt, S.R., Fedorov, D.G., and Slipchenko, L.V. (2012). Fragmentation methods: a route to accurate calculations on large systems. *Chem. Rev.* 112: 632–672.
- 5 Raghavachari, K. and Saha, A. (2015). Accurate composite and fragment-based quantum chemical models for large molecules. *Chem. Rev.* 115: 5643–5677.
- 6 Erlanson, D.A., Fesik, S.W., Hubbard, R.E. et al. (2016). Twenty years on: the impact of fragments on drug discovery. *Nat. Rev. Drug Disc.* 15: 605–619.
- 7 Mironov, V., Shchugoreva, I.A., Artyushenko, P.V. et al. (2022). Structure- and interaction-based design of anti-SARS-CoV-2 aptamers. *Chem. Eur. J.* 28: e202104481.
- 8 Fedorov, D.G. (2022). Parametrized quantum-mechanical approaches combined with the fragment molecular orbital method. *J. Phys. Chem.* in press.
- 9 Christensen, A.S., Kubar, T., Cui, Q., and Elstner, M. (2016). Semiempirical quantum mechanical methods for noncovalent interactions for chemical and biochemical applications. *Chem. Rev.* 116: 5301–5337.
- 10 Kitaura, K. and Morokuma, K. (1976). A new energy decomposition scheme for molecular interactions within the Hartree-Fock approximation. *Int. J. Quant. Chem.* 10: 325–340.
- 11 Chen, W. and Gordon, M.S. (1996). Energy decomposition analyses for many-body interaction and applications to water complexes. *J. Phys. Chem.* 100: 14316–14328.
- 12 Kitaura, K., Ikee, E., Asada, T. et al. (1999). Fragment molecular orbital method: an approximate computational method for large molecules. *Chem. Phys. Lett.* 313: 701–706.
- 13 Fedorov, D.G. (2017). The fragment molecular orbital method: theoretical development, implementation in GAMESS, and applications. *WIREs: Comput. Mol. Sc.* 7: e1322.
- 14 Fukuzawa, K. and Tanaka, S. (2022). Fragment molecular orbital calculations for biomolecules. *Curr. Opin. Struct. Biol.* 72: 127–134.
- 15 Mochizuki, Y., Tanaka, S., and Fukuzawa, K. (ed.) (2021). *Recent Advances of the Fragment Molecular Orbital Method*. Singapore: Springer.
- 16 Suenaga, M. (2008). Development of GUI for GAMESS/FMO calculation. *J. Comput. Chem. Jap.* 7: 33–54. (in Japanese).
- 17 Sawada, T., Fedorov, D.G., and Kitaura, K. (2010). Binding of influenza a virus hemagglutinin to the sialoside receptor is not controlled by the homotropic allosteric effect. *J. Phys. Chem. B* 114: 15700–15705.
- 18 Nakamura, S., Akaki, T., Nishiwaki, K. et al. (2023). System truncation accelerates binding affinity calculations with the fragment molecular orbital method: a benchmark study. *J. Comput. Chem.* 44:824–831.

- 19 Heifetz, A., Trani, G., Aldeghi, M. et al. (2016). Fragment molecular orbital method applied to lead optimization of novel interleukin-2 inducible T-cell kinase (ITK) inhibitors. *J. Med. Chem.* 59: 4352–4363.
- 20 Fedorov, D.G., Alexeev, Y., and Kitaura, K. (2011). Geometry optimization of the active site of a large system with the fragment molecular orbital method. *J. Phys. Chem. Lett.* 2: 282–288.
- 21 Fedorov, D.G. (2022). Polarization energies in the fragment molecular orbital method. *J. Comput. Chem.* 43: 1094–1103.
- 22 Nakata, H. and Fedorov, D.G. (2020). Geometry optimization, transition state search, and reaction path mapping accomplished with the fragment molecular orbital method. In: *Quantum Mechanics in Drug Discovery*, A. Heifetz (Ed.), *Methods in Molecular Biology*, vol. Vol. 2114, 87–104. New York: Springer.
- 23 Komeiji, Y., Mochizuki, Y., Nakano, T., and Fedorov, D.G. (2009). Fragment molecular orbital-based molecular dynamics (FMO-MD), a quantum simulation tool for large molecular systems. *J. Mol. Str. (THEOCHEM)* 898: 2–7.
- 24 Nakata, H. and Fedorov, D.G. (2016). Efficient geometry optimization of large molecular systems in solution using the fragment molecular orbital method. *J. Phys. Chem. A* 120: 9794–9804.
- 25 Nishimoto, Y. and Fedorov, D.G. (2016). The fragment molecular orbital method combined with density-functional tight-binding and the polarizable continuum model. *Phys. Chem. Chem. Phys.* 18: 22047–22061.
- 26 Nishimoto, Y., Nakata, H., Fedorov, D.G., and Irle, S. (2015). Large-scale quantum-mechanical molecular dynamics simulations using density-functional tight-binding combined with the fragment molecular orbital method. *J. Phys. Chem. Lett.* 6: 5034–5039.
- 27 Nakata, H., Fedorov, D.G., Yokojima, S. et al. (2014). Simulations of Raman spectra using the fragment molecular orbital method. *J. Chem. Theory Comput.* 10: 3689–3698.
- 28 Nakata, H. and Fedorov, D.G. (2020). Analytic first and second derivatives of the energy in the fragment molecular orbital method combined with molecular mechanics. *Int. J. Quantum Chem.* 120: e26414.
- 29 Fedorov, D.G., Kitaura, K., Li, H. et al. (2006). The polarizable continuum model (PCM) interfaced with the fragment molecular orbital method (FMO). *J. Comput. Chem.* 27: 976–985.
- 30 Fedorov, D.G. (2018). Analysis of solute-solvent interactions using the solvation model density combined with the fragment molecular orbital method. *Chem. Phys. Lett.* 702: 111–116.
- 31 Nishimoto, Y. and Fedorov, D.G. (2021). The fragment molecular orbital method combined with density-functional tight-binding and periodic boundary conditions. *J. Chem. Phys.* 154: 111102.
- 32 Fedorov, D.G. and Nakamura, T. (2022). Free energy decomposition analysis based on the fragment molecular orbital method. *J. Phys. Chem. Lett.* 13: 1596–1601.

- 33 Nakamura, T., Yokaichiya, T., and Fedorov, D.G. (2022). Analysis of guest adsorption on crystal surfaces based on the fragment molecular orbital method. *J. Phys. Chem. A* 126: 957–969.
- 34 Nakamura, T., Yokaichiya, T., and Fedorov, D.G. (2021). Quantum-mechanical structure optimization of protein crystals and analysis of interactions in periodic systems. *J. Phys. Chem. Lett.* 12: 8757–8762.
- 35 Nakamura, T. and Fedorov, D.G. (2022). The catalytic activity and adsorption in faujasite and ZSM-5 zeolites: the role of differential stabilization and charge delocalization. *Phys. Chem. Chem. Phys.* 24: 7739–7747.
- 36 Nishimoto, Y., Fedorov, D.G., and Irle, S. (2014). Density-functional tight-binding combined with the fragment molecular orbital method. *J. Chem. Theory Comput.* 10: 4801–4812.
- 37 Nishimoto, Y. and Fedorov, D.G. (2018). Adaptive frozen orbital treatment for the fragment molecular orbital method combined with density-functional tight-binding. *J. Chem. Phys.* 148: 064115.
- 38 Fedorov, D.G. and Kitaura, K. (2004). The importance of three-body terms in the fragment molecular orbital method. *J. Chem. Phys.* 120: 6832–6840.
- 39 Alexeev, Y., Mazanetz, M.P., Ichihara, O., and Fedorov, D.G. (2012). GAMESS as a free quantum-mechanical platform for drug research. *Curr. Top. Med. Chem.* 12: 2013–2033.
- 40 Fedorov, D.G. (2023). *Complete Guide to the Fragment Molecular Orbital Method in GAMESS*. Singapore: World Scientific.
- 41 Barca, G.M.J., Bertoni, C., Carrington, L. et al. (2020). Recent developments in the general atomic and molecular electronic structure system. *J. Chem. Phys.* 152: 154102.
- 42 Fedorov, D.G., Brekhov, A., Mironov, V., and Alexeev, Y. (2019). Molecular electrostatic potential and electron density of large systems in solution computed with the fragment molecular orbital method. *J. Phys. Chem. A* 123: 6281–6290.
- 43 Ozono, H., Mimoto, K., and Ishikawa, T. (2022). Quantification and neutralization of the interfacial electrostatic potential and visualization of the dispersion interaction in visualization of the interfacial electrostatic complementarity. *J. Phys. Chem. B* 126: 8415–8426.
- 44 Fedorov, D.G. and Kitaura, K. (2007). Pair interaction energy decomposition analysis. *J. Comput. Chem.* 28: 222–237.
- 45 Fedorov, D.G. and Kitaura, K. (2012). Energy decomposition analysis in solution based on the fragment molecular orbital method. *J. Phys. Chem. A* 116: 704–719.
- 46 Green, M.C., Fedorov, D.G., Kitaura, K. et al. (2013). Open-shell pair interaction energy decomposition analysis (PIEDA): formulation and application to the hydrogen abstraction in tripeptides. *J. Chem. Phys.* 138: 074111.
- 47 Fedorov, D.G. (2020). Three-body energy decomposition analysis based on the fragment molecular orbital method. *J. Phys. Chem. A* 124: 4956–4971.
- 48 Fedorov, D.G. (2019). Solvent screening in zwitterions analyzed with the fragment molecular orbital method. *J. Chem. Theory Comput.* 15: 5404–5416.

- 49 Thirman, J. and Head-Gordon, M. (2015). An energy decomposition analysis for second-order Møller-Plesset perturbation theory based on absolutely localized molecular orbitals. *J. Chem. Phys.* 143: 084124.
- 50 Fedorov, D.G. and Kitaura, K. (2014). Use of an auxiliary basis set to describe the polarization in the fragment molecular orbital method. *Chem. Phys. Lett.* 597: 99–105.
- 51 Fedorov, D.G., Kromann, J.C., and Jensen, J.H. (2018). Empirical corrections and pair interaction energies in the fragment molecular orbital method. *Chem. Phys. Lett.* 702: 111–116.
- 52 Nakano, T., Mochizuki, Y., Yamashita, K. et al. (2012). Development of the four-body corrected fragment molecular orbital (FMO4) method. *Chem. Phys. Lett.* 523: 128–133.
- 53 Watanabe, C., Fukuzawa, K., Okiyama, Y. et al. (2013). Three- and four-body corrected fragment molecular orbital calculations with a novel subdividing fragmentation method applicable to structure-based drug design. *J. Mol. Graphics Modell.* 41: 31–42.
- 54 Mazanetz, M.P., Chudyk, E., Fedorov, D.G., and Alexeev, Y. (2016). Applications of the fragment molecular orbital method to drug research. In: *Computer Aided Drug Discovery* (ed. W. Zhang), 217–255. York: Springer, New.
- 55 Heifetz, A., Chudyk, E.I., Gleave, L. et al. (2016). The fragment molecular orbital method reveals new insight into the chemical nature of GPCR-ligand interactions. *J. Chem. Inf. Model.* 56: 159–172.
- 56 Chudyk, E.I., Sarrat, L., Aldeghi, M. et al. (2018). Exploring GPCR-ligand interactions with the fragment molecular orbital (FMO) method. In: *Computational Methods for GPCR Drug Discovery* (ed. A. Heifetz), 179–195. New York: Humana Press.
- 57 Heifetz, A., Morao, I., Babu, M.M. et al. (2020). Characterizing interhelical interactions of G-protein coupled receptors with the fragment molecular orbital method. *J. Chem. Theory Comput.* 16: 2814–2824.
- 58 Mazanetz, M.P., Ichihara, O., Law, R.J., and Whittaker, M. (2011). Prediction of cyclin- dependent kinase 2 inhibitor potency using the fragment molecular orbital method. *J. Cheminf.* 3: 2.
- 59 Yoshida, T. and Hirono, S. (2019). A 3D-QSAR analysis of CDK2 inhibitors using FMO calculations and PLS regression. *Chem. Pharm. Bull.* 67: 546–555.
- 60 Tokutomi, S., Shimamura, K., Fukuzawa, K., and Tanaka, S. (2020). Machine learning prediction of inter-fragment interaction energies between ligand and amino-acid residues on the fragment molecular orbital calculations for Janus kinase-inhibitor complex. *Chem. Phys. Lett.* 757: 137883.
- 61 Lim, H., Jeon, H.-N., Lim, S. et al. (2022). Evaluation of protein descriptors in computer-aided rational protein engineering tasks and its application in property prediction in SARS-CoV-2 spike glycoprotein. *Comp. Str. Biotechn. J.* 20: 788–798.
- 62 Heifetz, A., Aldeghi, M., Chudyk, E.I. et al. (2016). Using the fragment molecular orbital method to investigate agonist-orexin-2 receptor interactions. *Biochem. Soc. Trans.* 44: 574–581.

- 63 Morao, I., Fedorov, D.G., Robinson, R. et al. (2017). Rapid and accurate assessment of GPCR-ligand interactions using the fragment molecular orbital-based density-functional tight-binding method. *J. Comput. Chem.* 38: 1987–1990.
- 64 Takaba, K., Watanabe, C., Tokuhisa, A. et al. (2022). Protein-ligand binding affinity prediction of cyclin-dependent kinase-2 inhibitors by dynamically averaged fragment molecular orbital-based interaction energy. *J. Comput. Chem.* 43: 1362–1371.
- 65 Handa, C., Yamazaki, Y., Yonekubo, S. et al. (2022). Evaluating the correlation of binding affinities between isothermal titration calorimetry and fragment molecular orbital method of estrogen receptor beta with diarylpropionitrile (DPN) or DPN derivatives. *J. Ster. Biochem. Mol. Biol.* 222: 106152.
- 66 Fedorov, D.G. and Kitaura, K. (2018). Pair interaction energy decomposition analysis for density functional theory and density-functional tight-binding with an evaluation of energy fluctuations in molecular dynamics. *J. Phys. Chem. A* 122: 1781–1795.
- 67 Nakanishi, I., Fedorov, D.G., and Kitaura, K. (2007). Molecular recognition mechanism of FK506 binding protein: an all-electron fragment molecular orbital study. *Proteins: Struct., Funct. Bioinf.* 68: 145–158.
- 68 Ozawa, M., Ozawa, T., Nishio, M., and Ueda, K. (2017). The role of CH/ π interactions in the high affinity binding of streptavidin and biotin. *J. Mol. Graph. Model.* 75: 117–124.
- 69 Maruyama, K., Sheng, Y., Watanabe, H. et al. (2018). Application of singular value decomposition to the inter-fragment interaction energy analysis for ligand screening. *Comp. Theor. Chem.* 1132: 23–34.
- 70 Green, M.C., Nakata, H., Fedorov, D.G., and Slipchenko, L.V. (2016). Radical damage in lipids investigated with the fragment molecular orbital method. *Chem. Phys. Lett.* 651: 56–61.
- 71 Li, S., Qin, C., Cui, S. et al. (2019). Discovery of a natural-product-derived pre-clinical candidate for once-weekly treatment of type 2 diabetes. *J. Med. Chem.* 62: 2348–2361.
- 72 Mai, X., Higashi, K., Fukuzawa, K. et al. (2022). Computational approach to elucidate the formation and stabilization mechanism of amorphous formulation using molecular dynamics simulation and fragment molecular orbital calculation. *Int. J. Pharmaceutics* 615: 121477.
- 73 Sladek, V., Tokiwa, H., Shimano, H., and Shigeta, Y. (2018). Protein residue networks from energetic and geometric data: are they identical? *J. Chem. Theory. Comput.* 14: 6623–6631.
- 74 Doi, H., Okuwaki, K., Mochizuki, Y. et al. (2017). Dissipative particle dynamics (DPD) simulations with fragment molecular orbital (FMO) based effective parameters for 1-palmitoyl-2-oleoyl phosphatidyl choline (POPC) membrane. *Chem. Phys. Lett.* 684: 427–432.
- 75 Monteleone, S., Fedorov, D.G., Townsend-Nicholson, A. et al. (2022). Hotspot identification and drug design of protein-protein interaction modulators using the fragment molecular orbital method. *J. Chem. Info. Model.* 62: 3784–3799.

- 76 González-Olvera, J.C., Zamorano-Carrillo, A., Arreola-Jardón, G., and Pless, R.C. (2022). Residue interactions affecting the deprotonation of internal guanine moieties in oligodeoxyribonucleotides, calculated by FMO methods. *J. Mol. Model.* 28: 43.
- 77 Lim, H., Hong, H., Hwang, S. et al. (2022). Identification of novel natural product inhibitors against matrix metalloproteinase 9 using quantum mechanical fragment molecular orbital-based virtual screening methods. *Int. J. Mol. Sci.* 23: 4438.
- 78 Hengphasatporn, K., Wilasluck, P., Deetanya, P. et al. (2022). Halogenated baicalein as a promising antiviral agent toward SARS-CoV-2 main protease. *J. Chem. Inf. Model.* 62: 1498–1509.
- 79 Paciotti, R., Agamennone, M., Coletti, C., and Storchi, L. (2020). Characterization of PD-L1 binding sites by a combined FMO/GRID-DRY approach. *J. Comput.-aided Mol. Des.* 34: 897–914.
- 80 Acharya, A., Agarwal, R., Baker, M.B. et al. (2020). Supercomputer-based ensemble docking drug discovery pipeline with application to covid-19. *J. Chem. Inf. Model.* 60: 5832–5852.
- 81 Gaus, M., Goez, A., and Elstner, M. (2013). Parametrization and benchmark of DFTB3 for organic molecules. *J. Chem. Theory Comput.* 9: 338–354.
- 82 Fedorov, D.G. (2020). Partition analysis for density-functional tight-binding. *J. Phys. Chem. A* 124: 10346–10358.
- 83 Fedorov, D.G. (2021). Partitioning of the vibrational free energy. *J. Phys. Chem. Lett.* 21: 6628–6633.
- 84 Sladek, V. and Fedorov, D.G. (2022). The importance of charge transfer and solvent screening in the interactions of backbones and functional groups in amino acid residues and nucleotides. *Int. J. Mol. Sci.* 23: 13514.
- 85 Fedorov, D.G. and Kitaura, K. (2016). Subsystem analysis for the fragment molecular orbital method and its application to protein-ligand binding in solution. *J. Phys. Chem. A* 120: 2218–2231.
- 86 Paciotti, R., Coletti, C., Marrone, A., and Re, N. (2022). The FMO2 analysis of the ligand- receptor binding energy: the biscarbene-gold(I)/DNA G-quadruplex case study. *J. Comput. Aided Mol. Des.* 36: 851–866.
- 87 Zhang, Q., Zhao, N., Meng, X. et al. (2022). The prediction of protein-ligand unbinding for modern drug discovery. *Exp. Op. Drug Disc.* 17: 191–205.
- 88 Sato, M., Yamataka, H., Komeiji, Y. et al. (2008). How does an S_N2 reaction take place in solution? Full ab initio MD simulations for the hydrolysis of the methyl diazonium ion. *J. Am. Chem. Soc.* 130: 2396–2397.
- 89 Sato, M., Yamataka, H., Komeiji, Y. et al. (2010). Does amination of formaldehyde proceed through a zwitterionic intermediate in water? Fragment molecular orbital molecular dynamics simulations by using constraint dynamics. *Chem. Eur. J.* 16: 6430–6433.
- 90 Nakata, H., Fedorov, D.G., Nagata, T. et al. (2015). Simulations of chemical reactions with the frozen domain formulation of the fragment molecular orbital method. *J. Chem. Theory Comput.* 11: 3053–3064.

- 91 Steinmann, C., Fedorov, D.G., and Jensen, J.H. (2013). Mapping enzymatic catalysis using the effective fragment molecular orbital method: towards all ab initio biochemistry. *PLoS ONE* 8: e60602.
- 92 Pruitt, S.R. and Steinmann, C. (2017). Mapping interaction energies in chorismate mutase with the fragment molecular orbital method. *J. Phys. Chem A* 121: 1798–1808.
- 93 Ishida, T., Fedorov, D.G., and Kitaura, K. (2006). All electron quantum chemical calculation of the entire enzyme system confirms a collective catalytic device in the chorismate mutase reaction. *J. Phys. Chem. B* 110: 1457–1463.
- 94 Ito, M. and Brinck, T. (2014). Novel approach for identifying key residues in enzymatic reactions: proton abstraction in ketosteroid isomerase. *J. Phys. Chem. B* 118: 13050–13058.
- 95 Abe, Y., Shoji, M., Nishiya, Y. et al. (2017). The reaction mechanism of sarcosine oxidase elucidated using FMO and QM/MM methods. *Phys. Chem. Chem. Phys.* 19: 9811–9822.
- 96 Tribedi, S., Kitaura, K., Nakajima, T., and Sunoj, R.B. (2021). On the question of steric repulsion versus noncovalent attractive interactions in chiral phosphoric acid catalyzed asymmetric reactions. *Phys. Chem. Chem. Phys.* 23: 18936–18950.

Part III

Artificial Intelligence in Pre-clinical Drug Discovery

9

The Role of Computer-Aided Drug Design in Drug Discovery

Storm van der Voort¹, Andreas Bender², and Bart A. Westerman¹

¹Department of Neurosurgery, Amsterdam UMC, location VUMC, Cancer Center, Amsterdam, the Netherlands

²Yusuf Hamied Department of Chemistry, University of Cambridge, Lensfield Rd, Cambridge, United Kingdom

9.1 Introduction to Drug–Target Interactions, Hit Identification

The therapeutic effect of drugs is dependent on their interactions with their target molecules, such as kinases, GPCRs, phosphodiesterases, nuclear receptors, or ion channels [1]. Drug–target interactions (DTIs) are not only responsible for therapeutic efficacy but could also lead to adverse events that might conflict with clinical benefits [2]. Therefore, accurate assessment of DTIs is an important step in the drug discovery process, allowing researchers to probe the target properties, efficacy, and safety of a drug, thereby propelling it into various stages of the drug development process.

Hit identification is commonly based on phenotypic assays based on high-throughput screening (HTS) using compounds or fragment-based phenotypic assays, which can subsequently be linked to structural information such as X-ray crystallography or NMR structural information. Subsequently, assessment of DTIs is often done through *in vitro* methods, although these methods have practical limitations when considering the enormous number of potential small-molecule-to-target interactions. Virtual screening is a computational method used to identify potential drug candidates by screening large databases of compounds against a target protein. Therefore, high throughput *in silico* DTI prediction can facilitate the matching of a wide variety of compounds against an array of targets – after which the most promising drug–target combinations can be verified experimentally [3]. Recently, AlphaFold, a neural network (NN)-based predictor of 3D protein structure from the sequence [1], won the 14th Critical Assessment of Protein Structure Prediction [1–4]. This was followed by the publication of 350,000 models of protein structures generated by AlphaFold, showing the potential that NN-based methods have within drug discovery [5], although only a single, *holo*, structure is generated for each protein [6]. Here we will describe how DTI predictors, as well as other complementary predictive models, can assist in the development of new drugs. Figure 9.1 shows that DTI prediction is used at various stages in the clinical development

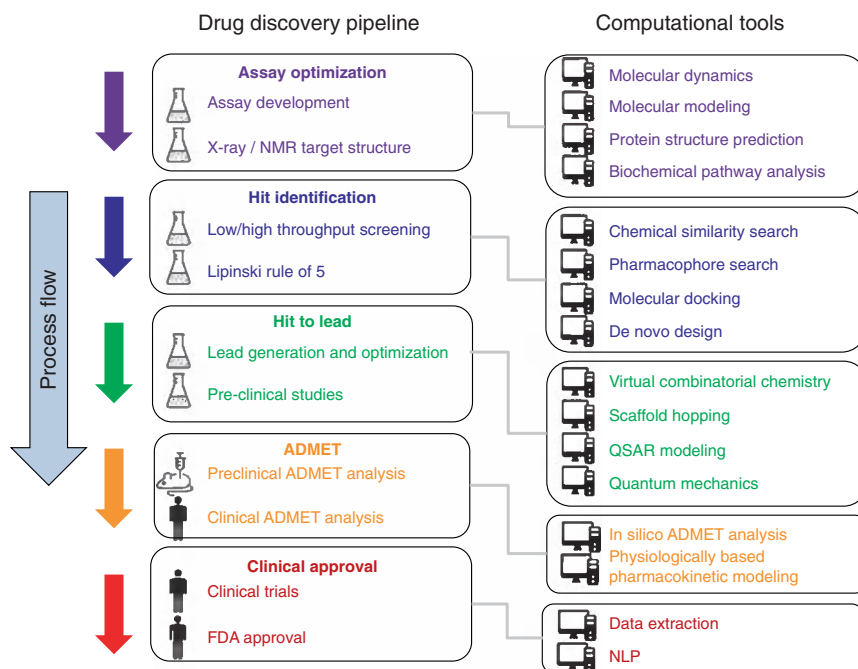


Figure 9.1 Overview of drug discovery pipeline vs. computer-aided design. The figure is partially based on Schaduangrat et al. [7].

pipeline, where each drug discovery stage (left) can be informed by their respective bioinformatic tools (right).

Box 9.1 Hit identification, single or more targets?

An interesting concept in drug discovery is the use of multi-target approaches, relevant for many diseases, such as cancer. A specific multi-target approach is the utilization of polypharmacology, or the use of a single drug to target multiple proteins at once, thereby potentially enhancing its efficacy [4]. This concept can for instance be applied to kinase inhibitors since kinase mutations often drive the process of carcinogenesis [8]. Most of the kinase inhibitors prevent downstream signal transduction by binding to the highly conserved ATP binding site of kinase to block the binding of ATP [9]. Consequently, the binding of a kinase inhibitor to kinases is not highly selective. Given that only approximately 1–2% of kinase inhibitors' targets are known [10–12], proper DTI prediction could uncover this as yet undisclosed target space.

One of the popular techniques in virtual screening is fragment-based screening, which involves breaking down larger molecules into smaller fragments to search for potential binding sites on the target protein. Binding affinity prediction is also a crucial step in drug discovery, where various computational tools are used to predict

the strength of binding between a potential drug candidate and the target protein. The success of drug development depends on understanding protein–ligand interactions, which determine the stability and specificity of the DTI. Therefore, a thorough understanding of these methods is essential for efficient and effective drug discovery.

9.2 Lead Identification and Optimization: QSAR and Docking-Based Approaches

The most established methods for *in silico* DTI prediction are *ligand-based* and *docking-based* approaches [1]. In ligand-based methods like quantitative structure-activity relationships (QSAR), a large collection of confirmed binders to a certain target are collected, and linear (regression) models are built to correlate certain structural features to biological activity [5–7]. This model can then be used to predict the activity of untested molecules. An advantage of QSAR methods is that they do not require structural information about the target protein, making them suitable for protein classes where structural information is scarce, such as GPCRs [8]. However, QSAR modeling has pitfalls such as data overfitting, poor generalizability, and inadequate model validation [13, 14].

In contrast, docking-based approaches use structural information of the protein target to “fit” a molecule into the active site [15, 16]. Here, affinity is typically predicted by assessing the free energy gain upon placing the ligand in the active site using a “scoring function.” However, docking has pitfalls too, as it requires structural information [17], model accuracy is highly dependent on the scoring function used [18–21], and it often fails to incorporate receptor structural flexibility [22]. Furthermore, it is computationally expensive compared to ligand-based methods and therefore less suitable for probing polypharmacology [23].

Matched molecular pair analysis is a common tool used to identify and analyze structural modifications in drug molecules [24]. Knowledge of molecules with similar physical and chemical properties can be used to improve the pharmacokinetics and pharmacodynamics of a drug. Solubility issues are an important issue in drug development, as poorly soluble drugs may have limited bioavailability. Features such as LogD, i.e. the logarithm of the partition coefficient between a drug and water/octanol, allow one to predict a drug’s solubility and distribution [25]. Furthermore, pK_a , the negative logarithm of the acid dissociation constant, affects the drug’s solubility and permeability [26]. A thorough understanding of these parameters can help identify potential drug candidates and optimize their properties for clinical use.

9.3 DTI Machine Learning Methods

Machine learning (ML) methods have recently been aimed to address the shortcomings of traditional ligand-based and docking-based approaches. ML methods in DTI prediction learn from a set of known data points to predict whether a compound

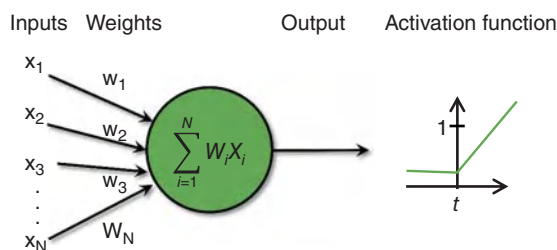


Figure 9.2 (a) An overview of the McCulloch-Pitts model for a neuron, which receives N inputs x with weights w . The neuron sums the inputs and weights to obtain the total input value and activates if the input value exceeds a certain threshold. (b) An example of a neural network containing 7 inputs, one hidden layer, and one output layer. Adapted from Krogh [35].

binds to a target [27]. Generally, a model (a set of rules) is “trained” by making predictions about labeled data. One of the simplest ML methods is linear regression (also used in QSAR), where a line is fit through a set of known data points, and predictions are made by extrapolating from this fitted line [28]. ML approaches have received increased interest over the last years because of their low computational cost, high performance, and applicability to proteins without structural information [29–32].

One of the more recent methods used in ML is deep learning [33, 34] (DL). DL makes use of artificial NNs, thereby mimicking the neural structure of the human brain to generate complex predictive models [35]. NNs are built from neurons, which are connected via links (Figure 9.2, a). NNs are trained by making predictions about a labeled training set, after which the “loss” is defined (the error between the predictions and the actual data labels). The weights of all the neurons are optimized so that the loss function, and therefore the prediction error, is minimized.

9.4 Supervised, Non-supervised and Semi-supervised Learning Methods

ML methods require “training,” i.e. they must be taught how to make predictions based on the input data. There are two main ways of training ML models, supervised and unsupervised learning [36]. The main difference between these training methods is the presence of labeled data [37]. Supervised learning uses labeled training data: data has been pre-labeled (e.g. whether a specific molecule binds to the protein target or not). The model is then trained by making predictions about the input data and comparing the predictions to the label. The variables within the model (connections, thresholds, and weights) are subsequently optimized such that they increase the accuracy of the prediction. A limitation of supervised learning is the need for high-quality labeled data, which can be time-consuming to generate as this is often done through manual curation [37, 38]. Furthermore, since labels are defined in advance, this method has limited capacity to discover new patterns outside the realm of this predefined angle [37].

Unsupervised methods do not use labeled training data and can be used to find patterns or commonalities in data [39]. Unsupervised learning can facilitate to find unique patterns in relation to common patterns within datasets [39]. Unsupervised learning can also be used to preprocess a dataset and drastically cut down on the effort required to label data for supervised learning. This type of combined unsupervised and supervised learning is a form of semi-supervised learning, which is positioned between supervised and unsupervised learning and aims to address the shortcomings of both [39]. Semi-supervised learning can also be applied by training a model on partially labeled data, letting the model infer labels based on partial training during the training process [40].

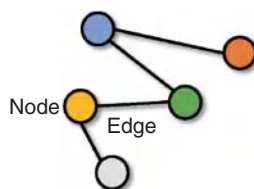
9.5 Graph-Based Methods to Label Data for DTI Prediction

Typical ML methods can be applied to two- or three-dimensional data sources that can be mapped to Euclidian coordinates and can therefore be referred to as Euclidian data [41, 42]. In contrast, graphs have a non-Euclidian structure, featuring nodes (points on the graph) and edges (connections between nodes, Figure 9.3).

Graphs make it possible to examine inputs that are not well represented in Euclidian space. For example, the DTI network is naturally represented as a graph. Drugs and targets make up nodes, and the observed interactions between them are represented as edges (Figure 9.4a). Chemical and protein structures can also be represented as graphs, with atoms or amino acids acting as nodes and chemical bonds or interactions between molecules represented as edges (Figure 9.4b). This more “natural” translation of the DTI network or chemical structures compared to string representations has the potential to make graph-based methods for DTI prediction superior over traditional Euclidian methods.

Still, graphs present additional challenges for computation. Traditional NNs use string representations to perform calculations. However, the irregular nature of a graph makes it hard to perform these calculations. GNNs typically learn by generating representations of the graphs, which are lower-order strings that incorporate information on nodes, edges, or a graph. An example is shown in Figure 9.4c. Here, a representation of target node A is generated by taking its connectivity information, showing it is connected to nodes B, D, and C. An extra layer of the GNN then considers the connectivity information of nodes B, D, and C. This connectivity information is embedded into a lower-dimension representation that can be used as the input for NN-based learning [42, 43].

Figure 9.3 A general overview of the structure of a graph. The nodes (points on the graph) can be connected to each other via edges (connections).



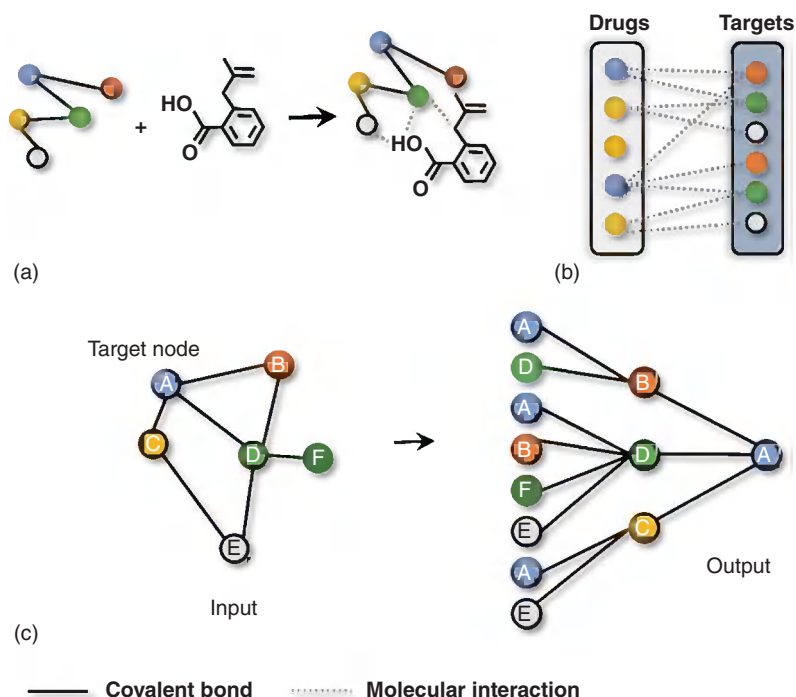


Figure 9.4 (a) Low-order graph methods model the protein and the drug as graphs, using atoms/residues as nodes. (b) High-order methods model the DTI network as a graph, using drugs and targets as nodes. (c) An example of how calculations can be performed on graph-based inputs. Adapted from Zhang et al. [42].

9.6 The Importance of Explainable ML Methods: Linking Molecular Properties to Effects

The main criticism of using ML methods, and especially NNs, for predicting DTIs is their “black-box” nature [44, 45]. Often, the intricate nature of the hidden layer(s) of the NN, combined with the automated nature in which the weights and connections within an NN are optimized, means they are not understood and are therefore not transparent [44]. Since the predictions that NNs make on DTIs are used for predicting drug efficacies and safeties, the models used must be understood by humans. This has been stressed by the recent debate surrounding the “right to explain” language used in the European General Data Protection Regulation, which could imply that “black-box” NNs will not be allowed for important automated decisions [46, 47]. Therefore, it is necessary to have human-understandable or “explainable” ML methods in DTI predictions. Furthermore, explainable ML methods for DTI prediction can offer more insights beyond simply predicting whether a particular compound will bind well. This has the potential to improve the rational design of drugs for a particular protein target, advancing the field of medicinal chemistry. However, this will require a shift away from optimizing ML methods for predictive accuracy and toward ML methods that can be explained.

In order to explain the DTI molecular mechanisms, one would like to pinpoint individual atoms that explain the interaction. However, this molecular interaction assessment is performed in the context of general drug properties required for the solubility and membrane permeability of compounds, hence, this aspect cannot be uncoupled from features that explain the molecular specificity of the compound. Optimal general features consist of physicochemical properties that have values according to Lipinski and others [48, 49], consisting of restrictions of logD, pKa, the number of hydrogen bond acceptors (HBA), intramolecular hydrogen bonds (IHMB) [50], hydrogen bond donors (HBD) [51], and rotational bonds, as well as the polar surface area (PSA). Rather than physicochemical properties, molecular properties are more suited to explain molecular interactions. Different methods have been developed to describe small-molecule compounds, ranging from relatively simple to more extended descriptors. Simple descriptors such as MACCS fingerprints (166 features) have the advantage that the results are relatively easy to understand, at the cost of limited overlap with the enormous amount of possibilities in the chemical space. More complex graph-based descriptors, such as ECFP fingerprints (Extended Connectivity Fingerprint) [52], describe molecules with a fixed number of bits, commonly between 1024 and 4096 bits, as a binary vector representation. Each bit corresponds to a certain molecular feature where the total number of bits is used to describe the molecular structure. A recent development is the use of NNs, including graph convolutional networks (GCNs), that can go beyond the complexity of ECFP fingerprints. Physical-chemical properties and molecular descriptors are therefore valuable to make DTI models explainable, and this is currently a field of innovation [53].

9.7 Predicting Therapeutic Responses

When possible, the identified therapeutics should be customized to patients' molecular profiles to achieve better therapeutic effects [54]. Many studies have shown the value of precision medicine in a clinical setting, especially in the field of cancer [55, 56], especially for kinase inhibitors, which are commonly mutated and therefore patient-specific targets for cancer treatment. With the increasing availability of HTS data and rapid development in the field of genomics, researchers are able to employ computational methods to build drug response prediction models with good performance and interpretability for monotherapy [57, 58] as well as combination therapy [59, 60]. Therefore, the ability to make links between drug effects and patient-specific molecular profiles, and drug sensitivity data required for certain patients to a certain drug is both scientifically as well as clinically relevant. Molecular markers that predict a therapeutic response in the clinic commonly consist of single or multiple defined mutations [61]. In addition, mutations are considered the primary cause of abnormal growth in cancer cells and are associated with drug responses [61, 62]. In spite of that, drug sensitivity predictions based on mutation as well as based on other molecular data such as mRNA expression profiles generally do not perform well [57, 63–66].

9.8 ADMET-tox Prediction

Computer-based predictions of a drug's absorption, distribution, metabolism, excretion, and toxicity (ADMET) form a significant step in drug discovery. The bioavailability of a drug, or its ability to reach the bloodstream, is dependent on its absorption and metabolism. Highly specific compounds should have good evolvability (e.g. favorable molecular and physical properties) to ensure oral bioavailability, half-life, metabolism, transporters, cell permeability, distribution, and solubility. In silico models can predict how well a drug is absorbed in the gut, how quickly it is metabolized, and whether it can cross cell membranes. Cell permeability is also a critical factor in drug discovery, as drugs must be able to penetrate cells to reach their targets. Transporters also play a significant role in drug distribution, and in silico models can predict how a drug interacts with these proteins.

The suitability of a compound for further clinical implementation depends in large part on the potential adverse events. Without proper insight into adverse events, it is possible that therapy could do more harm than good for the patient. The explainability of adverse events based on drug targets would particularly help to provide insight into this balance. So, for a drug therapy to be used safely and effectively, models that can predict the toxicity to the patient of a given therapy would be highly valuable. To create models that can predict the toxicity of compounds or combinations thereof, several data sources can be used, among which are clinicaltrials.gov and the FDA's Adverse Event Reporting System (FAERS) database. Both databases contain millions of patient observations for adverse events linked to drug therapies. Using Natural Language Processing (NLP), structured data can be extracted from these data sources and converted to a structured Common Terminology Criteria for Adverse Events (CTCAE) format [67]. hERG, a potassium ion channel, is a common target for drug-induced cardiotoxicity, and in silico models can predict a drug's risk of causing hERG inhibition [68].

9.9 Challenging Aspects of Using Computational Methods in Drug Discovery

Over the last decades, computational methods have been employed at least since the 1980s, with the first docking programs such as DOCK [69] becoming available. Another peak in the utilization of computational methods happened around 2000, coinciding with the publication of the draft sequence of the human genome [70], which has led to high (and probably also inflated) expectations regarding the impact on drug discovery in subsequent years. As an article from 2001 [71] states, there was the expectation that there would be "3,000–10,000 targets compared with 483," which were targeted at the time of writing – however, an article from 2017 [72] put this number only at or around 667. Hence, it is important to keep in mind also the limitations of novel technologies, such as computational methods in drug discovery.

9.9.1 What are Those Limitations?

Firstly, most computational methods in drug discovery are based on data – which has the advantage that the data that has been generated is used in models (such as HTS data, measurements of and physicochemical properties). On the other hand, the use of data for algorithms reinforces a focus on areas that have been explored already, hence potentially slowing down explorations of novel areas of chemical and biological target space (if such algorithms are based on historic data only). On the other hand, computational approaches can be based on modeling, so on understanding the underlying system, e.g. when modeling receptors and their interaction with a ligand or their activation mechanisms, or, on a scale probably more relevant for physiology, cellular, and higher-level systems. The crux is, though, that biological systems are often poorly understood, rendering this engineering-style approach not immediately feasible in many cases [73].

Secondly, there is the question of *in vivo* relevance of the resulting predictions (an area that has been extensively discussed in a set of two recent review articles [74, 75]). Drug discovery at its core is not about preclinical assays or computational methods – rather, all that matters is the efficacy and safety of a compound in a clinic, for a patient cohort we are able to identify via diagnosis. However, much of the data that is available for generating predictive models at the current point in time is of pre-clinical nature, such as bioactivity data against isolated targets and physicochemical characterization of compounds. This data can be labeled, and hence predictive models can be built – however, the question is whether those predictions translate to clinically relevant endpoints. For example, we are able to predict protein targets of small molecules reasonably well [31] – but whether this translates to *in vivo* relevant effects also depends on the dose and compound pharmacokinetics (reaching the target tissue, etc.). Hence, just predicting unconditional compound properties is not sufficient to generate *in vivo*-relevant computational models. On the other hand, *in vivo*-relevant data, such as clinical readouts, is much more difficult to label, and hence to generate models for – effects obtained depend on the dose, patient genotype, disease endotype, sex, age, co-medication, etc. In addition, such data is available on a much smaller scale, usually only in the number of hundreds or in the small thousands of compounds. So, when it comes to *in vivo*-relevant endpoints, we are dealing with small datasets, which are difficult to label – and hence, it is difficult to generate meaningful models for them. Therefore, there is a disconnect between the data we need to translate into the clinic (*in vivo* relevant data) and the labelable, large-scale data (which is largely *in vitro* data, far removed from clinical relevance).

Thirdly, there are problems with performing meaningful validation of computational models in the context of drug discovery. Unlike biological sequence space, where we are able to assign transition likelihoods for mutations for, e.g. BLOSUM matrices, chemical space is large (on the order of 10^{60} possible compounds) and difficult to characterize. This has severe consequences for all subsequent steps of model generation – since we are unable to identify biases in the data or perform any type of meaningful statistics on the results obtained. Model validation in itself is also insufficient to impact drug discovery as a whole – computational models do not take

any project context into account, and neither do they take into account downstream assays that can (or cannot) be performed for experimental validation of the results obtained. However, those are key factors that matter in drug discovery – when a model predicts toxicity of a certain type; so what does this mean in the context of disease (lifestyle disease or terminal cancer?), dose, reaching the target tissue, etc.? Is the prediction of the model *relevant* in the given context, and *how can the output be confirmed (or refuted)*? It is less likely that projects are stopped just because of a prediction, so integration with the *process* is key here, as illustrated in Figure 9.5 below. Just establishing model performance metrics by themselves, even if they are better than preceding metrics, does not address this aspect of how a model translates to decision-making in a project in practice.

Finally, the generic performance metrics of models, which are frequently found in publications are often irrelevant in practical settings. Performance such as AUC and class-averaged accuracy, are generic – but in a practical setting there is a context of to what extent follow-up assays can be performed, and whether the model operates in an “abundance of options” setting (discovery), or in a “scarcity of options” setting (e.g. in later stage optimization). In the former case, usually sufficient recall in the top few percentages of any ranked library is what is needed (but in absolute terms, this recall may be small, given we are in a situation with many options, say finding active compounds); while in the latter case, a much bigger attention needs to be paid to avoiding false-positive predictions in say a toxicity prediction setting (since losing compounds is very costly and needs to be avoided). All of this depends on the context of how the model is used and what the local experimental follow-up capabilities of a model are – and generic performance metrics are generally not sufficient to be able to translate to this real-world situation.

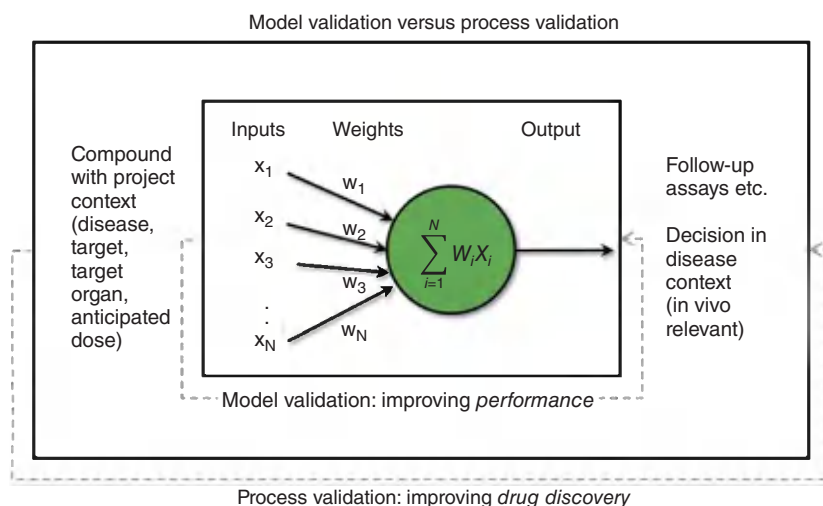


Figure 9.5 Model validation (center) in itself is insufficient to improve drug discovery as a process, since the project context, available follow-up assays, and use of situation-relevant performance metrics also need to be taken into account to have a real-world impact.

References

- 1 Jumper, J. et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596: 583.
- 2 Ozden, B., Structure, A.K. et al. (2021). Function, undefined, and, undefined & 2021, undefined. Assessment of the CASP14 assembly predictions. *Wiley Online Libr.* 89: 1787–1799.
- 3 Pereira, J. et al. (2021). High-accuracy protein structure prediction in CASP14. *Wiley Online Libr.* 89: 1687–1699.
- 4 Jumper, J. et al. (2021). Applying and improving AlphaFold at CASP14. *Proteins Struct. Funct. Bioinform.* 89: 1711–1721.
- 5 Thornton, J.M., Laskowski, R.A., and Borkakoti, N. (2021). AlphaFold heralds a data-driven revolution in biology and medicine. *Nat. Med.* 2710 (27): 1666–1669.
- 6 Perrakis, A. and Sixma, T.K. (2021). AI revolutions in biology. *EMBO Rep.* 22: e54046.
- 7 Schaduangrat, N. et al. (2020). Towards reproducible computational drug discovery. *J. Cheminform.* 12: 9.
- 8 Bhullar, K.S. et al. (2018). Kinase-targeted cancer therapies: progress, challenges and future directions. *Mol. Cancer* 17: 48.
- 9 Ahn, N.G. and Resing, K.A. (2005). Lessons in rational drug design for protein kinases. *Science* (80–) 308: 1266–1267.
- 10 Fabian, M.A. et al. (2005). A small molecule–kinase interaction map for clinical kinase inhibitors. *Nat. Biotechnol.* 23: 329–336.
- 11 Kanev, G.K. et al. (2019). The landscape of atypical and eukaryotic protein kinases. *Trends Pharmacol. Sci.* 40: 818–832.
- 12 Roskoski, R. (2019). Properties of FDA-approved small molecule protein kinase inhibitors. *Pharmacol. Res.* 144: 19–50.
- 13 Kubinyi, H. (2004). Validation and predictivity of QSAR models. In: *European Symposium on QSAR & Molecular Modelling*.
- 14 Cherkasov, A. et al. (2014). QSAR modeling: Where have you been? Where are you going to? *J. Med. Chem.* 57: 4977–5010.
- 15 Pinzi, L. and Rastelli, G. (2019). Molecular docking: shifting paradigms in drug discovery. *Int. J. Mol. Sci.* 20: 4331.
- 16 Ferreira, L.G., Dos Santos, R.N., Oliva, G., and Andricopulo, A.D. (2015). Molecular docking and structure-based drug design strategies. *Molecules* 20: 13384–13421.
- 17 Lavecchia, A. and Di Giovanni, C. (2013). Virtual screening strategies in drug discovery: a critical review. *Curr. Med. Chem.* 20 (23): 2839–2860.
- 18 Neves, M. A. C., Totrov, M., Llc, M. & Abagyan, R. Docking and scoring with ICM: the benchmarking results and strategies for improvement. doi:<https://doi.org/10.1007/s10822-012-9547-0>
- 19 Park, H., Lee, J., and Lee, S. (2006). Critical assessment of the automated AutoDock as a new docking tool for virtual screening. *Proteins Struct. Funct. Bioinform.* 65: 549–554.

- 20 Ha, S., Andreani, R., Robbins, A., and Muegge, I. (2000). Evaluation of docking/scoring approaches: A comparative study based on MMP3 inhibitors. *J. Comput. Mol. Des.* 145 (14): 435–448.
- 21 Mysinger, M.M., Carchia, M., Irwin, J.J., and Shoichet, B.K. (2012). Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J. Med. Chem.* 55: 6582–6594.
- 22 De Vivo, M. and Cavalli, A. (2017). Recent advances in dynamic docking for drug discovery. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 7: e1320.
- 23 Jakhar, R., Dangi, M., Khichi, A., and Chhillar, A.K. (2019). Relevance of molecular docking studies in drug designing. *Curr. Bioinform.* 15: 270–278.
- 24 Griffen, E., Leach, A.G., Robb, G.R., and Warner, D.J. (2011). Matched molecular pairs as a medicinal chemistry tool. *J. Med. Chem.* 54: 7739–7750.
- 25 Hsieh, C.-M., Wang, S., Lin, S.-T., and Sandler, S.I. (2011). A predictive model for the solubility and octanol–water partition coefficient of pharmaceuticals. *J. Chem. Eng. Data* 56: 936–945.
- 26 Navo, C.D. and Jiménez-Osés, G. (2021). Computer prediction of pKa values in small molecules and proteins. *ACS Med. Chem. Lett.* 12: 1624–1628.
- 27 Vamathevan, J. et al. (2019). Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* 186 (18): 463–477.
- 28 Freedman, D.A. (2009). Statistical models: theory and practice answers to selected exercises the labs. *Statistics (Ber)*. 442.
- 29 Ru, X. et al. (2021). Current status and future prospects of drug–target interaction prediction. *Brief. Funct. Genomics* 20: 312–322.
- 30 Mousavian, Z. and Masoudi-Nejad, A. (2014). Drug-target interaction prediction via chemogenomic space: Learning-based methods. *Expert Opin. Drug Metab. Toxicol.* 10: 1273–1287.
- 31 Mayr, A., Klambauer, G., Unterthiner, T., et al. (2018). Large-scale comparison of machine learning methods for drug target prediction on ChEMBL, undefined. pubs.rsc.org
- 32 Bagherian, M., Sabeti, E., Wang, K. et al. (2021). Machine learning approaches and databases for prediction of drug–target interaction: a survey paper, undefined. academic.oup.com
- 33 Carpenter, K.A., Cohen, D.S., Jarrell, J.T., and Huang, X. (2018). Deep learning and virtual drug screening. *Future Med. Chem.* 10: 2557–2567.
- 34 D’Souza, S., Prema, K.V., and Balaji, S. (2020). Machine learning models for drug–target interactions: current knowledge and future directions. *Drug Discov. Today* 25: 748–756.
- 35 Krogh, A. (2008). What are artificial neural networks? *Nat. Biotechnol.* 262 (26): 195–197.
- 36 Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A. & Aljaaf, A. J. A systematic review on supervised and unsupervised machine learning algorithms for data science. 3–21 (2020). doi:https://doi.org/10.1007/978-3-030-22475-2_1
- 37 Rajoub, B. Supervised and unsupervised learning. *Biomed. Signal Process. Artif. Intell. Healthc.* 51–89 (2020). doi:<https://doi.org/10.1016/B978-0-12-818946-7.00003-2>

- 38 Tuia, D., Volpi, M., Copa, L. et al. (2011). A survey of active learning algorithms for supervised remote sensing image classification. *IEEE J. Sel. Top. Signal Process.* 5: 606–617.
- 39 Usama, M. et al. (2019). Unsupervised machine learning for networking: techniques, applications and research challenges. *IEEE Access* 7: 65579–65615.
- 40 Chapelle, O., Schölkopf, B., and Zien, A. (2006). *Semi-Supervised Learning*, 2. Cambridge, MA: MIT Press.
- 41 Bronstein, M.M., Bruna, J., Lecun, Y. et al. (2017). Geometric deep learning: going beyond Euclidean data. *IEEE Signal Process. Mag.* 34: 18–42.
- 42 Zhang, Z. et al. (2022). Graph neural network approaches for drug-target interactions. *Curr. Opin. Struct. Biol.* 73: 102327.
- 43 Zhou, J. et al. Graph neural networks: a review of methods and applications (2021). doi:<https://doi.org/10.1016/j.aiopen.2021.01.001>
- 44 Li, O., Liu, H., Chen, C., and Rudin, C. (2018). Deep learning for case-based reasoning through prototypes: a neural network that explains its predictions. *Proc. AAAI Conf. Artif. Intell.* 32.
- 45 Angelov, P. and Soares, E. (2020). Towards explainable deep neural networks (xDNN). *Neural Netw.* 130: 185–194.
- 46 Ratner, M. (2018). FDA backs clinician-free AI imaging diagnostic tools. *Nat. Biotechnol.* 36: 673–674.
- 47 Selbst, A. D. & Powles, J. Meaningful information and the right to explanation. doi:<https://doi.org/10.1007/s13347-017-0263-5>
- 48 Lipinski, C.A. (2004). Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discov. Today Technol.* 1: 337–341.
- 49 Veber, D.F. et al. (2002). Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.* 45: 2615–2623.
- 50 Kuhn, B., Mohr, P., and Stahl, M. (2010). Intramolecular hydrogen bonding in medicinal chemistry. *J. Med. Chem.* 53: 2601–2611.
- 51 Kenny, P.W. (2022). Hydrogen-bond donors in drug design. *J. Med. Chem.* 65: 14261–14275.
- 52 Thomas, M., Smith, R., Boyle, N.M.O. et al. Comparison of structure- and ligand-based scoring functions for deep generative models: a GPCR case study. In: , 1–39.
- 53 Askr, H., Elgeldawi, E., Aboul Ella, H. et al. Deep learning in drug discovery: an integrative review and future challenges. *Artif Intell Rev.* <https://doi.org/10.1007/s10462-022-10306-1>. Epub ahead of print.
- 54 Chiu, Y.-C. et al. (2019). Predicting drug response of tumors from integrated genomic profiles by deep neural networks. *BMC Med. Genomics* 12: 18.
- 55 Pishvaian, M.J. et al. (2020). Overall survival in patients with pancreatic cancer receiving matched therapies following molecular profiling: a retrospective analysis of the know your tumor registry trial. *Lancet Oncol.* 21: 508–518.
- 56 van der Velden, D.L. et al. (2019). The drug rediscovery protocol facilitates the expanded use of existing anticancer drugs. *Nature* 574: 127–131.
- 57 Iorio, F. et al. (2016). A landscape of pharmacogenomic interactions in cancer. *Cell* 166: 740–754.

- 58 Yang, W. et al. (2013). Genomics of drug sensitivity in cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.* 41: D955–D961.
- 59 Bansal, M. et al. (2014). A community computational challenge to predict the activity of pairs of compounds. *Nat. Biotechnol.* 32: 1213–1222.
- 60 Narayan, R.S. et al. (2020). A cancer drug atlas enables synergistic targeting of independent drug vulnerabilities. *Nat. Commun.* 11: 2935.
- 61 Mina, M. et al. (2017). Conditional selection of genomic alterations dictates cancer evolution and oncogenic dependencies. *Cancer Cell* 32: 155–168.e6.
- 62 Martínez-Jiménez, F. et al. (2020). A compendium of mutational cancer driver genes. *Nat. Rev. Cancer* 20: 555–572.
- 63 Costello, J.C. et al. (2014). A community effort to assess and improve drug sensitivity prediction algorithms. *Nat. Biotechnol.* 32: 1202–1212.
- 64 Jang, I. N. S., Neto, E. C., Guinney, J., Friend, S. H. & Margolin, A. A. Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. *Biocomputing 2014* (2013). doi:https://doi.org/10.1142/9789814583220_0007
- 65 Garnett, M.J. et al. (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* 483: 570–575.
- 66 Liu, H., Zhao, Y., Zhang, L., and Chen, X. (2018). Anti-cancer drug response prediction using neighbor-based collaborative filtering with global effect removal. *Mol. Ther. Nucleic Acids* 13: 303–311.
- 67 Trotti, A., Colevas, A.D., Setser, A. et al. (2003). CTCAE v3.0: development of a comprehensive grading system for the adverse effects of cancer treatment. *Semin. Radiat. Oncol.* 13: 176–181.
- 68 Creanza, T.M. et al. (2021). Structure-based prediction of hERG-related cardiotoxicity: a benchmark study. *J. Chem. Inf. Model.* 61: 4758–4770.
- 69 Kuntz, I.D., Blaney, J.M., Oatley, S.J. et al. (1982). A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* 161: 269–288.
- 70 Lander, E.S. et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409: 860–921.
- 71 Reiss, T. (2001). Drug discovery of the future: the implications of the human genome project. *Trends Biotechnol.* 19: 496–499.
- 72 Santos, R. et al. (2016). A comprehensive map of molecular drug targets. *Nat. Rev. Drug Discov.* 16: 19–34.
- 73 Lazebnik, Y. (2002). Can a biologist fix a radio?—Or, what I learned while studying apoptosis. *Cancer Cell* 2: 179–182.
- 74 Bender, A. & Cortés-Ciriano, I. Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 1: Ways to make an impact, and why we are not there yet. *Drug Discov. Today* (2021). 26(2):511–524.
- 75 Bender, A. and Cortes-Ciriano, I. (2021 Apr). Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 2: a discussion of chemical and biological data. *Drug Discov. Today* 26 (4): 1040–1052.

10

AI-Based Protein Structure Predictions and Their Implications in Drug Discovery

Tahsin F. Kellici¹, Dimitar Hristozov², and Inaki Morao³

¹Department of Computational Drug Discovery, Evotec UK Ltd., Milton Park, Abingdon, Oxfordshire OX14 4RZ, United Kingdom

²Department of In Silico Research and Development, Evotec UK Ltd., Milton Park, Abingdon, Oxfordshire OX14 4RZ, United Kingdom

³Department of Protein Homeostasis, Evotec UK Ltd., Milton Park, Abingdon, Oxfordshire OX14 4RZ, United Kingdom

10.1 Introduction

Proteins perform many different functions within organisms, including catalyzing metabolic reactions, DNA replication, providing structure to cells, responding to stimuli, and transporting molecules. The particular function performed by a protein is determined by its three-dimensional (3D) structure, which in turn is encoded in its 1D amino acid sequence. A computational method that predicts the 3D structure from the 1D amino acid sequence has been long sought. Such a method would open up new doors for both protein function prediction and rational protein design, thus accelerating the discovery of new drugs [1].

Traditionally, knowledge-based methods or physics-based protein-structure prediction have been used. Both methods are appealing but prone to computational issues when applied to even moderate-sized proteins. On the other hand, recent advances in the field of artificial intelligence (AI) have brought the data-driven approach to the fore. Therefore, new tools and software packages (RoseTTAFold [2], AlphaFold [3], trRosetta [4] and trRosettaX [5], ESMFold [6], RGN2 [7], among others) have been recently developed and have produced in silico protein structure predictions of unrivaled accuracy.

In 2020, at CASP 14 (a community-wide blind competition to predict 3D structures of not yet publicly available proteins) [8], AlphaFold2 [3] – a deep neural network approach devised by DeepMind – took the world by storm. It was the top-ranked protein structure prediction method by a large margin, producing predictions with high accuracy and achieving a median GlobalDistanceTest_TotalScore (GDT_TS) of 92.4 (out of 100) across all the proteins to be solved in the competition. GDT_TS over 90 is comparable with the results of laborious experimental determinations of

protein structures and is considered as a correct solution that scientists can rely on with confidence.

A detailed discussion of the technical aspects of the AlphaFold2 implementation is beyond the scope of this chapter. Briefly, the method makes use of the growing availability of public data (PDB) [9, 10] and incorporates novel neural network architectures and training procedures. This allows the simultaneous tuning of the model parameters in order to optimize the final 3D structure. More details can be found in the AlphaFold2 manuscript [3], its extensive support information, and in the GitHub repository [11].

An important feature of AlphaFold2 model is its ability to assign a confidence score per residue to its own predictions. This score is termed the “predicted local-distance difference test” (pLDDT). pLDDT estimates how well the prediction would agree with an experimental structure based on the local distance difference test C α (IDDT-C α). It has been shown to be well-calibrated and to be a competitive predictor of disordered regions [3].

The source code for the AlphaFold2 model, trained weights, and inference script have been publicly released [11]. This has allowed researchers to use and extend the original model. In addition, DeepMind teamed up with the European Bioinformatics Institute (EMBL-EBI) to create the AlphaFold Protein Structure Database [12]. As of August 2022, there are 214 684 311 structures available on the AlphaFold DB website, including 48 complete proteomes for bulk download.

Inspired by the AlphaFold2 success and with the goal of increasing protein structure prediction accuracy for structural biology research and advancing protein design, Baek et al. [2] developed RoseTTAFold – a 3-track deep neural network model that achieved similar performance as AlphaFold2 but with significantly lower hardware requirements. This model allows the generation of 3D protein structures on a single-GPU workstation. In addition, due to its architecture, RoseTTAFold offers the potential to predict complexes of unknown structure that possess more than three chains.

Guided by the advances in natural language processing, a few methods (ESM-Fold [6] and RGN2 [7]) that do not require multiple sequence alignment (MSE) have been recently proposed. Both methods offer comparable performances in some cases while significantly improving the inference time.

In order to render AI-generated protein structures suitable for driving structure-based drug design projects (e.g. via virtual screening, lead optimization [13], etc.), it is usually necessary to reorganize the binding site to accommodate a given ligand series or the generation of a biologically significant conformational ensemble for the protein. This chapter starts with a review of state-of-the-art methods for combining deep-learning structural models with experimental data. Such a combination allows the refinement of the models produced by deep learning alone, making them more suitable for structure-based design. An overview of the combination of AI-based methods with computational approaches follows, leading to a summary of the advances and some of the remaining challenges in using deep learning structural models for drug design and discovery.

10.2 Impact of AI-Based Protein Models in Structural Biology

The accurate prediction of protein structure achieved by the deep learning methods discussed in the introduction has opened new possibilities in integrative structural biology. Experimental models created by methodologies such as cryo-EM and X-Ray crystallography are being combined with predicted models (Figure 10.1) [14]. In X-ray crystallography, predicted models can help with the “phase problem”, an issue that is frequently solved by using molecular replacement with previously solved experimental structures. In the case of nuclear magnetic resonance (NMR), combining models and experimental restraints would offer a direct path to determining the multiple conformational states a protein can assume in solution. In cryo-EM, theoretical modeling methods can be useful to either accelerate model building by providing an initial model or fill parts of the EM map that are at a lower resolution. The remainder of this section presents recent examples of the use of experimental data in combination with deep learning models to provide a set of optimized conformations of the target protein. Such optimized conformations can be further used in drug discovery projects.

10.2.1 Combination of AI-Based Predictions with Cryo-EM and X-Ray Crystallography

Dramatic advances in the technology of electron microscopy (EM) at cryogenic temperatures (cryo-EM) have resulted in the production of an abundance of structures at near-atomic or better resolution [15]. Continuous developments in applying

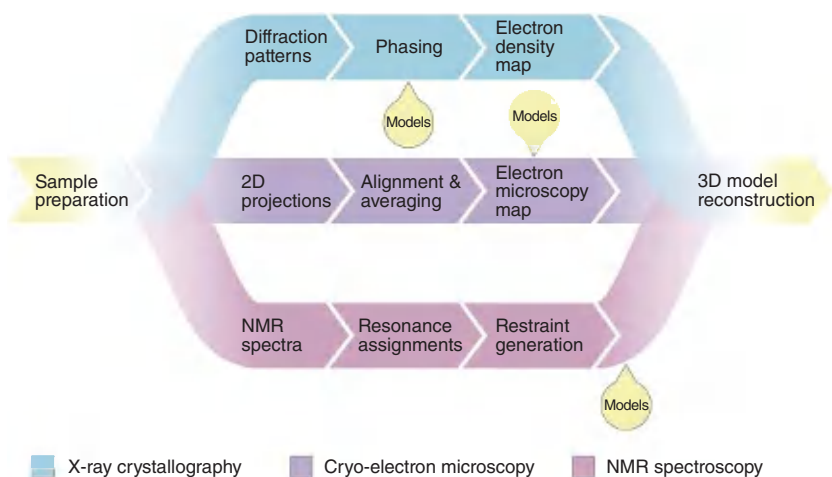


Figure 10.1 Exploiting deep learning models for accurate experimental structure determination of proteins (Figure from Ref. [14]). Deep learning models can help with the phasing of crystal structures, electron microscopy maps, and the generation and interpretation of restraints coming from NMR or other techniques.

cryo-EM to smaller proteins will further advance the power and applicability of this technology for structure-based drug design [16]. The key advantages of cryo-EM over crystallography relate to lower sample requirements, no need for crystal formation, and the fact that cryo-EM allows visualization of samples in various conformational states [17]. Due to these advantages, cryo-EM has found various uses in drug discovery [18], including:

- Handling of cases where the ligand binding site is unknown, and where induced fit phenomena bring large conformational changes to the protein [15]
- Elucidation of targets that involve one or two proteins in oligomeric structures
- Dealing with targets that are resistant to crystallization (thus hindering the establishment of the structure–activity relationship [SAR] and/or hindering the achievement of the target profile by ligand optimization alone).

Computational techniques and predictive modeling play an important role in building atomic models in cryo-EM density maps [19]. Cryo-EM images have extremely low signal-to-noise levels because biological macromolecules are highly radiation-sensitive, requiring low-dose imaging, and because the molecules are poor in contrast [20]. The improved accuracy of deep learning-based protein structure prediction has inspired fresh approaches for model building in cryo-EM density maps, some of which are summarized in Table 10.1. As can be seen from the cited examples, the combination of deep learning methods with cryo-EM data has enhanced the processing of cryo-EM density maps. Two new tools called `phenix.process_predicted_model` and `ISOLDE` further help with the integration of AlphaFold models in the experimental structure determination by X-ray crystallography and cryo-EM [32]. The `phenix.process_predicted_model` tool down-weights or removes low-confidence residues and can break a model into confidently predicted domains in preparation for molecular replacement or cryo-EM docking. These confidence metrics are further used in `ISOLDE` to weight torsion and atom–atom distance restraints, allowing the complete AlphaFold model to be interactively rearranged to match the docked fragments and reducing the need for the rebuilding of connecting regions [32].

A detailed study of the importance of theoretical models in cryo-EM was provided by Hryc and Baker [33]. The authors examined a set of community-established standard density maps representing 12 unique biological datasets ranging from 1.8 to 4.5 Å in resolution. Those targets represent a variety of macromolecular complexes, making them ideal candidates to evaluate the accuracy of predictive modeling in the context of cryo-EM density maps. It was found that AlphaFold2 produced models of superior or equal quality to the state-of-the-art methods used to model the density map for the major capsid protein gp7 of ϵ 15, an infectious bacteriophage, as well as for most of the proteins of Cyanophage Syn5 [33]. The authors also found that the AlphaFold2 prediction confidence score (pLDDT) is a good guide to the overall model quality and accuracy. However, there were a number of examples where pLDDT scores were not sufficient indicators of potential model accuracy [33].

In a similar analysis, Kryshafovych et al. [34] described the solution of four experimental structures using AlphaFold2 models submitted to CASP14 [35] (crystal structure of the inner membrane reductase FoxB; subunits of phage AR9 non-virion RNA polymerase; the crystal structure of the baseplate anchor

Table 10.1 AlphaFold2 models are being used in combination with cryo-EM density maps and crystallography data in order to provide optimized 3D structures.

Target	Description	Electron microscopy database (EMD) accession codes and PDB ID	Ref.
NALCN-FAM155A-UNC79-UNC80 channel complex	NALCN channel mediates voltage-modulated sodium leak currents, which can be blocked by extracellular calcium. The functional NALCN channel is a hetero-tetrameric channelosome. In order to predict the structure of the tetramer, the models of UNC79-UNC80 head and tail regions were predicted by AlphaFold2, docked into a cryo-EM map, and manually adjusted using Coot [21].	EMD-32344 and 7W7G	[22]
FtsH-HflKC AAA protease complex	The membrane-bound AAA protease FtsH is the key player controlling protein quality in bacteria. The predicted FtsHTM hexamer from AlphaFold 2 was manually fitted to the FtsHPD+TM map. Models of FtsHPD+TM-HflKC and FtsHCD were subjected to the Phenix real-space refinement [23].	7W13	[24]
Nuclear ring (NR) and cytoplasmic ring (CR) from the <i>Xenopus laevis</i> NPC	The authors combined "sideview" particles and "tilt-view" particles to overcome the insufficient Fourier space sampling problem and used AlphaFold2 to predict all nucleoporin structures.	EMD32056, EMD-32060, EMD-32061, 7VOP	[25]
Pentameric assembly of the Kv2.1 tetramerization domain	The T1-domain sequences from Kv2.1 and Kv8.2 (in a 3 : 1 ratio) were submitted to the ColabFold notebook to generate the heterotetramer using AlphaFold2-Multimer [26]. The models were inspected, and the top-ranked model was used for the addition of zinc.	7RE5	[27]
Structure of the human glucose transporter GLUT4	An initial structure model for GLUT4 was generated by AlphaFold2. The structure was docked into the density map and manually adjusted and rebuilt by COOT [21].	EMD-32760; EMDB-32761; 7WSM; 7WSN	[28]
Ternary complex of insulin-like growth factor 1 (IGF1) with IGF-binding protein 3 (IGFBP3) and acid-labile subunit (ALS)	Model building of ambiguously resolved parts was aided by a protein model generated from AlphaFold2 and a post-processed map generated from DeepEMhancer [29].	EMD-32735, 7WRQ	[30]
Crystal structure of the Ars2-Red1 complex	The structure of the complex was determined by molecular replacement using the AlphaFold2 Ars2 model (AF-094326) and refined to a Rfree of 30.2% and a Rwork of 24.5%. AlphaFold was also used to model the dimeric structure of the Red1 C-terminus.	7QUU	[31]

and partner TSP assembly region of TSP4 from Bacteriophage CBA120; crystal structure of Afl503 transmembrane receptor) [34]. The authors also reported that the AlphaFold2 models helped improve the structure of an already solved target (the bacterial exo-sialidase Sia24) [34]. Although molecular replacement is a very well-established technique, high-accuracy models are needed, and until recently, this always required the availability of templates with high levels of sequence identity. As the accounts in this paper show, the models provided by deep learning methods are indeed powerful and can be used for molecular replacement [34]. The provided results for the monomeric models of subunits (the phage AR9 nvRNA polymerase, the tail spike protein TSP4-N from bacteriophage CBA120, and the Afl503 receptor) allowed the assembly of complex folds that reflect in large parts the experimentally determined oligomeric structures. Exceptions are the flexible linkers and loops without a defined secondary structure that introduce errors.

10.2.2 Combination of AI-Based Predictions with NMR Structures

NMR spectroscopy is now a well-established technique to elucidate the structure, interaction, and dynamics of molecules in solution and is also used to guide structure-based lead discovery campaigns [36]. The advantages of this technique include

- both chemical compounds and proteins of interest give NMR signals
- the binding mode between ligand and protein, such as conformational transitions upon ligand binding and interaction interface, can be determined at an atomic resolution
- NMR performs well for weak intermolecular interactions with a dissociation constant (Kd) in the $\mu\text{M}/\text{mM}$ range.

Fast NMR data acquisition has led to remarkable improvements in the throughput of high-resolution and sensitive NMR methodologies. These improvements allow the identification of new fragments, thus creating a new avenue for fragment-based drug discovery and development (FBDD) [37].

In order to determine the 3D structure of a protein by NMR, a large number of distance restraints derived from nuclear Overhauser effect data together with spin-spin coupling constants are collected. The most accurate solution structures of proteins, however, are obtained when residual dipolar couplings (RDCs), preferentially for different internuclear vectors, are measured and included in the refinement of the 3D structure. RDCs improve the local backbone geometry of NMR-based solution structures of proteins and can accurately define the relative orientation of secondary structure elements and protein domains. In a recent article, Zweckstetter compared AlphaFold2 structures with NMR structures for the third IGG-binding domain from streptococcal protein G (GB3), the DNA damage-inducible protein (DinI), and ubiquitin [38]. Particularly useful for the comparison of RDCs and RDC-derived solution structures with models predicted by AlphaFold2 is GB3, because of the small rigid domain, the existence of a high-resolution crystal structure, and multiple high-resolution NMR structures using a large number of RDCs [38]. The study shows that 3D structures predicted by AlphaFold2 can be highly representative of the solution conformation of proteins. The excellent agreement of a large number of RDCs with the structures predicted

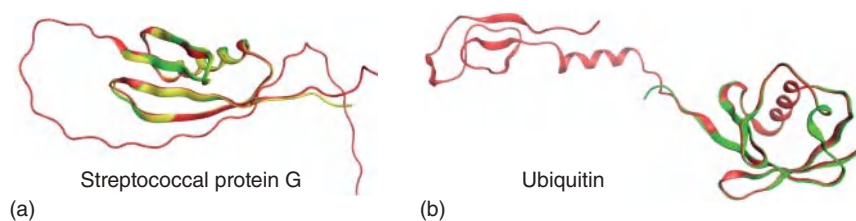


Figure 10.2 a Streptococcal protein G RDC-derived NMR structure (green; PDB id: 1P7F), 1.1 Å X-ray structure (yellow; PDB id: 1IGD), AlphaFold2-structure (red). RMSD value for the AlphaFold model compared to the NMR structure is 0.42 Å; b Ubiquitin: RDC-derived NMR structure (green; PDB id: 2MJB), AlphaFold2-model (red). The structures have an RMSD of 0.65 Å.

by AlphaFold2 for GB3, DinI, and ubiquitin demonstrates the high accuracy of the predicted structures both in terms of local geometry and relative orientation of secondary structure elements (Figure 10.2), that is, the global structure [38]. These proteins provide appropriate cases for a successful AlphaFold2 prediction since they are very small and several high-resolution structures are available in the PDB. Thus were used in the training of the AlphaFold2 neural network. Problems could arise, however, for proteins that do not have these advantages. These problems are alleviated when AlphaFold2 models are combined with RDCs: either the AlphaFold2 model that best fits to the experimental RDCs can be selected (e.g. N-terminal domain of Ca²⁺ + ligated calmodulin) or the AlphaFold2 model can be used as starting structure for RDC-based refinement calculations [38].

Similar encouraging results are reported for the 68-kDa SARS-CoV-2 M^{pro} enzyme, where measured RDCs, using a new, highly precise TROSY-AntiTROSY Encoded RDC (TATER) experiment, are compared with values derived from both high-resolution X-ray structures and AlphaFold2 models [39]. The highest pLDDT-scoring model of the full AlphaFold2 implementation fits RDCs better than 92% of all X-ray structures. Relative to the best X-ray structures, AlphaFold2 M^{pro} models agree more closely with solution RDCs for residues that are part of regular secondary structure than the remainder. This result indicates that catalytic scaffolds are well defined by AlphaFold2. The authors further hypothesize that new opportunities for combining experimentation with molecular dynamics simulations as solution RDCs provide highly precise input for QM/MM simulations of substrate binding/reaction trajectories [39].

In another study, Fowler et al. used the program ANSURR (Accuracy of NMR Structures Using RCI and Rigidity) [40], a software that measures the accuracy of solution structures, and showed that AlphaFold2 tends to be more accurate than NMR ensembles that have been calculated from chemical shifts [41]. In some cases of dynamic structures, however, like the EF-hand domain of human polycystin 2 or the transmembrane and juxtamembrane domains of the epidermal growth factor receptor in dodecylphosphocholine (DPC) micelles, the NMR ensembles are more accurate, and AlphaFold2 had low confidence. The authors found that AlphaFold2 could be used as the model for NMR-structure refinements and that AlphaFold2 structures validated by ANSURR may require no further refinement

[41]. A similar conclusion was reached by Tejero et al. [42]. The team used 12 data sets available for nine protein targets, and the results showed that the AlphaFold2 models have a remarkably good fit to the experimental NMR data. Across a wide range of structure validation methods, including both knowledge-based validations of backbone/sidechain dihedral angle distribution and packing scores, and model vs. data validation against experimental NOESY and RDC data, the AlphaFold2 models have similar, and in some cases, better structure quality scores compared with models generated using conventional structure generation methods in the hands of experts using these same NMR data [42].

The examples provided so far show that in most cases, deep learning models can predict the structures of small, relatively rigid, single-domain proteins in solution without the use of structural templates. These theoretical models could be used for construct optimization, surface analysis for buffer optimization, and site-directed mutagenesis to improve spectral quality by interpreting chemical shift perturbations. At the same time, NMR data can be used for the refining of AlphaFold2 models against RDC, sparse NOE, and chemical shift, as this data takes into account the multiple conformational states of proteins [42].

10.2.3 Combination of AI-Based Predictions with Other Experimental Restraints

Methods like mass spectrometry (MS), fluorescence resonance energy transfer (FRET), double electron–electron resonance (DEER), and electron paramagnetic resonance (EPR) can also help determine the structure of a protein by providing sets of experimental restraints. In the case of native MS, the method allows the topological investigation of intact protein complexes with high sensitivity and a theoretically unrestricted mass range. MS offers the crucial advantage of being able to provide structural data on the proteome scale. For example, proteome-wide crosslinking studies can help to filter biologically irrelevant interactions. Native MS started from a few laboratories in the 1990s, which demonstrated that noncovalent interactions could be preserved in the gas phase for analysis, enabling information on subunit stoichiometry, binding partners, protein complex topology, protein dynamics, and even binding affinities from a single mass spectrometric analysis [43]. Native MS does not yield detailed molecular structure information, but it has some major advantages over traditional structural biology methods, like speed, selectivity, sensitivity, and the ability to simultaneously measure several species present in a mixture [44]. These characteristics make native MS a method that requires just a fraction of the sample needed to solve structures by NMR spectroscopy or X-ray crystallography [44].

In a recent article, Allison et al. assessed whether native MS can be used in order to verify the plausibility of structural models generated by AlphaFold2 [45]. Three protein complexes whose interactions involve disordered regions, ligands, and point mutations were selected for evaluation and analysis: the structure of dihydroorotate dehydrogenase (DHODH), the small heat shock proteins (HSP) 17.7 and 18.1 from *Pisum sativum*, and the N-terminal domain (NT) of the spider silk protein Major ampullate Spidroin 1 (MaSp1) from *Euprosthenops australis*. In the case of DHODH, the protein contains a central cavity, which in the experimental structures is occupied by the cofactor flavin mononucleotide (FMN). Overlaying the ligand binding

sites of the AlphaFold2 prediction and the X-ray structure revealed a nearly identical arrangement of the residues that coordinate FMN. Native MS, in combination with crosslinking and ion mobility (IM) measurements, showed that the human protein cannot maintain the correct conformation in the absence of FMN in MS, which strongly supports that FMN is required to adopt a stable conformation. Native MS can inform about the role of the cofactor in promoting the correct fold of DHODH, a role that is not evident from the ML-based prediction alone. In the case of HSP 17.7 and 18.1, AlphaFold2 could correctly predict both homodimers but also the hypothetical HSP 17.7–18.1 heterodimer with an equal per-residue confidence score. Native MS, however, revealed homodimer formation, while at the same time suggesting that heterodimerization is practically impossible. The last example, MASP1 is monomeric above, and dimeric below, pH 6.5. This pH sensitivity is partially due to a conserved salt bridge between Asp39/Asp40 and Lys65 on the opposing subunit. AlphaFold2 was used to predict the structure of the dimeric wild-type protein, as well as a point mutant with a weakened salt bridge, Asp40Asn. AlphaFold2 predicts with the same confidence the structure of the dimer in both the wild-type and the mutant. However, native MS analysis of both proteins at pH 6.0 showed that the Asp40Asn mutation abolished dimerization nearly completely, showing that the impact of losing this salt bridge on dimer formation requires experimental validation [45].

Other sparse experimental data like DEER restraints have been used in combination with Rosetta [46] in order to model conformational changes in proteins. In order to integrate these restraints, a new tool was created called ConfChangeMover (CCM). The performance of CCM was evaluated in both soluble and membrane proteins using simulated or experimental distance restraints, respectively. The main advantage of CCM over other methods stems from its ability to automatically identify, group, and move secondary structural elements (SSEs) as rigid bodies, a task that can be combined nicely with Rosetta [47]. More recently, del Alamo et al. [48], reported an investigation of the conformational dynamics of amino acid-polyamine-organocation transporter (GadC), a protein aiding the exchange of γ -aminobutyric acid (GABA) with extracellular Glu, using DEER spectroscopy. The analysis was assisted by generating an ensemble of structural models in multiple conformations using AlphaFold2, as described in [49]. The observed correspondence between conformational changes predicted by AlphaFold2 and distance changes observed by DEER is striking. AlphaFold2 predicted that the transmembrane helix 10 (TM10) acts as an extracellular thin gate. The ensemble of models coupled with DEER data suggested that the motion of TM10 was tightly coupled to that of TM9. Additionally, the dynamics of TM10 could distinguish between glutamate and GABA [48].

10.2.4 Impact of Deep Learning Models in Other Areas of Structural Biology

AI-based predicting methods can also help with protein construct design and engineering protein surfaces for protein crystallization. The predicted structure of a fold and the pLDDT score for each residue can be utilized to locate the less compact and disordered regions. Often, omitting less ordered regions from a protein sequence is

beneficial to design well-behaved recombinant proteins for structural studies [50]. Furthermore, these methods can aid with mutational surface engineering, a method used to create patches with low conformational entropy in order to achieve an enhancement in the resolution of a crystal structure [51]. In this direction, DeepREx-WS, a webserver that assists with the identification of residues to be varied in protein surface engineering processes, provides helpful results with AI-based protein structures [52].

The examples provided in this section suggest that AlphaFold2 predictions are generally highly accurate; however, as Terwilliger et al. show, many parts of these predictions are incompatible with experimental data from corresponding crystal structures [53]. The combination of deep learning models with electron density or electron microscopy maps resulted in very high-quality models. Table 10.1 shows that theoretical models derived from AlphaFold2 are becoming quite popular in aiding the determination of complex biological macromolecules. When it comes to methods like native MS, DEER spectroscopy, etc., even though they do not provide direct structural details, they can detect a wide variety of protein interactions and aid in the generation of reliable models. In the future, methods like hydrogen/deuterium exchange (HDX) MS should be combined with ML for monitoring structural and dynamic aspects of proteins in solution. Restraints derived from NMR, EPR, or other techniques can be exploited either by defining the modeling question a priori or by employing the experimental data to identify a likely model a posteriori. In addition to the examples listed previously, restraints may be obtained from data collected with EPR and other sources of geometrical restraints such as FRET and cross-linking. Ultimately, the combination of experimental techniques and 3D protein models derived from deep learning produces more reliable models, which can be further exploited in drug discovery projects. It must be noted, however, that the given examples do not specifically address the influence of bound ligands, flexible regions, and point mutations on protein interactions and that further investigation is required in this direction.

10.3 Combination of AI-Based Methods with Computational Approaches

In the past years, physics-based refinement and force-field-based methods have served their purpose as an orthogonal approach for improving the quality of protein models that were predicted by informatics-based approaches [54]. Until recently, deep learning methods were considered to be a complement to simulation in macromolecular modeling and not a way of replacing them. That was mostly the result of the fact that deep neural networks typically have a multitude of parameters that must be optimized during training. This created the danger of overfitting: training on limited data that could result in a model perfectly tuned to reproduce the training set but unable to correctly predict new inputs [54]. The emergence of highly accurate structure prediction by machine learning is now raising questions about the limitations and future role of physics-based refinement. The machine

learning-based models still have deficiencies, but further refinement has become much harder, even though there are still multiple issues to be solved. For example, the sampling problem continues to be a major challenge, and it is still difficult to create different conformational states from a given initial model [55, 56].

One of the first attempts to optimize deep learning models and create a conformational ensemble from AlphaFold2 structures was performed by Heo et al. [55]. The overall refinement protocol used by this group at CASP14 challenge consisted of three major components, as illustrated in Figure 10.3. The pre-sampling step consisted of oligomeric state prediction, putative binding ligands, and the possibility

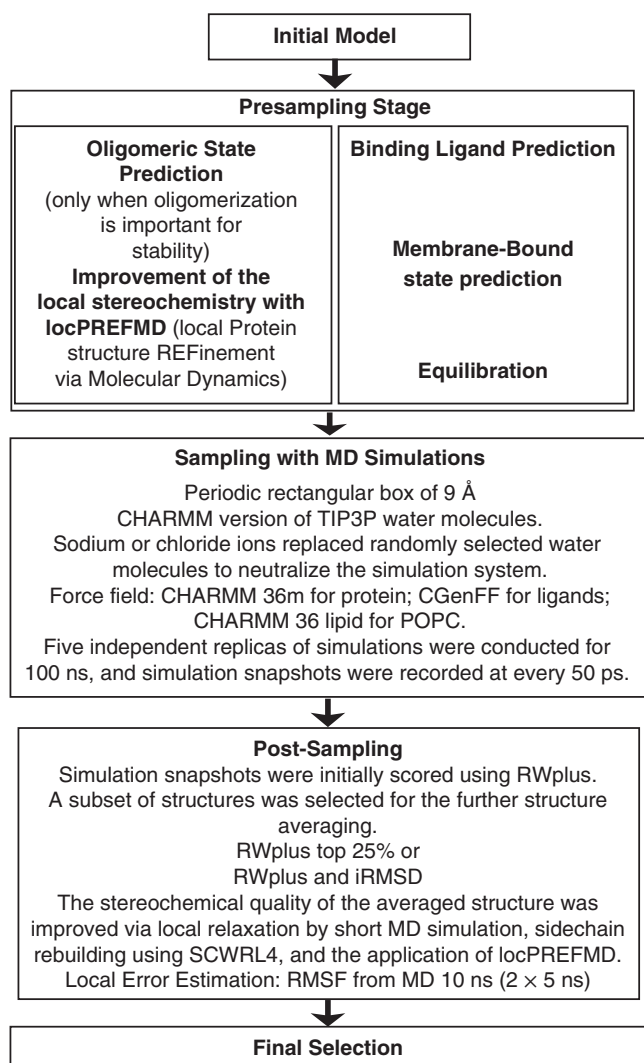


Figure 10.3 Overview of the refinement protocol as established. Source: Adapted from Heo et al. [55].

of membrane interactions. These were predicted manually based on homologous structures searched by HHsearch [57]. For the sampling step, simulation systems for non-membrane-bound proteins were constructed in an explicit water box. A periodic rectangular box was constructed with a minimal distance from any protein atom to the closest box edge of 9 Å. Empty spaces in the box were filled with the CHARMM version of TIP3P water molecules [58]. Protein conformations were sampled via molecular dynamics simulations. In principle, the protocol described by Heo et al. reached near-atomistic accuracy for many targets. However, this required long enough simulations to sample the native state with an assumption that the native state remains the lowest free energy state as other non-native conformations are being explored. This protocol was able to refine models from other predictors, including many models generated based on machine learning methods. However, when it came to improving AlphaFold2 models, the group experienced significant difficulty. A simple explanation may be that AlphaFold2 models already had very high accuracy to begin with. In terms of the global distance test – high accuracy (GDT-HA) metric [59], one of the most frequently used in literature and in CASP experiments, only 26 out of 87 TS domains had less than 70 GDT-HA units. Thus, not many models required refinement. However, even for models that had significant errors that should have been fixed, refinement was not very successful. The Markov State Model analysis and relaxation of the experimental structures revealed several issues with the current refinement protocol. First, there was still a sampling problem with the MD-refinement protocol. To reach the native state from the initial model state via MD simulations, it had to overcome several kinetic energy barriers for partial unfolding and refolding. However, the time required for the transitions was much longer than simulations for refinement, and state transitions were prohibited by the restraints. As a result, the sampled structures during the refinement simulations hardly deviated from the initial AlphaFold2 models. These were some interesting results, taking into account that a similar protocol when applied by the same group on AlphaFold models from CASP13 showed that physics-based refinement improved the accuracy of machine-learning models to exceed the accuracy of any other available method based on the targets that were assessed [56].

The applicability of AlphaFold2-generated protein conformations in virtual screening experiments – an important part of most early drug discovery projects – has been investigated in [60] with its performance being evaluated across a range of targets using the DUD-E dataset [61]. Out of the box, many AlphaFold2 structures produced low enrichment, hinting that the AlphaFold2 predictions may need further refinement before being used in the context of virtual screening [62]. These results could be the consequence of low confidence loops occluding part of the binding site, missing co-factors, and uncertainty in relative domain orientation. Where a comparison could be made, the authors found that unrefined AlphaFold2 structures deliver similar enrichments to those of an apo experimentally derived structure, significantly below the enrichments using an experimentally derived holo structure [60]. Meanwhile, the application of induced fit docking coupled with molecular dynamics (IFD-MD), a method that combines ligand-based pharmacophore docking, rigid receptor docking, and protein structure prediction with explicit solvent

molecular dynamics simulations [63], can induce a binding site conformation that delivers enrichments much closer to the holo structure. This is also supported by the finding that the average binding site volume of the IFD-MD refined AlphaFold2 structure is closer to the holo structure than the raw AlphaFold2 structure [60].

Encouraged by these results [60], the authors went one step further by investigating a total of 14 protein targets, each of which consists of a congeneric set of active ligands along with a co-crystallized structure with one of those ligands [13]. Seven of the data sets are taken from the 2015 paper in which the FEP+ methodology was introduced [64], plus one homology model of PDE10A, which was used as an isolated test case; the remaining six come from internal Schrodinger drug discovery projects. In each case, the authors evaluated the performance of IFD-MD for several different homology models based on templates with differing sequence identities (roughly 30%, 40%, and 50%, although templates in all three of these categories are not available for every target). For this task, they used the ligand for which a co-crystallized structure is available for the IFD-MD calculations (so as to be able to evaluate the RMSD from the experimentally determined structure). The authors exported the top 5 poses produced by IFD-MD and carried out FEP calculations for the entire congeneric series of ligands for each pose. The final pose is selected using a scoring function, which combines several performance metrics from the FEP calculations (correlation coefficient, RMS error) as well as the absolute binding free energy. This protocol (as shown in Figure 10.4(a)) could be generalized and applied to many potential structure-based drug discovery projects, requiring experimental binding affinity data for a congeneric series obtained either from the literature (publication or patent) or in-house experiments [13]. A key aspect of this refinement protocol (Figure 10.4(a)) is the use of binding data from a ligand series to select the most appropriate protein-ligand complex structure. The ambiguity and noise that are present in a typical homology model or from deep learning methods could be addressed by differentiating proposed options with ligand-based information. These results suggest that the IFD-MD and FEP calculations provide a way to combine protein structure prediction and ligand binding information [13]. In a similar way, Beuming et al. [65] used AlphaFold2 models in order to evaluate the performance of FEP+. The authors, in order to generate the MSA, employed three databases: BFD [66], Mgnify [67], and Uniref90 [68]. The ligand was introduced into the *apo* model coming from AlphaFold2 by aligning the model with the crystal structure used for the original FEP calculations (Figure 10.4(b)). The Mean Unsigned Error of the individual perturbations for calculations done with AlphaFold2 is comparable with those performed with crystal structures, and in many cases, the R^2 values are similar to the expected values for well-behaving FEP calculations. It needs to be highlighted, however, that in this method, the introduction of the aligned ligand was performed through superposition with crystal structures. As a result, the conclusions are highly dependent on whether the initial binding poses are accurate enough.

Another approach [69] combined deep learning approaches with mechanistic modeling for a set of proteins that experimentally showed conformational changes by using trRosetta [4] as a deep learning predictive platform (Figure 10.5). By combining DeepMSA [70], with deep residual-convolutional network trRosetta

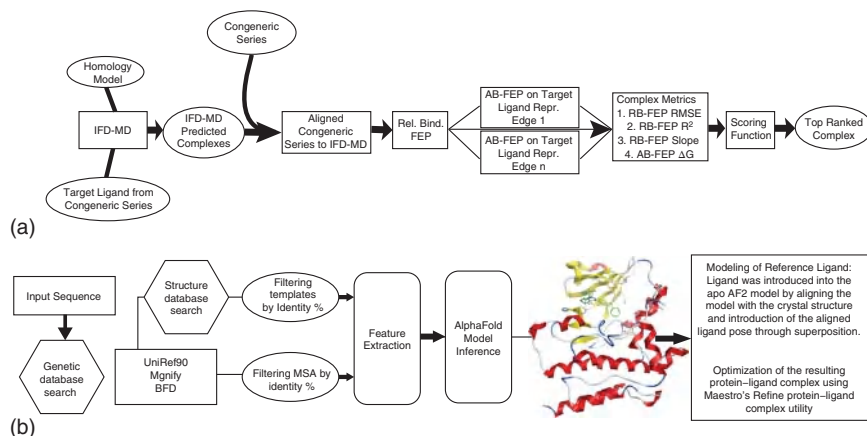


Figure 10.4 AlphaFold2 models in combination with Induced-Fit Docking or other techniques enable accurate free energy perturbation calculations [13, 60, 65].

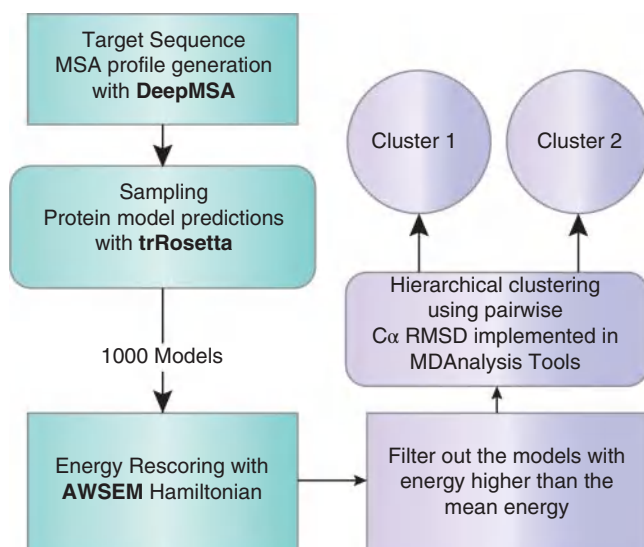


Figure 10.5 Workflow of the protein-folding pipeline used in [69]. Source: Adapted from Audagnotto et al. [69].

and the AWSEM force field [71], the authors observed that both X-ray structures of the different protein states and the similar intermediate states explored by the MD simulation were predicted. To test the ability of the pipeline to predict protein conformational ensembles, the authors investigated only X-ray structures with a maximum length of roughly 200 amino acids, a resolution equal to or less than 2.40 Å, and where more than one conformation was available in the PDB for the same sequence [69]. Four test cases were taken into consideration (Adenylate kinase, α I-domains of LFA-1, Myoglobin protein, T4 lysozyme, and Tetrahymena

thermophila-BIL2). It was observed that the conformational space explored by the predicted models was similar to the one observed during the MD simulations of the experimental protein structure. In particular, for the adenylate kinase and the T4 lysozyme predictions, it was possible to predict the active and inactive structures as well as the intermediate conformations observed during the MD simulations. Although the trRosetta algorithm has been trained on static PDB structures, it was able to reproduce protein flexibility. Limitations of the technique have been observed in the case where metal ions (LFA1 test case) and cofactor (Myoglobin test case) were present influencing the conformational equilibrium. It is important to notice that the ability to predict protein flexibility is correlated to the number of available structures for the different protein conformations in the PDB, and this is the strongest limitation of the current machine-learning techniques. Another important aspect is the quality of the initial MSA profile. The authors have noted that the choice of the MSA algorithm can hamper the model prediction by favoring one conformation [69].

In a similar attempt, del Alamo et al. used a set of eight membrane proteins (transporters and GPCRs) representing different structural classes and mechanisms of action to set up an approach driving AlphaFold2 to sample their alternative conformations [49]. These included five unique transporters (LAT1, ZnT8, MCT1, STP10, and ASCT2), whose structures had been previously experimentally determined in both inward- and outward-facing conformations, and three representative G-protein-coupled receptors (CGRPR, PTH1R, and FZD7), whose structures had been solved experimentally in active and inactive states. None of these proteins were part of the original AlphaFold2 training set, which included structures located in the protein data bank (PDB) [49, 72]. The sequences of all targets were truncated at the N- and C-termini to remove large soluble and/or intrinsically disordered regions that represent a challenge for AlphaFold2. Prediction runs were executed using AlphaFold v2.0.1. The pipeline used in this study differs from the default AlphaFold2 pipeline in several aspects. First, all MSAs were obtained using the MMSeqs2 server [73]. Second, the search for a template was disabled, except when explicitly performed with specific templates of interest. Third, the number of recycles was set to one, rather than three by default. Finally, models were not refined following their prediction. This study utilized all five neural networks when predicting structures without templates, with 10 predictions per neural network per MSA size. The results indicated that AlphaFold2 could be manipulated to accurately model alternative conformations of transporters and GPCRs whose structures were not available in the training set. The use of shallow MSAs was instrumental to obtaining structurally diverse models in most proteins, and in one case (MCT1) accurate modeling of alternative conformations also required the manual curation of template structures. Thus, while the results presented here provide a blueprint for obtaining AlphaFold2 models of alternative conformations, they also argue against an optimal one-size-fits-all approach for sampling the conformational space of every protein with high accuracy. Moreover, this approach showed limited success when applied to transporters whose structures were used to train AlphaFold2, hinting at the possibility that traditional methods may still be required to capture alternative conformers [49]. The work extends the scope of AlphaFold2 beyond the structure

prediction of a single state to the exploration of the conformational diversity of proteins. Even though determining the populations of alternative conformations and the interconversion pathways between them still appears to be out of reach, this work represents a crucial step toward describing the dynamic nature of proteins with modern artificial intelligence-based structure predictors [72].

The structures of membrane proteins, especially GPCRs, have been thoroughly compared with deep learning models in multiple recent articles. He et al. evaluated the performance of AlphaFold2 on GPCRs by analyzing 29 GPCR structures released after the publication of the AlphaFold2 database, thereby making sure that the prediction for these GPCR structures did not involve the experimental structural information [74]. The study was focused on subdomain assembly, ligand binding, and functional state of the receptors. Even though AlphaFold2 achieves a C α RMSD accuracy of ~ 1 Å in protein structure prediction, the resulting deep learning models cannot be used directly for structure-based drug design of GPCRs. AlphaFold2 shows limitations in predicting the assembly between extracellular and trans-membrane domains and the transducer-binding interface of a GPCR. Molecular docking was performed against the orthosteric sites in their predicted models and experimental structures. Different binding poses of ligands were observed between AlphaFold2 models and experimental structures with large RMSDs (especially in the case of 5-hydroxytryptamine receptor 1F) due to distinct sidechain conformations. Although C α RMSD in the predicted model is very low, the predicted model has a narrower pocket compared with that of the experimental structure.

These findings are further confirmed by a recent report from Nicoli et al. [75], in which they use the AlphaFold2 model of OR5K1 [a member of the odorant receptors (OR)] for analysis and compare it to structures coming from traditional template-based homology modeling (HM). This allowed the authors to evaluate the use of AlphaFold2 OR structures for ligand-protein interaction studies. AlphaFold2 and homology models have differences in the backbone that unavoidably affect the binding site conformations. A difference between HM and AlphaFold2 models is the activation state. The prevalence of GPCR models in the inactive state in AlphaFold2 has been addressed in a recent paper by Heo et al. [76], and the authors found that this may also affect the accuracy of binding site predictions. The refinement process of the AlphaFold2 model included multiple consequent steps of IFD with the most potent compound. This step was needed not only to improve the performance, as for the homology model, but also to open the orthosteric binding site and allow the docking of agonists [75].

10.3.1 Combination of Structure Prediction with Other Computational Approaches

AI-based protein structures are being exploited for multiple purposes that have an impact on drug design. Binding site prediction provides important information to uncover protein functions and to direct structure-based approaches. Cavity prediction tools like PrankWeb 3 [77], CavitySpace [78], GraphSite [79], and ProBiS-Fold [80] incorporate now AlphaFold-generated models and provide data on the

shape and physicochemical properties of ligand binding sites and help with druggability assessment. The ligand site prediction can lead to the comparison of pockets for drug repurposing and the prediction of off-target activities. In that direction tools like PrePCI [81], a web server that predicts the interactions between proteins and small molecules and uses, among other sources, AlphaFold structures, and DrugMAP [82] are useful resources for generating such information as well as lead candidate selection and identification of metabolites involved in mediating cellular processes. Models coming from AlphaFold have also been helpful in the analysis of cysteines in chemoproteomic datasets [83]. This reactive cysteine profiling plays an important role in covalent drug discovery.

Deep learning methods in combination with Markov Chain Monte Carlo optimization have shown great promise in protein design [84]. Anishchenko et al. showed that the trRosetta deep neural network trained using multiple sequence information could predict 3D protein structures for *de novo* designed proteins from a single sequence even in the complete absence of co-evolution information [85]. *De novo* protein design is the next frontier when it comes to drug discovery. Innovations like designing mimetics of natural immune proteins with augmented therapeutic affinity and activity but diminished immunogenicity and toxicity can be improved and expedited by using these methods [86].

AI-based methods have also been deployed in the difficult task of protein–protein docking. Protein–protein interactions are responsible for a number of key physiological processes. Modulators can target the interfaces of these interactions, called “hotspots.” Structure-based design techniques can be applied to design PPI modulators once a three-dimensional structure of the protein complex is available. Protocols like Fold-and-Dock (based on trRosetta) [87], FoldDock (based on AlphaFold) [88], AlphaFold Multimer [26], and AF2Complex [89, 90] have all shown promising results when compared to methods that are based on shape complementarity and template-based docking [88].

10.4 Current Challenges and Opportunities

The achievements of AlphaFold2 and the rest of the deep learning methods that predict a 3D structure of a protein from its 1D amino acid sequence have been impressive. However, it needs to be noted that knowledge-based methods lack a fundamental energy framework. Such approaches use available data to make extrapolative predictions regarding related biological and chemical systems [91]. This limits the application of deep learning to cases in which very large numbers of training examples exist and requires careful testing of generalization error (using validation examples not used in training) and manual tuning of model hyperparameters controlling the training to minimize overfitting. The fact that the AlphaFold2 algorithm was able to create biologically-relevant conformational ensembles of proteins via workflows described in this chapter (Figures 10.3–10.5) does not help in finding the most appropriate structure for assisting in a medicinal chemistry project. The output, as shown above, will be multiple equivalent structures that

need to be assessed and evaluated. One further item to be noted is that the studies that were referenced in this chapter have been retrospective in nature, and that in cases where there is not much previous knowledge on the target, these approaches will not facilitate the drug discovery process. What could help is combining the output of these deep learning methods with experimental data coming from NMR, EPR, MS, FRET, etc., as shown in the first part of this chapter.

Another question that needs to be answered is whether pLDDT and related model quality metrics are sufficient for judging the quality of the model. The accuracy of predictive models must still be evaluated based on their agreement with experimental data. With a growing training set as more and more structures become available in RCSB, predictive modeling may eventually achieve the level of accuracy needed to model dynamic protein structures and complexes. Until then, AlphaFold2 and other predictive modeling techniques, despite all their successes, cannot replace experimental methods [33].

A major issue that limits the applicability of AlphaFold2 and related deep-learning methods in drug design projects is the fact that the predicted protein conformations do not take into ligands account. In the future, these deep learning methods could be used to reliably predict the structures of protein–ligand interactions [92]. For example, AlphaFill [93, 94] is an algorithm based on sequence and structure similarity that aims to “transplant” such “missing” small molecules and ions from experimentally determined structures into predicted protein models. These publicly available structural annotations are mapped to predicted protein models, to help scientists interpret biological function and design experiments. Co-folding algorithms are also being developed, allowing the generation of protein–ligand complexes. A possible workaround is the use of binding data from a ligand series to choose between multiple options for the protein–ligand complex structure. The ambiguity and noise that are present in a typical homology model, even with the most recent advances, could be addressed by differentiating proposed options with ligand-based information. The examples given above show that an approach that combines IFD and/or MD with deep learning models could provide useful insights into the binding mode and prioritization of ligands.

Some other issues affecting the quality of deep learning structural models include:

- *Intrinsically disordered proteins/regions (IDPs/IDRs)*: As noticed by Wilson et al. [95], considering a residue that pertains to a helix, strand, or β -turn in an AlphaFold2 structure, as ordered, and otherwise as disordered, results in an overestimation of disorder content and a poor prediction of disordered regions. While this may seem like a trivial observation, the abundance of AlphaFold2 structures generated for disordered proteins has made such a pitfall increasingly likely for researchers who are less familiar with IDPs and structural prediction methods. Another issue is predicting the structural dynamics and transitions (i.e. order-to-disorder, disorder-to-order, disorder-to-disorder) that an IDP may undergo [95]. The development of machine-learning methods for IDPs and polypeptide aggregation cannot be ruled out; however, at the current stage, more structural data is needed for these proteins and to link the different structures

to their stabilities [96]. Intrinsically disordered regions are not the only parts where AlphaFold2 predictions struggle. It has also been observed that modeling loops remains difficult when using neural network-based methods [97]. The poor prediction of these regions has in some cases a strong impact on the quality of the models.

- *Fold-switching proteins*: Fold-switching proteins respond to cellular stimuli by remodeling their secondary structures and changing their functions. Contrasting IDPs/IDRs, which are natively unstructured, fold-switching proteins have regions that either assume distinct stable secondary and tertiary structures under different cellular conditions or populate two stable folds at equilibrium [98]. 94% of AlphaFold2 predictions captured one experimentally determined conformation but not the other. Despite these biased results, AlphaFold2's estimated confidences were moderate-to-high for 74% of fold-switching residues [98].
- *Glycosylation*: The absence of cofactors and of co- or post-translational modifications in the models in the AlphaFold protein structure database is of particular importance when it comes to glycosylation. This issue might be remediated using sequence and structure-based comparative studies. It appears that the space where glycosylation happens is somehow preserved in AlphaFold2 models. This allows for these structural features to be directly grafted onto a model [99].
- *Folding pathways*: Outeiral et al. investigated whether state-of-the-art protein structure prediction methods can provide any insight into protein folding pathways [100]. The team generated tens of thousands of folding trajectories with seven protein structure prediction programs, obtained a set of AlphaFold2 trajectories, and used them to determine major features of folding using a simple set of statistical rules. It was found that protein structure prediction methods can in some cases distinguish the folding kinetics (two-state versus multistate) of a chain better than a random baseline, but not significantly better, and often significantly worse, than a simple, sequence-agnostic linear classifier using only the number of amino acids in the chain. In a recent opinion by Chen et al., the results of AlphaFold are compared with “interpreting a movie by fast-forwarding to the final scene without first watching the previous two hours” [101]. Scientists can see the result of the folding process but not the actual process.
- *Mutations*: Understanding the impact that missense mutations have on protein structure helps to reveal their biological effects [102, 103]. Recent papers from Sen et al. and Buel et al. showed that AlphaFold2 could not predict the full extension of the impact of a mutation. For example, alanine substitution causes the ubiquitin-associated domains (UBAs) to become intrinsically disordered; however, AlphaFold2 predicted alanine-substituted UBA1 or UBA2 to be structurally equivalent to WT UBA with only minor differences in the fold.

Although what has been listed here is a brief outline of the shortcomings of deep learning structural models, these issues also provide insights into the problems of current experimental methods. In order to build improved models, better and more training data are needed. Experiments and modeling methods are required to sample the entire conformational space of proteins. The machine learning

methods themselves will also have to evolve to include ligands, post-translational modifications, and complexes of different types of molecules.

10.5 Conclusions

The ability to reliably predict the 3D structure of a protein from its amino acid sequence has potentially far-reaching consequences in many scientific fields. This is demonstrated by the rapidly growing interest in AlphaFold2 ever since the publication of the initial article [3]. It has the potential to revolutionize our understanding of biology, allowing us to derive function from structure; predict protein variants/mutations; design new proteins [85]; study the evolution of proteins and the origins of life. In traditional drug discovery, the availability of high-quality computational models, usually augmented with experimental data, has already made a big impact. However, uncertainty about the accuracy of the predictions in active sites and the inability to define the conformational state of a protein remain key limitations [92]. In addition, AlphaFold models in combination with other related methods help in enabling pocket prediction, binding site comparison for drug repurposing, off-target predictions, ligandability assessment, engineering protein surfaces for protein crystallization, protein design, and protein–protein docking.

The availability of the AlphaFold Protein Structure Database by DeepMind and the ESM Metagenomic Structure Atlas by Meta-AI as openly accessible, extensive repositories [12, 104], as well as the implementation of ColabFold [105], could support a plethora of projects, including rare diseases research programs [106]. Rare diseases in particular, are often overlooked by research investors mainly because of unfavorable costs/patient ratios, might significantly benefit from such an approach [107]. Furthermore, models coming from AlphaFold and RosettaFold are now considered trusted external resources/data content and are fully integrated with PDB data [108].

Machine learning-based fold predictions are a game changer for structural bioinformatics and experimentalists alike, with exciting possibilities ahead [109]. In the field of drug discovery, the jury on the impact of AlphaFold2 and related methods is still out. However, there is no doubt that those methods have opened up a myriad of new avenues for exciting research and have brightened up the outlook for the future of drug discovery.

References

- 1 Dill, K.A. and MacCallum, J.L. (2012). The protein-folding problem, 50 years on. *Science* 338 (6110): 1042–1046. <https://doi.org/10.1126/science.1219021>.
- 2 Baek, M. et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 373 (6557): 871–876. <https://doi.org/10.1126/science.abj8754>.
- 3 Jumper, J. et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596 (7873): 583–589. <https://doi.org/10.1038/s41586-021-03819-2>.

- 4 Yang, J. et al. (2020). Improved protein structure prediction using predicted interresidue orientations. *Proc. Natl. Acad. Sci. U. S. A.* 117 (3): 1496–1503. <https://doi.org/10.1073/pnas.1914677117>.
- 5 Su, H. et al. (2021). Improved protein structure prediction using a new multi-scale network and homologous templates. *Adv Sci (Weinh)* 8 (24): e2102592. <https://doi.org/10.1002/advs.202102592>.
- 6 Lin, Z. et al. (2022). Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv* doi: 10.1101/2022.07.20.500902.
- 7 Chowdhury, R. et al. (2021). Single-sequence protein structure prediction using language models from deep learning. *bioRxiv* doi: 10.1101/2021.08.02.454840.
- 8 Kryzhtafovych, A. et al. (2021). Critical assessment of methods of protein structure prediction (CASP)—round XIV. *Proteins: Structure, Function, and Bioinformatics* 89 (12): 1607–1617. <https://doi.org/10.1002/prot.26237>.
- 9 Burley, S.K. et al. (2019). RCSB protein data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res.* 47 (D1): D464–D474. <https://doi.org/10.1093/nar/gky1004>.
- 10 Goodsell, D.S. et al. (2020). RCSB protein data Bank: enabling biomedical research and drug discovery. *Protein Sci.* 29 (1): 52–65. <https://doi.org/10.1002/pro.3730>.
- 11 Available from: <https://github.com/deepmind/alphafold>.
- 12 Varadi, M. et al. (2022). AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* 50 (D1): D439–D444. <https://doi.org/10.1093/nar/gkab1061>.
- 13 Xu, T. et al. (2022). Induced-fit docking enables accurate free energy perturbation calculations in homology models. *J. Chem. Theory Comput.* 18 (9): 5710–5724. <https://doi.org/10.1021/acs.jctc.2c00371>.
- 14 Masrati, G. et al. (2021). Integrative structural biology in the era of accurate structure prediction. *J. Mol. Biol.* 433 (20): 167127. <https://doi.org/10.1016/j.jmb.2021.167127>.
- 15 Van Drie, J.H. and Tong, L. (2020). Cryo-EM as a powerful tool for drug discovery. *Bioorg. Med. Chem. Lett.* 30 (22): 127524. <https://doi.org/10.1016/j.bmcl.2020.127524>.
- 16 Scapin, G., Potter, C.S., and Carragher, B. (2018). Cryo-EM for small molecules discovery, design, understanding, and application. *Cell. Chem. Biol.* 25 (11): 1318–1325. <https://doi.org/10.1016/j.chembiol.2018.07.006>.
- 17 de Oliveira, T.M. et al. (2021). Cryo-EM: the resolution revolution and drug discovery. *SLAS Discov* 26 (1): 17–31. <https://doi.org/10.1177/2472555220960401>.
- 18 Renaud, J.P. et al. (2018). Cryo-EM in drug discovery: achievements, limitations and prospects. *Nat. Rev. Drug Discov.* 17 (7): 471–492. <https://doi.org/10.1038/nrd.2018.77>.
- 19 Topf, M. et al. (2008). Protein structure fitting and refinement guided by Cryo-EM density. *Structure* 16 (2): 295–307. <https://doi.org/10.1016/j.str.2007.11.016>.

- 20 Palmer, C.M. and Aylett, C.H.S. (2022). Real space in cryo-EM: the future is local. *Acta Crystallogr. D Struct. Biol.* 78 (Pt. 2): 136–143. <https://doi.org/10.1107/S2059798321012286>.
- 21 Emsley, P. et al. (2010). Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* 66 (Pt 4): 486–501. <https://doi.org/10.1107/S0907444910007493>.
- 22 Kang, Y. and Chen, L. (2022). Structure and mechanism of NALCN-FAM155A-UNC79-UNC80 channel complex. *Nat. Commun.* 13 (1): 2639. <https://doi.org/10.1038/s41467-022-30403-7>.
- 23 Liebschner, D. et al. (2019). Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in phenix. *Acta Crystallogr. D Struct. Biol.* 75 (Pt 10): 861–877. <https://doi.org/10.1107/S2059798319011471>.
- 24 Qiao, Z. et al. (2022). Cryo-EM structure of the entire FtsH-HflKC AAA protease complex. *Cell Rep.* 39 (9): 110890. <https://doi.org/10.1016/j.celrep.2022.110890>.
- 25 Tai, L. et al. (2022). 8 a structure of the outer rings of the *Xenopus laevis* nuclear pore complex obtained by cryo-EM and AI. *Protein Cell* <https://doi.org/10.1007/s13238-021-00895-y>.
- 26 Evans, R. et al. (2022). Protein complex prediction with AlphaFold-Multimer. *bioRxiv* <https://doi.org/10.1101/2021.10.04.463034>.
- 27 Xu, Z. et al. (2022). Pentameric assembly of the Kv2.1 tetramerization domain. *Acta Crystallogr. D Struct. Biol.* 78 (Pt 6): 792–802. <https://doi.org/10.1107/S205979832200568X>.
- 28 Yuan, Y. et al. (2022). Cryo-EM structure of human glucose transporter GLUT4. *Nat. Commun.* 13 (1): 2671. <https://doi.org/10.1038/s41467-022-30235-5>.
- 29 Sanchez-Garcia, R. et al. (2021). DeepEMhancer: a deep learning solution for cryo-EM volume post-processing. *Commun. Biol.* 4 (1): 874. <https://doi.org/10.1038/s42003-021-02399-1>.
- 30 Kim, H. et al. (2022). Structural basis for assembly and disassembly of the IGF/IGFBP/ALS ternary complex. *Nat. I.D.A.A. Commun.* 13 (1): <https://doi.org/10.1038/s41467-022-32214-2>.
- 31 Foucher, A.-E. et al. (2022). Structural analysis of Red1 as a conserved scaffold of the RNA-targeting MTREC/PAXT complex. *Nat. I.D.A.A. Commun.* 13 (1): <https://doi.org/10.1038/s41467-022-32542-3>.
- 32 Oeffner, R.D. et al. (2022). Putting AlphaFold models to work with phenix.process_predicted_model and ISOLDE. *Acta Crystallographica Section D* 78 (11): 1303–1314. <https://doi.org/10.1107/S2059798322010026>.
- 33 Hryc, C.F. and Baker, M.L. (2022). AlphaFold2 and CryoEM: revisiting CryoEM modeling in near-atomic resolution density maps. *iScience* <https://doi.org/10.1016/j.isci.2022.104496>.
- 34 Kryshtafovych, A. et al. (2021). Computational models in the service of X-ray and cryo-electron microscopy structure determination. *Proteins* 89 (12): 1633–1646. <https://doi.org/10.1002/prot.26223>.
- 35 Jumper, J. et al. (2021). Applying and improving AlphaFold at CASP14. *Proteins: Structure, Function, and Bioinformatics* 89 (12): 1711–1721. <https://doi.org/10.1002/prot.26257>.

- 36 Dias, D.M. and Ciulli, A. (2014). NMR approaches in structure-based lead discovery: recent developments and new frontiers for targeting multi-protein complexes. *Prog. Biophys. Mol. Biol.* 116 (2–3): 101–112. <https://doi.org/10.1016/j.pbiomolbio.2014.08.012>.
- 37 Sugiki, T. et al. (2018). Current NMR techniques for structure-based drug discovery. *Molecules* 23 (1): <https://doi.org/10.3390/molecules23010148>.
- 38 Zweckstetter, M. (2021). NMR hawk-eyed view of AlphaFold2 structures. *Protein Sci.* 30 (11): 2333–2337. <https://doi.org/10.1002/pro.4175>.
- 39 Robertson, A.J. et al. (2021). Concordance of X-ray and AlphaFold2 models of SARS-CoV-2 main protease with residual dipolar couplings measured in solution. *J. Am. Chem. Soc.* 143 (46): 19306–19310. <https://doi.org/10.1021/jacs.1c10588>.
- 40 Fowler, N.J., Sljoka, A., and Williamson, M.P. (2020). A method for validating the accuracy of NMR protein structures. *Nat. Commun.* 11 (1): 6321. <https://doi.org/10.1038/s41467-020-20177-1>.
- 41 Fowler, N.J. and Williamson, M.P. (2022). The accuracy of protein structures in solution determined by AlphaFold and NMR. *Structure* <https://doi.org/10.1016/j.str.2022.04.005>.
- 42 Tejero, R. et al. (2022). AlphaFold models of small proteins rival the accuracy of solution NMR structures. *Front. Mol. Biosci.* 9: 877000. <https://doi.org/10.3389/fmolb.2022.877000>.
- 43 Leney, A.C. and Heck, A.J. (2017). Native mass spectrometry: what is in the name? *J. Am. Soc. Mass Spectrom.* 28 (1): 5–13. <https://doi.org/10.1007/s13361-016-1545-3>.
- 44 Heck, A.J. (2008). Native mass spectrometry: a bridge between interactomics and structural biology. *Nat. Methods* 5 (11): 927–933. <https://doi.org/10.1038/nmeth.1265>.
- 45 Allison, T.M. et al. (2022). Complementing machine learning-based structure predictions with native mass spectrometry. *Protein Sci.* 31 (6): e4333. <https://doi.org/10.1002/pro.4333>.
- 46 Rohl, C.A. et al. (2004). Protein structure prediction using Rosetta. In: *Methods in Enzymology*, 66–93. Academic Press.
- 47 Sala, D. et al. (2022). Modeling of protein conformational changes with Rosetta guided by limited experimental data. *Structure* <https://doi.org/10.1016/j.str.2022.04.013>.
- 48 del Alamo, D. et al. (2022). Integrated AlphaFold2 and DEER investigation of the conformational dynamics of a pH-dependent APC antiporter. *Proc. Natl. Acad. Sci. U S A* 119 (34): e2206129119. <https://doi.org/10.1073/pnas.2206129119>.
- 49 del Alamo, D. et al. (2022). Sampling alternative conformational states of transporters and receptors with AlphaFold2. *eLife* 11: e75751. <https://doi.org/10.7554/eLife.75751>.
- 50 Edich, M. et al. (2022). The impact of AlphaFold on experimental structure solution. *bioRxiv* doi: 10.1101/2022.04.07.487522.

- 51 Derewenda, Z.S. and Vekilov, P.G. (2006). Entropy and surface engineering in protein crystallization. *Acta Crystallogr. Sec. D* 62 (1): 116–124. <https://doi.org/10.1107/S0907444905035237>.
- 52 Manfredi, M. et al. (2021). DeepREx-WS: a web server for characterising protein–solvent interaction starting from sequence. *Comput. Struct. Biotechnol. J.* 19: 5791–5799. <https://doi.org/10.1016/j.csbj.2021.10.016>.
- 53 Terwilliger, T.C. et al. (2022). AlphaFold predictions: great hypotheses but no match for experiment. *bioRxiv* doi: 10.1101/2022.11.21.517405.
- 54 Mulligan, V.K. (2021). Current directions in combining simulation-based macro-molecular modeling approaches with deep learning. *Expert Opin. Drug Discov.* 16 (9): 1025–1044. <https://doi.org/10.1080/17460441.2021.1918097>.
- 55 Heo, L., Janson, G., and Feig, M. (2021). Physics-based protein structure refinement in the era of artificial intelligence. *Proteins* 89 (12): 1870–1887. <https://doi.org/10.1002/prot.26161>.
- 56 Heo, L. and Feig, M. (2020). High-accuracy protein structures by combining machine-learning with physics-based refinement. *Proteins* 88 (5): 637–642. <https://doi.org/10.1002/prot.25847>.
- 57 Steinegger, M. et al. (2019). HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinform.* 20 (1): 473. <https://doi.org/10.1186/s12859-019-3019-7>.
- 58 Jorgensen, W.L. et al. (1983). Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* 79 (2): 926–935. <https://doi.org/10.1063/1.445869>.
- 59 Kryshtafovych, A. et al. (2018). Evaluation of the template-based modeling in CASP12. *Proteins: Struct. Funct. Bioinform.* 86 (S1): 321–334. <https://doi.org/10.1002/prot.25425>.
- 60 Zhang, Y. et al. (2022). Benchmarking refined and unrefined AlphaFold2 structures for hit discovery. *ChemRxiv* <https://doi.org/10.26434/chemrxiv-2022-kcn0d-v2>.
- 61 Mysinger, M.M. et al. (2012). Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J. Med. Chem.* 55 (14): 6582–6594. <https://doi.org/10.1021/jm300687e>.
- 62 Scardino, V., Di Filippo, J.I., and Cavasotto, C.N. (2022). How good are AlphaFold models for docking-based virtual screening? *iScience* 105920. <https://doi.org/10.1016/j.isci.2022.105920>.
- 63 Miller, E.B. et al. (2021). Reliable and accurate solution to the induced fit docking problem for protein–ligand binding. *J. Chem. Theory Comput.* 17 (4): 2630–2639. <https://doi.org/10.1021/acs.jctc.1c00136>.
- 64 Wang, L. et al. (2015). Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *J. Am. Chem. Soc.* 137 (7): 2695–2703. <https://doi.org/10.1021/ja512751q>.
- 65 Beuming, T. et al. (2022). Are deep learning structural models sufficiently accurate for free-energy calculations? Application of FEP+ to AlphaFold2-predicted

- structures. *J. Chem. Inf. Model.* 62 (18): 4351–4360. <https://doi.org/10.1021/acs.jcim.2c00796>.
- 66 Steinegger, M., Mirdita, M., and Soding, J. (2019). Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nat. Methods* 16 (7): 603–606. <https://doi.org/10.1038/s41592-019-0437-4>.
 - 67 Mitchell, A.L. et al. (2020). MGnify: the microbiome analysis resource in 2020. *Nucl. Acids Res.* 48 (D1): D570–D578. <https://doi.org/10.1093/nar/gkz1035>.
 - 68 Suzek, B.E. et al. (2007). UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* 23 (10): 1282–1288. <https://doi.org/10.1093/bioinformatics/btm098>.
 - 69 Audagnotto, M. et al. (2022). Machine learning/molecular dynamic protein structure prediction approach to investigate the protein conformational ensemble. *Sci. Rep.* 12 (1): <https://doi.org/10.1038/s41598-022-13714-z>.
 - 70 Zhang, C. et al. (2020). DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. *Bioinformatics* 36 (7): 2105–2112. <https://doi.org/10.1093/bioinformatics/btz863>.
 - 71 Davtyan, A. et al. (2012). AWSEM-MD: protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing. *J. Phys. Chem. B.* 116 (29): 8494–8503. <https://doi.org/10.1021/jp212541y>.
 - 72 Schlessinger, A. and Bonomi, M. (2022). Exploring the conformational diversity of proteins. *eLife* 11: e78549. <https://doi.org/10.7554/eLife.78549>.
 - 73 Steinegger, M. and Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* 35 (11): 1026–1028. <https://doi.org/10.1038/nbt.3988>.
 - 74 He, X.H. et al. (2022). AlphaFold2 versus experimental structures: evaluation on G protein-coupled receptors. *Acta Pharmacol. Sin.* <https://doi.org/10.1038/s41401-022-00938-y>.
 - 75 Nicoli, A. et al. (2022). Modeling the orthosteric binding site of the G protein-coupled odorant receptor OR5K1. *bioRxiv* doi: 10.1101/2022.06.01.494157.
 - 76 Heo, L. and Feig, M. (2022). Multi-state modeling of G-protein coupled receptors at experimental accuracy. *Proteins: Struct. Funct. Bioinform.* n/a(n/a) <https://doi.org/10.1002/prot.26382>.
 - 77 Jakubec, D. et al. (2022). PrankWeb 3: accelerated ligand-binding site predictions for experimental and modelled protein structures. *Nucl. Acids Res.* <https://doi.org/10.1093/nar/gkac389>.
 - 78 Wang, S. et al. (2022). CavitySpace: a database of potential ligand binding sites in the human proteome. *Biomolecules* 12 (7): <https://doi.org/10.3390/biom12070967>.
 - 79 Yuan, Q. et al. (2022). AlphaFold2-aware protein-DNA binding site prediction using graph transformer. *Brief. Bioinform.* 23 (2): <https://doi.org/10.1093/bib/bbab564>.
 - 80 Konc, J. and Janežič, D. (2022). ProBiS-fold approach for annotation of human structures from the AlphaFold database with no corresponding structure in the

- PDB to discover new druggable binding sites. *J. Chem. Inf. Model.* <https://doi.org/10.1021/acs.jcim.2c00947>.
- 81 Trudeau, S.J. et al. (2022). PrePCI: A structure- and chemical similarity-informed database of predicted protein compound interactions. *bioRxiv* doi: 10.1101/2022.09.17.508184.
 - 82 Li, F. et al. (2022). DrugMAP: molecular atlas and pharma-information of all drugs. *Nucl. Acids Res.* <https://doi.org/10.1093/nar/gkac813>.
 - 83 White, M.E.H., Gil, J., and Tate, E.W. (2022). Proteome-wide structure-based accessibility analysis of ligandable and detectable cysteines in chemoproteomic datasets. *bioRxiv* doi: 10.1101/2022.12.12.518491.
 - 84 Wicky, B.I.M. et al. (2022). Hallucinating symmetric protein assemblies. *Science* 378 (6615): 56–61. <https://doi.org/10.1126/science.add1964>.
 - 85 Anishchenko, I. et al. (2021). De novo protein design by deep network hallucination. *Nature* 600 (7889): 547–552. <https://doi.org/10.1038/s41586-021-04184-w>.
 - 86 Ding, W., Nakai, K., and Gong, H. (2022). Protein design via deep learning. *Brief. Bioinform.* 23 (3): <https://doi.org/10.1093/bib/bbac102>.
 - 87 Pozzati, G. et al. (2021). Limits and potential of combined folding and docking. *Bioinformatics* <https://doi.org/10.1093/bioinformatics/btab760>.
 - 88 Bryant, P., Pozzati, G., and Elofsson, A. (2022). Improved prediction of protein-protein interactions using AlphaFold2. *Nat. Commun.* 13 (1): 1265. <https://doi.org/10.1038/s41467-022-28865-w>.
 - 89 Gao, M. et al. (2022). AF2Complex predicts direct physical interactions in multimeric proteins with deep learning. *Nat. Commun.* 13 (1): 1744. <https://doi.org/10.1038/s41467-022-29394-2>.
 - 90 Gao, M., Nakajima An, D., and Skolnick, J. (2022). Deep learning-driven insights into super protein complexes for outer membrane protein biogenesis in bacteria. *Elife* 11. <https://doi.org/10.7554/eLife.82885>.
 - 91 Schlick, T. and Portillo-Ledesma, S. (2021). Biomolecular modeling thrives in the age of technology. *Nat. Comput. Sci.* 1 (5): 321–331. <https://doi.org/10.1038/s43588-021-00060-9>.
 - 92 Mullard, A. (2021). What does AlphaFold mean for drug discovery? *Nat. Rev. Drug Discov.* 20 (10): 725–727. <https://doi.org/10.1038/d41573-021-00161-0>.
 - 93 Hekkelman, M.L. et al. (2021). AlphaFill: enriching the AlphaFold models with ligands and co-factors. *bioRxiv* doi: 10.1101/2021.11.26.470110.
 - 94 Hekkelman, M.L. et al. (2022). AlphaFill: enriching AlphaFold models with ligands and cofactors. *Nat. Methods* <https://doi.org/10.1038/s41592-022-01685-y>.
 - 95 Wilson, C.J., Choy, W.Y., and Karttunen, M. (2022). AlphaFold2: a role for disordered protein/region prediction? *Int. J. Mol. Sci.* 23 (9): <https://doi.org/10.3390/ijms23094591>.
 - 96 Strodel, B. (2021). Energy landscapes of protein aggregation and conformation switching in intrinsically disordered proteins. *J. Mol. Biol.* 433 (20): 167182. <https://doi.org/10.1016/j.jmb.2021.167182>.
 - 97 Lee, C., Su, B.H., and Tseng, Y.J. (2022). Comparative studies of AlphaFold, RoseTTAFold and Modeller: a case study involving the use of G-protein-coupled receptors. *Brief. Bioinform.* <https://doi.org/10.1093/bib/bbac308>.

- 98 Chakravarty, D. and Porter, L.L. (2022). AlphaFold2 fails to predict protein fold switching. *Protein Sci.* 31 (6): e4353. <https://doi.org/10.1002/pro.4353>.
- 99 Bagdonas, H. et al. (2021). The case for post-predictional modifications in the AlphaFold protein structure database. *Nat. Struct. Mol. Biol.* 28 (11): 869–870. <https://doi.org/10.1038/s41594-021-00680-9>.
- 100 Outeiral, C., Nissley, D.A., and Deane, C.M. (2022). Current structure predictors are not learning the physics of protein folding. *Bioinformatics* <https://doi.org/10.1093/bioinformatics/btab881>.
- 101 Chen, S.J. et al. (2023). Opinion: protein folds vs. protein folding: differing questions, different challenges. *Proc. Natl. Acad. Sci. U. S. A.* 120 (1): e2214423119. <https://doi.org/10.1073/pnas.2214423119>.
- 102 Sen, N. et al. (2022). Characterizing and explaining the impact of disease-associated mutations in proteins without known structures or structural homologs. *Brief. Bioinform.* <https://doi.org/10.1093/bib/bbac187>.
- 103 Buel, G.R. and Walters, K.J. (2022). Can AlphaFold2 predict the impact of missense mutations on structure? *Nat. Struct. Mol. Biol.* 29 (1): 1–2. <https://doi.org/10.1038/s41594-021-00714-2>.
- 104 David, A. et al. (2022). The AlphaFold database of protein structures: a Biologist's guide. *J. Mol. Biol.* 434 (2): 167336. <https://doi.org/10.1016/j.jmb.2021.167336>.
- 105 Mirdita, M. et al. (2022). ColabFold: making protein folding accessible to all. *Nat. Methods* <https://doi.org/10.1038/s41592-022-01488-1>.
- 106 Ros-Lucas, A. et al. (2022). The use of AlphaFold for in silico exploration of drug targets in the parasite *Trypanosoma cruzi*. *Frontiers in cellular and infection. Microbiology* 12. <https://doi.org/10.3389/fcimb.2022.944748>.
- 107 Rossi Sebastiano, M. et al. (2021). AI-based protein structure databases have the potential to accelerate rare diseases research: AlphaFoldDB and the case of IAHSF/Alsin. *Drug Discov. Today* <https://doi.org/10.1016/j.drudis.2021.12.018>.
- 108 Burley, S.K. et al. (2022). RCSB protein data Bank: tools for visualizing and understanding biological macromolecules in 3D. *Protein Sci.* e4482. <https://doi.org/10.1002/pro.4482>.
- 109 Edich, M. et al. (2022). The impact of AlphaFold on experimental structure solution. *Faraday Discuss.* <https://doi.org/10.1039/d2fd00072e>.

11

Deep Learning for the Structure-Based Binding Free Energy Prediction of Small Molecule Ligands

Venkatesh Mysore, Nilkanth Patel, and Adegoke Ojewole

NVIDIA Corporation, 2788 San Tomas Expy, Santa Clara, CA 95051, USA

11.1 Introduction

The prediction of the binding affinity between a ligand and a protein is one of the hardest and most important problems in computational drug discovery. These predictions are employed at all stages of the pipeline, from hit identification, to hit-to-lead development and lead optimization. Throughput and speed become relevant while screening catalogs of billions of compounds to find viable hits, when accuracy becomes paramount when prioritizing analogs of a hit to be synthesized. Ultimately, any virtual screening method is only as good as its scoring function, which serves as an estimate of the ligand-binding free energy. Understanding its assumptions and factoring in its limitations will enable the design of a robust workflow, informing key decisions such as the number of compounds to be screened and the number of hits to be followed up experimentally.

At a high level, the binding free energy prediction approaches can be classified as ligand-based and structure-based. In ligand-based approaches, the two- or three-dimensional constraints that a ligand must satisfy are used to mine a large database for compounds. These requirements may be derived from a known active compound, the characteristics of the binding pocket, or other relevant aspects known a priori. The structure-based approach, on the other hand, is characterized by reasoning about the docked three-dimensional protein–ligand complex. The ligand has multiple conformations accessible through its rotatable bonds, while the protein has multiple conformations that could expose multiple binding sites, each with varying degrees of flexibility afforded by the backbone and side-chain atoms of the amino-acid chain. For the structure-based approach to be viable, a reliable model of the protein, specifically the binding site, is required. Additionally, a reliable protocol for predicting the bound pose of the ligand is required, which automatically necessitates a reliable predictor of the protein–ligand binding free energy as the bound pose typically minimizes this quantity. It is important to explicitly state the assumptions of this approach. The protein's binding site is

presumed to be rigid, with alternate conformations typically handled by multiple rounds of docking and scoring to a representative ensemble. While ligand binding is also a dynamic process with the ligand sampling multiple poses and conformations even in its bound state, a single bound pose is usually what gets scored by the binding free energy estimator. Scores of multiple poses, potentially with multiple binding site conformations, can be used to account for the true nature of the binding process.

In this chapter, we focus on the use of deep learning in the structure-based prediction of binding free energy. We do not cover the precursor steps of protein structure prediction, binding site prediction, or bound pose prediction, which have been covered in some recent reviews [1]. We restrict our attention to the binding affinity prediction of an assembled complex, building on recent reviews on the topic [2–5].

Traditional approaches for addressing this problem have involved fitting a scoring function using known protein–ligand complex structures. Physics-inspired functional forms with pre-factors capturing the statistical distribution of protein–ligand atoms in complexes have proven to be remarkably useful in high-throughput virtual screening applications. However, such methods suffer from some inherent shortcomings: the functional form of the interaction terms has to be fixed ahead of time, and only experimentally determined co-crystal structures can be utilized for fitting. Further, the process of fitting and evaluating the scoring function involves substantial human inspection and ad hoc decision-making in many cases.

Deep-learning-based approaches have attempted to address these problems in the following ways. (1) A sufficiently deep and wide neural network can approximate arbitrary mathematical functions, so the functional form need not be hard-coded ahead of time. The model architecture, be it a 3D-convolutional network or graph convolutional one, can be conferred with as many parameters as necessary by increasing the number of features, nodes, or layers, enabling the complexity needed to capture the end-point. (2) The training procedure is less sensitive to low-quality training examples (with potentially incorrect poses) as long as there are a very large number of examples. Docked poses predicted by one or more methods can be used to expand the training set, increasing the number of training examples by several orders of magnitude. (3) Well-established procedures exist for the automated training and evaluation of models. Carefully splitting the available labeled data into training, validation, and test (or hold-out) can allow reliable and reproducible model development and unbiased evaluation. These three advantages have fueled the widespread development and application of deep learning for protein–ligand binding affinity prediction. The development of sophisticated software packages and libraries that encapsulate breakthrough technologies in deep learning has made it incredibly easy for researchers working on binding affinity to benefit from advances in image or object recognition, natural language processing, or network analysis. This has enabled rapid iterative experimentation, resulting in numerous successful applications of deep learning for structure-based binding affinity prediction. See Figure 11.1 for the approaches used to predict binding affinity.

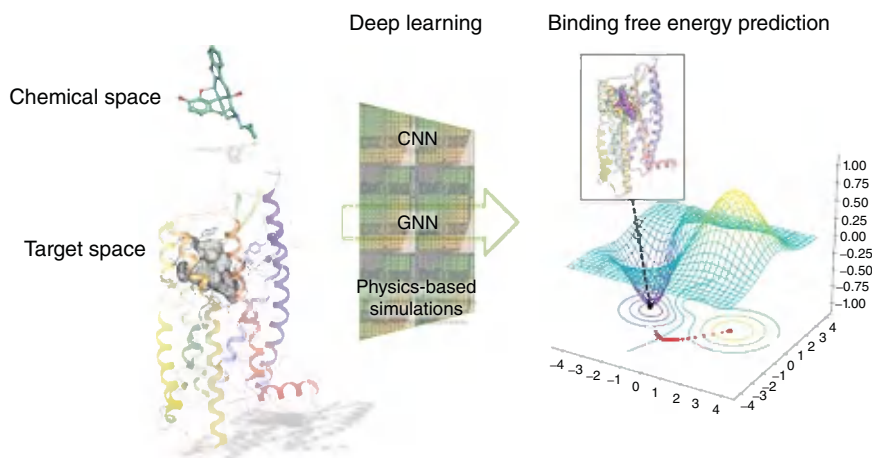


Figure 11.1 Schematic overview of the different models used for binding affinity prediction. Physics-based methods operate directly on ligand and target structures to predict binding affinity. Deep learning methods based on convolutional neural networks (CNNs) or graph neural networks (GNNs), for example, operate on compressed and learned representations of protein–ligand complexes to predict binding affinity.

In the following sections, we outline the current best practices in the application of these methods, providing a brief commentary on their evolution, limitations, and future.

11.2 Deep Learning Models for Reasoning About Protein–Ligand Complexes

The central challenge in adapting deep learning algorithms to the structure-based drug discovery domain is the representation of the protein–ligand complex in a format that can be digested as input by the deep learning framework. The protein–ligand interaction site features have to capture the 3-D structural and chemical aspects of the protein and ligand atoms involved. Specifically, it becomes crucial to capture the three-dimensional chemical neighborhood of a protein or ligand atom to facilitate the learning of a function that estimates binding free energy given the atoms of the protein–ligand complex. The decisions to be made center around whether explicit Cartesian coordinates, interatomic distances, or a coarser representation using grid points would capture this spatial context best. This representation of the input often dictates the model architecture capable of harnessing the locality information. An additional engineering challenge is incorporating information about the atoms, bonds, distances, and angles into the featurization of the atoms. In other words, all the information necessary to be able to make a reasonable prediction must be captured in the input representation and featurization. In this section, we go over the two broad classes of models – convolutional neural networks (CNNs) and graph neural networks (GNNs) – that have been successfully

employed to predict protein–ligand binding affinity. We quickly introduce the datasets before describing the models.

11.2.1 Datasets

Predicting protein–ligand binding affinities requires a training set comprising the ligand binding affinity values, such as K_i , K_d , and IC_{50} , against the target protein or a set of proteins. Several publicly available resources contain such information. For example, the Protein Data Bank and its derived datasets, PDBbind, BindingMOAD, BindingDB, and BindingDB subsets, are frequently used for CNN model development. Scientific community-driven standardized benchmarking initiatives, such as SAMPL, CASF, and CACHE, maintain datasets for model performance evaluations [6, 7]. While these datasets capture experimentally validated protein–ligand complexes, several data sources include experimental values of protein–ligand binding affinities without the 3-D details of the protein–ligand complexes (ChEMBL, PubChem, DUD-E, DOCKSTRING, ExCAPE, MoleculeNet, etc.). In the absence of experimental 3-D complex information, there are several approaches developed for utilizing such datasets for training a model. For example, it often includes employing computationally generated 3-D models of target proteins (using AlphaFold2 and RosettaFold) using protein sequence as an input in FASTA format, modeling the protein–ligand complex using docking (AutoDOCK), and more recently, using deep learning approaches for predicting the ligand binding sites in the proteins and predicting protein–ligand complexes [8–10].

11.2.2 Convolutional Neural Networks

11.2.2.1 Background

A CNN is a type of artificial neural network, mainly developed for computer vision tasks where the input features have locality information in their sequence. The reasoning to be performed should exploit the spatial proximity inherent in the input feature ordering. Identifying objects in an image would thus entail employing the same neural “filters” to smaller parts of the input. This is more parameter efficient than having a network where all input features had to interact with all others to enable the detection of some input object. As the depth of CNN layers increases, the model will be able to recognize increasingly complex objects.

11.2.2.2 Voxelized Grid Representation

Learnings from impactful CNN applications in computer vision were translated to tasks related to structure-based drug discovery by observing that the structure of the binding site and the compound could be treated as a three-dimensional grid of points where some are occupied by atoms [11]. The process of generating such features involved defining the ligand binding site as a 3-D box of a specific size and creating a grid within that box. An atom could be assigned to the nearest grid point or be spread out over a portion of the grid.

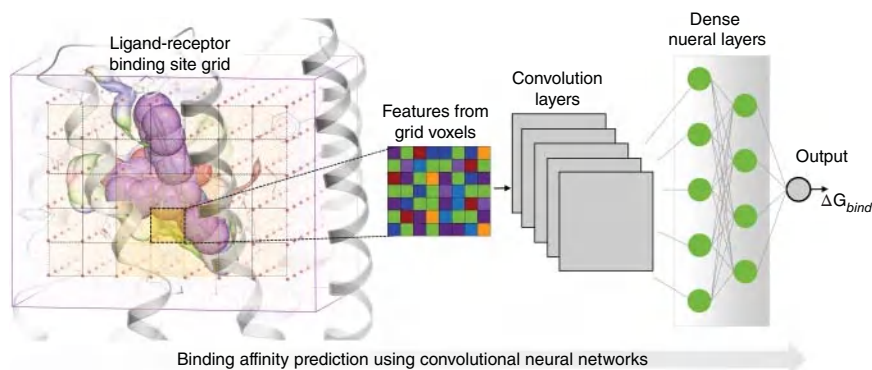


Figure 11.2 Featurization of a protein–ligand complex structure for binding affinity prediction using a convolutional neural network. CNN models for binding affinity can operate on 3-D tensors to capture the structural and physicochemical properties of a protein–ligand complex. The accuracy of 3-D CNNs is further enhanced by introducing random translations and rotations before applying convolutional filters.

While the initial CNN implementations for image recognition tasks used 2-D array representations of image segments as feature sets for input, advanced CNN models developed for drug discovery involve using 3-D tensors obtained from protein–ligand complexes (Figure 11.2). This representation also lends itself to data augmentation by random rotation and translation.

11.2.2.3 Descriptors

3D CNN models use various input features to represent the protein–ligand binding sites in a vectorized format similar to input tensors from images. These vectorized 3D grids of the binding site capture structural and/or physicochemical features. Such features include the presence or absence of atoms, atom types, and the electronic environment, including aromaticity, partial charge, protonation states, and atom interactions.

There are also approaches where the feature sets are generated from protein and ligand separately [12, 13]. Such feature sets include properties related to interaction surfaces – for example, solvent-accessible surface area, surface charge distribution, etc. Other surface representation approaches include utilizing and generating advanced geometric representations [14] and feeding them into a CNN model. Some recent models also have additional sources for feature calculations, such as the ions and solvents present or modeled at the ligand binding site and varying grid resolutions [15, 16].

11.2.2.4 Applications

Initial CNN-based models, such as AtomNet, GNINA, OnionNet, Kdeep, and MedusaNet, required an input protein–ligand binding pose to estimate the complex’s binding affinity. While these models show improved performance over the traditional molecular docking scoring programs and functions in the benchmark datasets, their ability is often limited by the docking step required to estimate the protein–ligand

binding pose. This step also impacts the throughput of the binding affinity prediction as the binding pose(s) generation step becomes the bottleneck compared to the interaction energy scoring step. A few models bypass the molecular docking step using amino acid sequences for proteins and SMILES format for ligands, followed by feature generation and binding affinity predictions. While bypassing the docking step may expedite the overall estimation process, it loses the input 3-D context of the interaction [6, 13, 15, 17–24].

11.2.3 Graph Neural Networks

11.2.3.1 Background

Graphs are represented by nodes (N) and edges (E) and a feature vector per node (f). Since the advent of social media and extensive data, graphs have become an important medium to capture and process such information efficiently. In the past decade, Neural Network architectures that operate on graph data structures have become helpful in drug discovery and molecular simulations.

11.2.3.2 Graph Representation

Protein–ligand interactions can be captured as a graph using various architectures. For example, a simple representation of a protein–ligand complex may have atoms from the binding pocket as nodes in the graph. In contrast, the edges represent the connectivity or vicinity of that atom. Graph representation captures the molecular features, along with its rotation and translation invariance features making it faster and more robust, while consuming much less memory than CNN.

11.2.3.3 Descriptors

While the choices for the atomic descriptors are the same as those for the CNNs, GNNs provide an additional avenue for encoding information about connectivity and neighborhood through the use of edge features. An edge between two atoms can carry information about the nature of the interaction (bonded or nonbonded), the distance, and other attributes. The graph convolutional operator aggregates information from all atoms connected to an atom; hence, the presence or absence of an edge can have a significant impact on the final prediction. The cut-off distance for deeming two atoms as interacting and the functional form of the distance-dependent term are some of the key design choices in a GNN. Unlike a CNN, whose input vector will be affected by rotation and translation of the complex, a GNN is invariant to these transformations as only internal distances are effectively used. Since there is no implicit order in which the edges of a node are to be traversed, chirality information is not automatically captured.

11.2.3.4 Applications

GNNs have been extensively applied to represent small molecules, for use in predicting compound properties and interactions. The first successful application of GNNs for protein–ligand binding affinity prediction was ?. Some of the initial GNN architectures for small molecules, such as PotentialNet, included features from the PLI

site with atoms and bonds from the protein–ligand complex. While other works, such as D-MPNN and graphDelta, involved features calculated by combining pre-trained fingerprints from QM datasets and fitting them into the extracted features from the PLI sites [25–27]. Cao and Shen [28] reported energy-based graph convolutional network for predicting intra- and inter-molecular interactions and related energies. Alternatively, Lim et al. [29] applied a mixed approach by using GAN for pose prediction and CNN for interaction-based scoring. Fusion models of 3D-CNN and GNN have shown better performance [30].

11.2.3.5 Extension to Attention Based Models

In the past few years, the field of NLP has seen a major shift with the advent of Transformer architecture, enabling the model to pay attention during self-supervised learning in order to capture long-range context. Following up on these advances, Liu et al. [14] employed the transformer architecture to develop intramolecular graph transformer (IGT) for binding pose and activity prediction. Similarly, Yuan et al. [31] developed a graph attention network with added custom mechanisms designed to capture bond features and node-level features. With the field of physics-informed ML gaining traction in recent years, Moon et al. [32] developed a model to incorporate physics-informed equations in addition to neural networks for predicting the protein–ligand binding affinity.

11.2.3.6 Geometric Deep Learning and Other Approaches

GNNs have started replacing 3D-CNNs in applications where rotational and translational invariance hold. Going beyond, a series of developments such as the SE(3)-transformer have led to more mathematically sophisticated models for protein structure prediction, binding pose prediction, and ultimately binding free energy. Of note is TANKBind from [33], which predicts both the bound pose and the binding affinity given a protein and a compound separately. Attempts have also been made to incorporate other mathematical concepts from geometry and topology to characterize the connectivity of atoms beyond graphs [34–36]; their wider adoption will aid their evaluation on real-world problems.

11.3 Deep Learning Approaches Around Molecular Dynamics Simulations

Structural biologists have long understood that the rigid body “lock-and-key” hypothesis to binding is an oversimplification and that modeling the flexibility of both the receptor and ligand is necessary to accurately predict binding affinity. From this understanding, molecular dynamics (MD) simulations have emerged as a powerful tool to investigate the dynamical properties of protein–ligand complexes. Despite the explosive growth in computational power in the last 30 years [37], long-time scale trajectories are required to probe the conformational changes that are relevant to binding under different conditions. Such MD simulations are computationally expensive and only apply to a small fraction of ligand–receptor pairs of

interest. In addition, classical molecular mechanics-based forcefields undersample biologically relevant high-energy transitions and binding states. Those states that are deeply sampled are explored in discrete steps. The theoretically correct approaches to computationally estimating ligand binding free energy typically employ explicit solvent MD simulations in some form [see other chapters on this topic].

Deep learning is beginning to permeate this sphere in multiple ways. Accurate force fields are essential for the MD simulations to be correct and for the quantities estimated to be realistic. This requires the computation of quantum mechanical potential energies for various ligand and peptide conformations. These computationally intensive calculations can now be performed using deep learning models trained on vast datasets of exact QM calculations. Going a step further, the neural networks developed to predict these potential energies can in fact drive the entire MD simulations. Machine learnt force-fields and their use in simulation are covered in a separate chapter. We highlight other important domains where deep learning is aiding MD simulation-based inference.

11.3.1 Enhanced Sampling

Approaches to marry deep learning and enhanced sampling are showing greater promise. Rufa et al. [38] describe a hybrid approach to perform alchemical free energy calculations correctly on a system where the ligand alone is simulated using neural potentials. Bertazzo et al. [39] describe a metadynamics-based approach for estimating binding free energy where machine learning is used to extract a free energy path from the initial steered-MD simulation.

11.3.2 Physics-inspired Neural Networks

Another family of methods incorporates force-field terms into their free energy prediction model. Hassan-Harrirou et al. [17] use the decomposition of MM/GBSA and other terms to featurize a voxelized 3D grid representing the ligand atoms. Guedes et al. [40] featurized a protein–ligand complex with force-field terms augmented with solvation, lipophilic, and ligand torsional terms, and then employed linear regression, random forests, and SVMs to predict the ligand binding free energy. Dong et al. [41] extract terms from MM / GBSA calculations, augmenting them with numerous descriptors before employing everything from linear regression to neural networks to predict free energy. Ji et al. [42] also employ MM / GBSA calculations and use the protein–ligand interaction energies as the featurization of the complex. Cain et al. [43] invoke the Generalized Born and Poisson–Boltzmann calculations for implicit solvent MD simulation, and propose a way of integrating those terms with a GNN representing the protein–ligand complex.

While machine-learnt force-fields concentrate on being able to replicate the potential energies employed in Hamiltonians of standard MD simulation force-fields, efforts to approximate the binding free energy directly using force-field-inspired neural networks have been underway in parallel. Zhu et al. [44] attempted to

develop a single neural network model that could predict the contribution to the binding free energy of a protein–ligand atom pair, given the force-field’s atom parameterization consisting of its partial charge and Lennard-Jones parameters. Cao and Shen [28] use a graph convolutional network modified to use operations inspired by the functional form of energy potentials. More recently, Moon et al. [32] used gated GNNs to learn atomic representations of a protein–ligand complex, subsequently feeding them pairwise into separate physics-inspired equations for different force-field terms.

11.3.3 Modeling Dynamics

Machine learning and deep learning are beginning to be evaluated for their ability to predict binding using conformational ensembles of protein–ligand complexes. These are based on approaches for predicting the course of an MD simulation, trying to reason over self-supervised representations of MD simulation frames. These exciting research areas are in the early stages of being adopted for industry-grade protein–ligand binding free energy prediction.

11.3.3.1 Applications

A variety of methods have been proposed to leverage protein–ligand conformational dynamics for binding affinity prediction. [45] began with 8888 initial candidate compounds and used a workflow consisting of physics-based flexible docking (Autodock Vina), followed by inference using the DeepBindBC ResNet binary classifier, and then a 100ns MD simulation step to predict 69 final candidate molecules. Of these, four were experimentally tested and shown to competitively inhibit TIPE2 (tumor necrosis factor- α -induced protein 8-like 2 protein) with μ M affinity for target PIP2. While the docking and MD steps explicitly computed dynamics, the deep learning step did not.

In contrast, Yasuda et al. [46] took a different approach predicated on the observation that binding affinity is associated with energy differences between the unbound and ligand-bound conformational ensembles. Their method begins with MD simulations of different ligands with a range of binding affinities for a specific binding pocket. These simulations, called the local dynamics ensemble (LDE), are defined as an ensemble of short-term trajectories of atoms of interest in the binding site. The authors used a multi-layer perceptron to compute the difference of LDE distributions between a ligand-bound and ligand-free system based on the Wasserstein Distance for all N pairs of bound and unbound systems. The resulting $N \times N$ matrix was embedded into points in a lower-dimensional space, and principal component analysis was performed on the embedded points. The first principal component was used as a proxy for ligand-binding energies.

Wang et al. [15] also begin by featurizing short-trajectory MD simulations. However, the authors trained random forest and LSTM-based models to predict the impact of active site point mutations on binding affinity. Structures of protein–ligand complexes were obtained from experimentally determined Platinum database and X-ray crystal structures (3Å or finer) from the PDB. Frames of nanosecond-scale

MD trajectories of wild-type and receptor mutants in complex with the ligands were featurized by the following descriptors: shape and topology, differences in estimated free energy, and local geometry (closeness, local surface area, orientation, contacts, and interfacial hydrogen bonds). The authors found that LSTM models trained on MD trajectories were more accurate than predictors based on energy estimation or descriptors alone.

Another approach is to predict binding affinity from an ensemble of protein–ligand structures not computed from an MD simulation. Intuitively, cross-docking a ligand to an ensemble of receptor conformations should provide a more comprehensive set of binding poses for structure-based virtual screening than single-pose docking does. However, two problems remain: 1) there is still no agreement on how to aggregate individual docking results into an ensemble docking score rank a ligand, and 2) ranking from traditional ensemble docking typically yields only modest improvement over docking to a rigid receptor. Ricci-Lopez et al. [47] propose to use ML to determine ensemble docking scores for four proteins: CDK2, FXa, EGFR, and HSP90. Receptor ensembles were prepared by docking ligands from standard libraries to crystal structures in the PDB. The authors used these ensembles to train binary classifiers using logistic regression and gradient-boosted decision trees and showed that these models significantly outperformed standard consensus docking at predicting binders and non-binders. Following a similar idea, Stafford et al. [48] propose AtomNet PoseRanker (ANPR), a graph convolutional network that predicts binding affinity from a collection of ligand poses. The input of ANPR is an ensemble of protein–ligand complexes computed from RosettaCM, which was used to sample low-energy structures in the vicinity of an input structure from PDBbind v2019. These structures are augmented with structures computed from docking the ligand to alternative structures of the same target protein. From this cross-docked data set, ANPR learns to recognize distinct ligand poses as valid in different receptor conformations. Ultimately, learning from ligand and receptor conformational diversity helps ANPR recognize a multitude of valid binding modes, improving ANPR’s binding predictions vs. Smina.

11.4 Modifying AlphaFold2 for Binding Affinity Prediction

AlphaFold2 [9] provided a step function improvement in protein structure prediction, generating unbound “apo” protein structures with near-experimental accuracy. This new capability created great interest among structural biologists. However, it is undetermined whether AlphaFold2 or similar models such as RosettaFold [8] can predict an arbitrary ligand binding site with accuracy sufficient to accommodate a variety of known binding partners. High accuracy in this task is necessary for computational drug discovery programs. The drug discovery community has, therefore, dedicated its efforts to determine how to modify AlphaFold2 to predict ligand binding.

11.4.1 Modifying AlphaFold2 Input Protein Database for Accurate Free Energy Predictions

Advances in classical force fields and sampling algorithms have made it possible to apply free energy perturbation (FEP) simulations to predict protein–ligand binding affinities with accuracies approaching those of biochemical and biophysical assays [49]. The improved accuracy of FEP methods shifts the burden of computational structure-based design to the availability of high-resolution protein–ligand complexes. It remains unanswered to what extent protein folds predicted by deep learning models such as AlphaFold2 can assist in free energy calculations. To partially address this knowledge gap, Beuming and colleagues [50] retrospectively substituted AlphaFold2 protein targets into Schrodinger FEP+ simulations [49] to reproduce previously validated studies. In this study, the AlphaFold2 database was modified to exclude template structures and sequences with more than 30% sequence identity to the query sequence at inference time. The modified model, called AF2₃₀, not only predicts protein active sites but also has substantial structural differences from the standard AlphaFold2 model. In a study limited to 14 protein targets, AF2₃₀ predicted relative changes in ligand affinities with an accuracy of 1.04 kcal/mol, which is statistically comparable to FEP+ calculations obtained using crystal structure complexes.

11.4.2 Modifying Multiple Sequence Alignment for AlphaFold2-Based Docking

An alternative is to modify AlphaFold2 to predict homo- and hetero-dimeric protein–protein interactions. The first approach to this end is AlphaFold-Multimer [51], an AlphaFold2 variant expressly trained on multimeric inputs. AlphaFold-Multimer incorporates additional changes to AlphaFold2, including modified losses to avoid superposing structures of different chains, to reduce steric clashes, and to account for permutation symmetry among homomeric chains. Other changes include paired multiple sequence alignments (MSAs) to capture interchain coevolutionary information and minor adjustments to the AlphaFold2 model architecture. On both small and large datasets of protein complexes, AlphaFold-Multimer exhibited superior accuracy to AlphaFold2 (input adapted with a flexible linker) per DockQ scores [52].

Bryant et al. [53] developed FoldDock, which uses a “fold and dock” approach with AlphaFold2 to predict binding between two different proteins each of length greater than 50 amino acids. A MSA is computed for the sequence of each binding partner, where each MSA is computed from a distinct database. A third MSA is computed by concatenating the first two MSAs. And a fourth “paired” MSA is constructed from the first two MSAs to determine inter-sequence coevolutionary information. In the FoldDock study, AlphaFold2 consumed these MSAs and predicted complexes with acceptable quality (DockQ ≥ 0.23) for 63% of 216 heterodimers. The predicted DockQ scores, used as a proxy for binding strength, also predicted interacting protein pairs from noninteracting pairs with a 1% false positive rate. However, FoldDock has

a greater empirical success rate with proteins from bacteria and archaea than those from eukaryotes and viruses.

An alternative approach by [54] is ColAttn. This method uses the MSA transformer to estimate a column attention score from the MSAs corresponding to the putative binding partners. AlphaFold-Multimer consumes this input to predict the complex structure. Overall, ColAttn may yield better complex structure prediction than AlphaFold-Multimer, particularly in eukaryotes.

11.5 Conclusion

11.5.1 New Models for Binding Affinity Prediction

Language-based models have recently been proposed to predict binding affinity. For example, Vielhaben et al. [55] use USMPep, an RNN architecture, to predict neopeptide binding affinity for Class 1 and Class 2 major histocompatibility complex (MHC) binding pockets. Vielhaben et al. [56] propose applying USMPep to predict viral peptides binding affinity to MHC. Cheng et al. [57] developed BERTMHC, a transformer-based multi-instance learning model to predict peptide-MHC Class 2 binding. Recently, Bachas et al. [58] proposed a BERT-style language model on antibody sequence data and binding affinity labels to quantitatively predict binding of unseen antibody sequence variants. Language-based approaches to binding affinity prediction promise to improve the productivity of early stages of drug discovery. Excitingly, transformer language models appear to improve in predictive performance on primary and downstream tasks as the architectures are scaled in parameter size from tens of millions to hundreds of billions [59]. This observation has not been tested for language models trained for binding affinity prediction. Doing so will require pretraining and fine-tuning large domain-specific models for DNA, RNA, proteins, and small molecules. Fortunately, recent advances in computing hardware, training frameworks, and inference frameworks have made the challenge tractable.

11.5.2 Retrospective from the Compute Industry

The foundational methods that have enabled the advancements in DL for binding affinity are GPU-accelerated algorithms from computational chemistry and popular machine learning architectures. The broad availability of application programming interfaces (APIs) such as CUDA [60] and OpenCL [61] has enabled these algorithms and architectures to leverage advances in GPU computing hardware. Over the last ten years, GPU acceleration has increased the size and complexity of MD-simulated systems by 100-fold [37]. The increased computational power afforded by GPUs has enabled more accurate ensemble computation in protein-ligand systems using density functional theory (DFT) [62] and hybrid quantum mechanics-molecular mechanics [63]. GPU-accelerated free-energy perturbation simulations [49] are routinely used to compute accurate relative binding affinities for biomedically and

industrially relevant protein–ligand systems. Machine-learned force-fields such as ANI [64] and AIMNet [65] have the potential to further accelerate these free energy simulations.

Graphics processing unit (GPU)-accelerated docking algorithms can also rapidly generate structural ensembles used to train DL models to predict binding affinity. Synthesizable molecules now number in the tens of billions, vastly expanding the scope of virtual screening and increasing the utility of massively parallel docking calculations [62]. For example, AutodockGPU was used to screen a billion molecules against a SARS-CoV-2 spike protein using the Summit supercomputer by parallelizing pose searching and scoring [66].

11.5.2.1 Future DL-Based Binding Affinity Computation will Require Massive Scalability

Despite impressive innovations in GPU architectures [handwiki.com], CPU codes that are amenable to GPU acceleration are precisely those that can be decomposed into a large number of thread blocks that are executed concurrently on GPU hardware. When optimization for a single GPU is insufficient to meet compute requirements, computation must either be scaled to multiple GPUs in the same compute node or distributed among multiple compute nodes with low-latency interconnects. Powerful yet simplified distributed training frameworks are necessary to make distributed supercomputing architectures routinely accessible for DL. The following discussion describes the emerging computational challenge and discusses the software and hardware frameworks that have been proposed to address the problem.

11.5.2.2 Single GPU Optimizations for DL

To take advantage of exponentially increasing datasets, the DL models for binding described in this chapter will need to radically expand in size. As a result, the GPU platforms and APIs that run these larger models must also increase in speed, complexity, and scalability. Owing to Moore's Law, individual GPUs will continue to scale in transistor count [handwiki.com], allowing for larger computational graphs to be mapped onto individual devices. GPUs will also be optimized at the hardware level to accelerate computational bottlenecks within DL architectures. For example, the recently announced Hopper GPU [67] will substantially accelerate transformer models by automatically altering tensor precision to maintain model accuracy while reaping the hardware performance benefits of representing data in smaller, faster numerical formats such as FP8.

11.5.2.3 Distributed DL Training and Inference

Data and Model Parallelism As data and model sizes have increased dramatically, it is common that DL models must be trained using multiple GPUs, which necessitates data or model parallelism. In the former case, different data batches, along with an identical set of model weights, are distributed to each GPU. Gradients are computed concurrently on each GPU and are returned to the main compute node, where they are aggregated and redistributed, along with new data batches, for the next iteration. Model parallelism is necessary when the model itself is so large that

it cannot be stored in the memory of a single GPU. In this case, the high memory requirement is managed by storing different layers of the model on different GPUs and transmitting the results of forward and backward propagation to the appropriate GPU(s). Efficient data and model parallelism require fast intra-node GPU interconnection technologies and fast internode networking, along with scalable, feature-rich, and user-friendly APIs such as Pytorch Lightning [68] and NeMo [69]. Inference frameworks such as Triton [70, 71] and associated SDKs such as TensorRT [72] are increasingly becoming necessary to deploy these large DL models with the requisite scalability and response time of a research and development environment.

Federated Learning A logical extension to data and model parallelism is Federated Learning (FL). FL frameworks, such as NVIDIA FLARE [73], apply the concepts of distributed learning to geographically dispersed sites. By design, FL has the potential to obviate the barriers to sharing proprietary data, leading to more accurate and robust models. In FL, participating sites perform localized training using their respective unshared data. Gradients and losses from these local models are subsequently aggregated using the previously described principles of data parallelism. And since no data leave their respective silos, the resulting global model can be shared beyond the individual sites.

References

- 1 Dhakal, A., McKay, C., Tanner, J.J., and Cheng, J. (2022). Artificial intelligence in the prediction of protein–ligand interactions: recent advances and future directions. *Brief. Bioinform.* 23 (1): bbab476. <https://doi.org/10.1093/bib/bbab476>.
- 2 Qin, T., Zhu, Z., Wang, X.S. et al. (2021). Computational representations of protein–ligand interfaces for structure-based virtual screening. *Expert Opin. Drug Discovery* 16 (10): 1175–1192. <https://doi.org/10.1080/17460441.2021.1929921>.
- 3 Anighoro, A. (2022). Deep learning in structure-based drug design. *Methods Mol. Biol.* 2390: 261–271. https://doi.org/10.1007/978-1-0716-1787-8_11. PMID: 34731473.
- 4 Kimber, T.B., Chen, Y., and Volkamer, A. (2021). Deep learning in virtual screening: recent applications and developments. *Int. J. Mol. Sci.* 22: 4435. <https://doi.org/10.3390/ijms22094435>.
- 5 Li, H., Sze, K.-H., Lu, G., and Ballester, P. (2020). Machine-learning scoring functions for structure-based drug lead optimization. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* 10: e1465. <https://doi.org/10.1002/wcms.1465>.
- 6 Zhang, H., Liao, L., Saravanan, K.M. et al. (2019). DeepBindRG: A deep learning based method for estimating effective protein–ligand affinity. *Peer J* 7: 2019. <https://doi.org/10.7717/peerj.7362>.
- 7 Ackloo, S., Al-Awar, R., Amaro, R.E. et al. (2022). CACHE (Critical Assessment of Computational Hit-finding Experiments): a public–private partnership benchmarking initiative to enable the development of computational methods for

- hit-finding. *Nature Rev. Chem.* 6 (4) Nature Research: 287–295. <https://doi.org/10.1038/s41570-022-00363-z>.
- 8 Baek, M., DiMaio, F., Anishchenko, I. et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science* (1979) 373 (6557): 871–876. <https://doi.org/10.1126/science.abj8754>.
 - 9 Jumper, J., Evans, R., Pritzel, A. et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596: 583. <https://doi.org/10.1038/s41586-021-03819-2>.
 - 10 Stärk, H., Ganea, O., Pattanaik, L. et al. (2022). Geometric deep learning for drug binding structure prediction. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 20503–20521.
 - 11 Wallach, I., Dzamba, M., and Heifets, A. (2015). AtomNet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. <http://arxiv.org/abs/1510.02855>.
 - 12 Ahmed, A., Mam, B., and Sowdhamini, R. (2021). DEELIG: a deep learning approach to predict protein-ligand binding affinity. *Bioinform. Biol. Insights* 15. <https://doi.org/10.1177/11779322211030364>.
 - 13 Jiménez, J., Škalič, M., Martínez-Rosell, G., and de Fabritiis, G. (2018). KDEEP: protein–ligand absolute binding affinity prediction via 3D-convolutional neural networks. *J. Chem. Inf. Model.* 58 (2): 287–296. <https://doi.org/10.1021/acs.jcim.7b00650>.
 - 14 Liu, Q., Wang, P.-S., Zhu, C. et al. (2021). OctSurf: Efficient hierarchical voxel-based molecular surface representation for protein-ligand affinity prediction. *J. Mol. Graph. Model.* 105: 107865. <https://doi.org/10.1016/j.jmgm.2021.107865>.
 - 15 Wang, Z., Zheng, L., Liu, Y. et al. (2021). OnionNet-2: a convolutional neural network model for predicting protein-ligand binding affinity based on residue-atom contacting shells. *Front. Chem.* 9. <https://doi.org/10.3389/fchem.2021.753002>.
 - 16 Gao, A. and Remsing, R.C. (2022). Self-consistent determination of long-range electrostatics in neural network potentials. *Nat. Commun.* 13 (1): 1–11. <https://doi.org/10.1038/s41467-022-29243-2>.
 - 17 Hassan-Harriour, H., Zhang, C., and Lemmin, T. (2020). RosENet: improving binding affinity prediction by leveraging molecular mechanics energies with an ensemble of 3D convolutional neural networks. *J. Chem. Inf. Model.* 60 (6): 2791–2802. <https://doi.org/10.1021/acs.jcim.0c00075>.
 - 18 Sunseri, J., King, J.E., Francoeur, P.G., and Koes, D.R. (2019). Convolutional neural network scoring and minimization in the D3R 2017 community challenge. *J. Comput. Aided Mol. Des.* 33 (1): 19–34. <https://doi.org/10.1007/s10822-018-0133-y>.
 - 19 Wu, Z., Ramsundar, B., Feinberg, E.N. et al. (2017). MoleculeNet: a benchmark for molecular machine learning. <http://arxiv.org/abs/1703.00564>.
 - 20 Cang, Z. and Wei, G. (2017). TopologyNet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions. *PLoS Comput. Biol.* 13 (7). <https://doi.org/10.1371/journal.pcbi.1005690>.

- 21 Stepniewska-Dziubinska, M.M., Zielenkiewicz, P., and Siedlecki, P. (2018). Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. *Bioinformatics* 34 (21): 3666–3674. <https://doi.org/10.1093/bioinformatics/bty374>.
- 22 McNutt, A.T., Francoeur, P., Aggarwal, R. et al. (2021). GNINA 1.0: molecular docking with deep learning. *J. Cheminform* 13: 43. <https://doi.org/10.1186/s13321-021-00522-2>.
- 23 Öztürk, H., Özgür, A., and Ozkirimli, E. (2018). DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics* 34 (17): i821–i829. <https://doi.org/10.1093/bioinformatics/bty593>.
- 24 Li, Y., Rezaei, M.A., Li, C., and Li, X. (2019). DeepAtom: a framework for protein–ligand binding affinity prediction. In: *Proceedings – 2019 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2019*, 303–310. <https://doi.org/10.1109/BIBM47256.2019.8982964>.
- 25 Yang, K., Swanson, K., Jin, W. et al. (2019). Analyzing learned molecular representations for property prediction. *J. Chem. Inf. Model.* 59: 3370–3388. <https://doi.org/10.1021/acs.jcim.9b00237>.
- 26 Feinberg, E.N., Sur, D., Wu, Z. et al. (2018). PotentialNet for molecular property prediction. *ACS Cent. Sci.* 4 (11): 1520–1530. <https://doi.org/10.1021/acscentsci.8b00507>.
- 27 Karlov, D.S., Sosnin, S., Fedorov, M.V., and Popov, P. (Mar. 2020). GraphDelta: MPNN scoring function for the affinity prediction of protein–ligand complexes. *ACS Omega* 5 (10): 5150–5159. <https://doi.org/10.1021/acsomega.9b04162>.
- 28 Cao, Y. and Shen, Y. (2020). Energy-based graph convolutional networks for scoring protein docking models. *Proteins: Struct. Function Bioinform.* 88. <https://doi.org/10.1002/prot.25888>.
- 29 Lim, J., Ryu, S., Park, K. et al. (2019). Predicting drug–target interaction using 3D structure-embedded graph representations from graph neural networks. <http://arxiv.org/abs/1904.08144>.
- 30 Knutson, C., Bontha, M., Bilbrey, J.A., and Kumar, N. (2022). Decoding the protein–ligand interactions using parallel graph neural networks. *Sci. Rep.* 12 (1). <https://doi.org/10.1038/s41598-022-10418-2>.
- 31 Yuan, H., Huang, J., and Li, J. (2021). Protein–ligand binding affinity prediction model based on graph attention network. *Math. Biosci. Eng.* 18 (6): 9148–9162. <https://doi.org/10.3934/mbe.2021451>.
- 32 Moon, S., Zhung, W., Yang, S. et al. (2022). PIGNet: a physics-informed deep learning model toward generalized drug–target interaction predictions. *Chem. Sci.* 13 (13): 3661–3673. <https://doi.org/10.1039/d1sc06946b>.
- 33 Lu, W., Wu, Q., Zhang, J. et al. (2022). TANKBind: trigonometry-aware neural networks for drug–protein binding structure prediction. *bioRxiv*. <https://doi.org/10.1101/2022.06.06.495043>.
- 34 Nguyen, D.Q., Nguyen, T.D., and Phung, D. (2019). Universal graph transformer self-attention networks. <https://arxiv.org/abs/1909.11855v1>
- 35 Meng, Z. and Xia, K. (2021). Persistent spectral–based machine learning (PerSpect ML) for protein–ligand binding affinity prediction. *Sci. Adv.* 7 (19). <https://doi.org/10.1126/sciadv.abc5329>.

- 36 Wee, J. and Xia, K. (2021). Ollivier persistent Ricci curvature-based machine learning for the protein–ligand binding affinity prediction. *J. Chem. Inf. and Model.* 61 (4): 1617–1626.
- 37 Pandey, M., Fernandez, M., Gentile, F. et al. (2022). The transformational role of GPU computing and deep learning in drug discovery. *Nature Mach. Intell.* 4 (3): 211–221.
- 38 Rufa, D.A., Bruce Macdonald, H.E., Fass, J. et al. (2020). Towards chemical accuracy for alchemical free energy calculations with hybrid physics-based machine learning/molecular mechanics potentials. *bioRxiv*. <https://doi.org/10.1101/2020.07.29.227959>.
- 39 Bertazzo, M., Gobbo, D., Decherchi, S., and Cavalli, A. (2021). Machine learning and enhanced sampling simulations for computing the potential of mean force and standard binding free energy. *J. Chem. Theory Comput.* 17 (8): 5287–5300. <https://doi.org/10.1021/acs.jctc.1c00177>.
- 40 Guedes, I.A., Barreto, A.M.S., Marinho, D. et al. (2021). New machine learning and physics-based scoring functions for drug discovery. *Sci. Rep.* 11: 3198. <https://doi.org/10.1038/s41598-021-82410-1>.
- 41 Dong, L., Qu, X., Zhao, Y., and Wang, B. (2021). Prediction of binding free energy of protein–ligand complexes with a hybrid molecular mechanics/generalized born surface area and machine learning method. *ACS Omega* 6 (48): 32938–32947. <https://doi.org/10.1021/acsomega.1c04996>.
- 42 Ji, B., He, X., Zhai, J. et al. (2021). Machine learning on ligand-residue interaction profiles to significantly improve binding affinity prediction. *Brief. Bioinf.* 22 (5). <https://doi.org/10.1093/bib/bbab054>.
- 43 Cain, S., Risheh, A., and Forouzes, N. (2022). A physics-guided neural network for predicting protein–ligand binding free energy: from host–guest systems to the PDBbind database. *Biomolecules* 12: 919. <https://doi.org/10.3390/biom12070919>.
- 44 Zhu, F., Zhang, X., Allen, J.E. et al. (2020). Binding affinity prediction by pairwise function based on neural network. *J. Chem. Inf. Model.* 60 (6): 2766–2772. <https://doi.org/10.1021/acs.jcim.0c00026>.
- 45 Zhang, H., Li, J., Saravanan, K.M. et al. (2021). An integrated deep learning and molecular dynamics simulation-based screening pipeline identifies inhibitors of a new cancer drug target TIPE2. *Front. Pharmacol.* 12.
- 46 Yasuda, I., Endo, K., Yamamoto, E. et al. (2022). Differences in ligand-induced protein dynamics extracted from an unsupervised deep learning approach correlate with protein–ligand binding affinities. *Commun. Biol.* 5 (1): 1–9.
- 47 Ricci-Lopez, J., Aguila, S.A., Gilson, M.K., and Brizuela, C.A. (2021). Improving structure-based virtual screening with ensemble docking and machine learning. *J. Chem. Inf. Model.* 61 (11): 5362–5376.
- 48 Stafford, K., Anderson, B.M., Sorenson, J., and van den Bedem, H. (2022). Atom-Net PoseRanker: enriching ligand pose quality for dynamic proteins in virtual high-throughput screens. *J. Chem. Inf. Model.* 62 (5): 1178–1189.
- 49 Abel, R., Wang, L., Harder, E.D. et al. (2017). Advancing drug discovery through enhanced free energy calculations. *Acc. Chem. Res.* 50: 1625–1632.
- 50 Beuming, T., Martín, H., Díaz-Rovira, A.M. et al. (2022). Are deep learning structural models sufficiently accurate for free-energy calculations? Application

- of FEP+ to AlphaFold2-predicted structures. *J. Chem. Inf. Model.* 62 (18): 4351–4360.
- 51 Evans, R., O'Neill, M., Pritzel, A. et al. (2021). Protein complex prediction with AlphaFold-Multimer. *BioRxiv*.
 - 52 Basu, S. and Wallner, B. (2016). DockQ: a quality measure for protein-protein docking models. *PLoS One* 11 (8): 1–9.
 - 53 Bryant, P., Pozzati, G., and Elofsson, A. (2022). Improved prediction of protein-protein interactions using AlphaFold2. *Nat. Commun.* 13: 1265.
 - 54 Chen, B., Xie, Z., Xu, J. et al. (2022). Improve the protein complex prediction with protein language models. *BioRxiv*.
 - 55 Vielhaben, J., Wenzel, M., Samek, W., and Strodtzoff, N. (2020). USMPep: universal sequence models for major histocompatibility complex binding affinity prediction. *BMC Bioinform.* 21 (1): 1–16.
 - 56 Vielhaben, J., Wenzel, M., Weicken, E., and Strodtzoff, N. (2021). Predicting the binding of SARS-CoV-2 peptides to the major histocompatibility complex with recurrent neural networks. *arXiv preprint arXiv:2104.08237*.
 - 57 Cheng, J., Bendjama, K., Rittner, K., and Malone, B. (2021). BERTMHC: improved MHC-peptide class II interaction prediction with transformer and multiple instance learning. *Bioinformatics* 37 (22): 4172–4179.
 - 58 Bachas, S., Rakocevic, G., Spencer, D. et al. (2022). Antibody optimization enabled by artificial intelligence predictions of binding affinity and naturalness. *bioRxiv*.
 - 59 Rae, J.W., Borgeaud, S., Cai, T. et al. (2021). Scaling language models: methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.
 - 60 Vingelmann, P. and Fitzek, F.H.P. (2020). CUDA release 10.2.89, NVIDIA.
 - 61 Stone, J.E., Gohara, D., and Shi, G. (2010). OpenCL: a parallel programming standard for heterogeneous computing systems. *Comput. Sci. Eng.* 12: 66–72.
 - 62 Grygorenko, O.O., Radchenko, D.S., Dziuba, I. et al. (2020). Generating multi-billion chemical space of readily accessible screening compounds. *iScience* 23: 101681.
 - 63 Yu, J.K., Liang, R., Liu, F., and Martínez, T.J. (2019). First-principles characterization of the elusive I fluorescent state and the structural evolution of retinal protonated Schiff base in bacteriorhodopsin. *J. Am. Chem. Soc.* 141: 18193–18203.
 - 64 Yoo, P., Sakano, M., Desai, S. et al. (2021). Neural network reactive force field for C, H, N, and O systems. *NPJ Comput. Mater.* 7: 9.
 - 65 Zubatyuk, R., Smith, J.S., Leszczynski, J., and Isayev, O. (2021). Accurate and transferable multitask prediction of chemical properties with an atoms-in-molecules neural network. *Sci. Adv.* 5, eaav6490.
 - 66 LeGrand, S., Scheinberg, A., Tillack, A.F. et al. (2020). GPU-accelerated drug discovery with docking on the summit supercomputer: porting, optimization, and application to COVID-19 research. *Proceedings of 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. <https://doi.org/10.1145/3388440.3412472>.
 - 67 Salvator, D. (2022). H100 Transformer Engine Supercharges AI Training, Delivering Up to 6x Higher Performance Without Losing Accuracy, NVIDIA.

- 68 Falcon, W. (2019). Pytorch lightning. GitHub. Note: <https://github.com/PyTorchLightning/pytorch-lightning> 3.6.
- 69 Kuchaiev, O., Li, J., Nguyen, H. et al. (2019). Nemo: a toolkit for building ai applications using neural modules. arXiv preprint arXiv:1909.09577.
- 70 Jahanshahi, A., Sabzi, H.Z., Lau, C., and Wong, D. (2020). Gpu-nest: Characterizing energy efficiency of multi-gpu inference servers. *IEEE Comput. Architect. Lett.* 19 (2): 139–142.
- 71 de Souza Pereira Moreira, G., Rabhi, S., Ak, R., and Schifferer, B. (2021). End-to-end session-based recommendation on GPU. In: *Fifteenth ACM Conference on Recommender Systems*, 831–833.
- 72 Jeong, E., Kim, J., Tan, S. et al. (2022). Deep learning inference parallelization on heterogeneous processors with TensorRT. *IEEE Embedd. Syst. Lett.* 14 (1): 15–18.
- 73 Han, W., Mawhirter, D., Wu, B. et al. (2019). FLARE: flexibly sharing commodity GPUs to enforce QoS and improve utilization. In: *International Workshop on Languages and Compilers for Parallel Computing*. Cham: Springer.

12

Using Artificial Intelligence for *de novo* Drug Design and Retrosynthesis

Rohit Arora¹, Nicolas Brosse², Clarisse Descamps², Nicolas Devaux²,
Nicolas Do Huu², Philippe Gendreau², Yann Gaston-Mathé², Maud Parrot²,
Quentin Perron², and Hamza Tajmouati²

¹Iktos Inc, 50 Milk St, Boston, MA 02110, USA

²Iktos SAS, 65 Rue de Prony, 75017, Paris, France

12.1 Introduction

12.1.1 Traditional Drug Design and Discovery Process Is Slow and Expensive

The drug discovery and development process is notably complex and iterative and is therefore time-consuming, arduous, and expensive. Typically, the process is initiated with the identification and validation of a potential therapeutic target (e.g. a protein) that is functionally implicated in a disease or multiple diseases [1]. Once the target has been identified and validated, the small molecules that can potentially interact with this target protein to either inhibit or activate its function (directly or otherwise) that has an impact on the disease state, must be identified. This launches the early stages of the drug design and discovery process (see Figure 12.1).

Initial hits against the biological target may be identified using a number of established methods, focusing primarily on compound activity against the target. These include (but are not limited to) high-throughput screening of chemical libraries, focused screening approaches, fragment-based drug design, virtual screening, and knowledge-based (both computer-based and medicinal chemistry-based from known chemical matter) drug design. Once hit molecules are identified, the process of identifying and prioritizing promising chemical series begins. Analogues of the original hit compounds may be synthesized and tested to investigate structure–activity relationships (SAR; physicochemical and absorption, distribution, metabolism, and excretion [ADME] properties in addition to activity) and identify lead compound series. Finally, lead optimization entails maintaining favorable activity, absorption, distribution, metabolism, excretion, and toxicity (ADMET) and physicochemical properties while tweaking the lead structure to ensure that the compound is successful in the downstream pre-clinical and clinical phases [2, 3].

Computational Drug Discovery: Methods and Applications, First Edition.

Edited by Vasanthanathan Poongavanam and Vijayan Ramaswamy.

© 2024 WILEY-VCH GmbH. Published 2024 by WILEY-VCH GmbH.

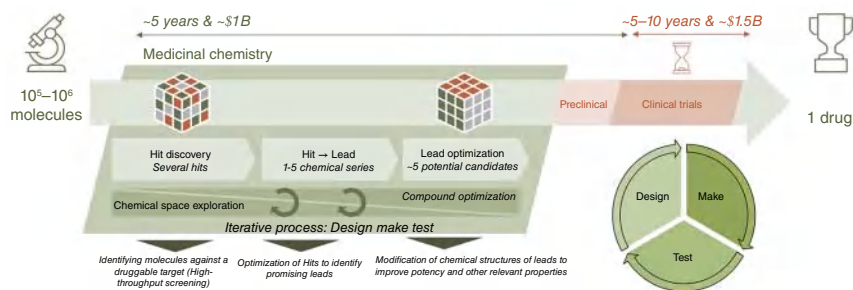


Figure 12.1 Early stages of drug design and discovery process.

The entire process follows the DMTA (Design-Make-Test-Analyze) iterative cycle at every phase to optimize the pharmacological and drug-like properties of potential drug candidates (Figure 12.1). Traditionally, in drug discovery campaigns, compound design and synthesis decisions are made with input from medicinal and synthetic chemistry teams and require multiple iterations. Owing to the iterative nature and challenges of each component of this cycle, bringing a new drug from the discovery stage to the market can take up to a decade and can cost well over a billion US dollars. This process can be accelerated by addressing the underlying challenges and streamlining the individual components of the DMTA cycle. This is a crucial issue that must be tackled at the discovery and optimization stages, before significant time and resources have been invested. Recently, some of the industry focus for *de novo* drug design [4] has shifted from lower-hanging fruits for chemical matter and toward using new modalities aimed toward drugging novel, first-in-class, hard-to-drug, or “undruggable” targets (e.g. PROTACs, molecular glues, RNA therapeutics, small-molecule modulators of protein–protein interactions, etc.) [5], that may go beyond Lipinski’s Rule of Five [6–8]. These new modalities introduce additional variables and therefore pose additional optimization challenges for each component of the DMTA cycle that are harder to tackle by medicinal chemists. It is essential to address them to accelerate the entire discovery process.

12.1.2 Success and Limitations of Standard Computational Methods

Over the last three decades or so, computational methods have been a mainstay of the drug discovery process and have helped in the development of a number of therapeutically relevant drug-like molecules. This is true for both small-molecule drugs and (more recently) peptide therapeutics [9, 10]. These methods include structure-based approaches like protein–ligand or protein–protein docking, virtual screening, molecular dynamics simulations, etc., and ligand-based approaches like quantitative structure–activity relationship (QSAR) and pharmacophore modeling.

From the perspective of *de novo* drug design, virtual screening of large libraries of molecules has been successfully deployed in a number of drug discovery campaigns [11]. Virtual screening involves either docking a large library of compounds to a known target (structure-based) or evaluating the similarity between compounds

in the large library and known actives against a target, especially in the absence of target structure information (ligand-based) [12]. The scale of these screening campaigns in terms of the compound library size has grown by orders of magnitude over the last few years – from $10^5 - 10^6$ to $10^8 - 10^9$ – in order to sample a larger chunk of the chemical space. This trend has been aided and accelerated by the advent of ultra-large chemical libraries of virtual compounds – e.g. Enamine REAL Space, Merck MASSIV library, GSKChemspace, and WuXi Apptec's GalaXi [13]. This has also encouraged development of a number of data analysis and predictive machine learning methods to manage these data, and to complete screening campaigns in a reasonable amount of time [14].

The impact of virtually screening ultra-large compound libraries has been significant – it has been shown to lead to a marked increase in the enrichment factor [15, 16]. This, however, does not address the inherent limitations of such methods. For *de novo* drug design, screening campaigns are limited to what is available in these compound libraries. Even the largest libraries ($\approx 10^9$ or larger) represent a fraction of the total size of the drug-like chemical space, which is estimated to be on the order of 10^{60} [17], and may not represent the overall diversity of this space. This drastically reduces the probability of finding the optimal compound for the target in question. Importantly, this approach does not address the issue of multi-parametric optimization (MPO) where the optimal drug candidate must be optimized across multiple objectives (solubility, potency, permeability, etc.). Prior knowledge of known actives can certainly improve this likelihood, but even with predictive methods, this problem can be fairly intractable. Screening methods – both ligand-based and structure-based – can evaluate existing ideas but cannot propose novel ideas. Put differently, these approaches can tell you what not to do, but they cannot tell you what to do. This is where artificial intelligence (AI)-based generative approaches can excel.

12.1.3 AI-Based Methods can Accelerate Medicinal Chemistry

AI-based workflows and pipelines are being adopted and deployed across a number of industries, including medical, pharmaceuticals, and life sciences, and are poised to have a significant impact on productivity, output, and the economy [18–20]. Over the last few years, the drug design and discovery sector, led by research and development by pharmaceutical industry and academic institutions, has also employed AI-based strategies to optimize a number of components in the pipeline (Figure 12.1). Specifically, this has impacted drug design, repurposing, screening, poly-pharmacology, and chemical synthesis [21].

Here we highlight the efforts that address the issues impacting the efficiency of the aforementioned DMTA cycle – drug design and chemical synthesis. Generative AI-based strategies are able to fulfill the promise of truly *de novo* drug design by being able to explore and sample the chemical space far more efficiently and comprehensively, subject to any medicinal chemistry constraints (target product profile, substructure or descriptor constraints, 3D scores, etc.), compared to traditional methods. Such generative approaches have found successful applications in tasks

such as image synthesis, language translation, text generation, and music generation [22–24]. Now these approaches have also found their way in chemistry. The methods based on these approaches generally do not rely solely on structural similarity, instead they learn the property similarity in the latent space, and are therefore able to design diverse set of ideas for the molecules that may be close in the physicochemical property space [25]. This approach has been extensively validated [26–28] and has yielded positive results in real-life case studies [29]. A key component of the generative AI in *de novo* drug design is synthetic accessibility, which has come in to a sharp focus more recently [30, 31]. Compounds obtained from generative chemistry can satisfy multiple objectives *in silico*, but at the same time can be difficult to synthesize [32]. Therefore, it is essential to ensure that these AI systems can optimize synthetic accessibility of the generated compounds and allow chemists to more efficiently choose the compounds to be synthesized and tested.

In this chapter, various components of the generative AI and synthetic accessibility that are implemented in *de novo* drug design and have been impacting drug discovery projects are discussed. Specifically we explore the state-of-the-art methods and algorithms that power these AI systems and platforms that have been commonly deployed across a variety of drug discovery and design programs. The aim of this chapter is to introduce the reader to the technical and functional aspects of these systems, and pave the way to further diversify the ecosystem that has been developing at the intersection of AI/ML, other computational techniques, chemistry, and biology, for the benefit of drug discovery.

12.2 Quantitative Structure-Activity Relationship Models

12.2.1 Introduction to QSAR Models

QSAR models are developed using one or more statistical model building tools, which may be broadly categorized into regression- and classification-based approaches. Like other regression models, QSAR regression models relate a set of “predictor” variables to the potency of the response variable, while classification QSAR models relate the predictor variables to a categorical value of the response variable. The mathematical form of a QSAR model is:

$$\text{Activity} = f(\text{physiochemical properties and/or structural properties}) + \text{error}$$

In the following Sections (12.2.2 and 12.2.3), we distinguish between the traditional machine learning-based techniques for QSAR models and the newer techniques based on deep neural networks (DNN).

12.2.2 QSAR Machine Learning Methods

Building statistical models to predict chemical properties is of paramount importance as it would allow to discard compounds without having to synthesize and

test them [33]. However, applying statistical modeling techniques to molecules face many hurdles. First, the molecules must be represented as fingerprints (vectors) to be processed by a statistical model [34]. Then, since the chemical space is very sparse, considerations about the applicability domain of the models are crucial. Finally, these models will make mistakes – no statistical model is perfect – and furthermore, the expert using them (e.g. medicinal or computational chemist) needs to be convinced of their usefulness. Therefore, explaining the results obtained through interpretability and not having black-box QSAR model is essential. Various fingerprints have been developed [35–39] not only for QSAR models but also for other operations on molecules, like similarity computation [40, 41] or clustering. The first class of fingerprints consists of listing molecular descriptors and building a vector from them, which in some cases shows good performance. These fingerprints relying on descriptors do not encode local features of the structure, but instead provide global information of the molecule. On the other hand, extended-connectivity fingerprints [42] enumerate the molecular features and directly encode the structure of the molecule. These fingerprints are fast-to-compute and generally make a good baseline for building QSAR models [43]. Finally, with the development of deep-learning methods a new category of fingerprints called Learned Representation has emerged [44]. It allows to encode molecular graphs or SMILES strings, and shows competitive results in QSAR model benchmarks [45]. These three kinds of molecular representations can be computed with the molecule alone (whether it be a small molecule or a peptide). It is worth mentioning that some fingerprints include the interaction between the molecule and a protein [46], which allows to build 3D QSAR models, especially useful in scaffold-hopping tasks [47].

Once molecules are converted to fingerprints, a statistical model is used to predict the property of interest. Depending on the model, varying performances can be obtained on the same task, but it is generally hard to know in advance which one to use. Use of multiple solutions like linear models [48, 49], random forests [50–52], support vector machines (SVM) [53, 54], neural networks, etc. [55] have been studied. Benchmarks exist to compare these models [49], [56–59]. However, these comparisons are limited to specific use cases and test sets, and generally do not allow to intrinsically rank the statistical models for QSAR. A summary diagram can be found in [37, Figure 1]. Aside from the model-measured performances, its applicability domain is critical to estimate. Though it is hard and is subject to many biases, efforts have been developed to measure the applicability domain and quantify the model's errors [60]. A simple yet efficient method to estimate the applicability domain is to evaluate the similarity of the estimated compound with the training dataset [61]. More sophisticated techniques take into account the defaults of the models aimed at improving these applicability domain metrics [62, 63]. The restrained applicability domain of a QSAR is the reason why it is generally very risky to use these models for scaffold hopping in very diverse chemical spaces. Efforts have been made to extend the applicability domain, thanks to federated learning [64], which consists of training a model on multiple private datasets while maintaining the privacy of the data involved [65]. This kind of approach is very promising, especially in the pharmaceutical industry, where data privacy is critical. However, there are, to-date,

few demonstrations [66] that the performance of the QSAR model is significantly improved.

Finally, the interpretability of the QSAR model is key for proper usage and understanding [67]. Most models are indeed black boxes, either because of the model (neural network) or the features (learned representations). Adding interpretability (for instance, using SHAP values [68–70]) allows an expert user to spot the biases in the model and results that stem from spurious correlations rather than causality.

12.2.3 QSAR Deep Neural Networks Methods

Numerous research studies have demonstrated that graph neural networks (GNN) may be a more effective modeling technique for predicting chemical properties than descriptor-based techniques [71–78]. Due to their remarkable ability to learn complex and often non-linear relationships between structures and properties, deep learning (DL) algorithms have more recently revolutionized the classic cheminformatics activity presented in Section 12.2.2.

The DL-based models largely fall into two main categories – descriptor-based models and graph-based models. In descriptor-based DL models, molecular descriptors and/or fingerprints commonly used in traditional QSAR models are used as the input, and then a specific DL architecture is deployed to train a model [79]. While in the graph-based DL models, the basic chemical information encoded by molecular graphs is used as the input, and then a graph-based DL algorithm, such as GNN, is used to train a model.

The GNN generalizes the convolution technique to the irregular molecular graph, which is a natural representation of chemical structures, in a manner similar to how convolutions are applied to regular data such as pictures and texts. A graph $G = (V, E)$ can be defined as the connectivity relationships between a collection of nodes (V) and a collection of edges (E). Naturally, one may also think of a molecule as a graph made up of a number of atoms (nodes) and a number of bonds (edges). In essence, GNN aims to learn the representations of each atom by aggregating the information from its nearby atoms encoded by the atom feature vector and the information of the connected bonds encoded by the bond feature vector through messages passing across the molecular graph repeatedly. This is done after the state updating of the central atoms and read-out operation. Through the read-out phase, the learned atom representations can then be applied to the prediction of molecular attributes. A key feature of GNN is its ability to automatically learn task-specific representations using graph convolutions while not requiring conventional manually created descriptors and/or fingerprints. An explanatory diagram is presented in [[80], Figure 1].

As of 2022, the most common Python libraries to manipulate GNNs are PyTorch Geometric, Deep Graph library, DIG, Spektral, and TensorFlow GNNs.

12.3 Modes of Generative AI in Chemistry

12.3.1 General Introduction

Generative modeling is a challenging problem to solve. It is fundamentally harder than discriminative modeling (e.g. statistical classification) because it models objects from arbitrary distributions of arbitrary nature and complexity (images, texts, graphs, audios, etc.). Being unsupervised, generative models benefit from large unlabeled datasets especially when designed as deep neural networks.

Generative models are practical not only for sampling novel data points from a given distribution but also for characterizing the likelihood of the sampled data points. When coupled with an optimization method for a given fitness score, they model the distribution of good-scored data points. This is the desired behavior of generative models applied to drug discovery – generate novel molecules optimizing a defined fitness score.

Generative models for molecular design can be characterized by three main features:

1. **Which molecular representation they use:** It can be either text (SMILES [81], SELFIES [82]), a graph [83], a set of fragments [84], or a synthesis tree [85].
2. **How they generate molecules:** The generation strategy can use a simple policy [86], add or remove atoms or bonds, for example. The most successful generative models in drug design in the literature rely on DNN from two main families of models.
 - **Deep autoregressive networks [87]:** Trained to retrieve a missing part of a piece of data knowing the rest of this piece of data and are compatible with reinforcement learning optimization.
 - **Autoencoder models [88]:** Trained to reconstruct the desired piece of data starting from an initial input (same piece of data or another type of piece of data). It is compatible with optimization in an Euclidean latent space and evolutionary algorithms.
3. **How they perform property optimization:** The property optimization strategy can be based on reinforcement learning [89], Bayesian optimization [90], or evolutionary algorithms [91].

12.3.2 Generative AI in Lead Optimization

MPO is required in a drug discovery project especially in the lead optimization stage in order to identify the rare compounds satisfying all the objectives of the project, such as activity across multiple biological assays, selectivity, (lack of) toxicity, pharmacokinetics, synthetic accessibility, and novelty, to name a few [92]. In the last five years, the development of AI approaches to drug discovery, and more specifically *de*

novo drug design through the use of deep generative models, has triggered a lot of interest in the computer-aided drug design (CADD) community. In this chapter, we briefly discuss the successful implementation of this approach.

The easiest way to generate molecules is to use a deep recurrent neural network (RNN), and more precisely, a deep long short-term memory (LSTM) [93], to generate molecules represented as SMILES [94]. The LSTM should first be trained on a big generic database such as ChEMBL or ZINC databases, using teacher forcing [95], to build a character-based language model for generating SMILES strings. Recall that the role of a language model p is to model the next character probability distribution given the sequence of previous characters:

$$p(x_{t+1}|x_1x_2x_t) = LSTM(x_{t+1}|x_1x_2x_t) \quad (12.1)$$

SMILES are generated by iteratively sampling the next character from its inferred past conditioned distribution $p(x_{t+1}|x_1x_2x_t)$ generating a SMILES starts and ends, respectively, with the special tokens of the vocabulary “START” and “END.” Molecules in ChEMBL database are transformed into their canonical achiral RDKit version. No data augmentation is performed either by enumerating the different ways of writing a SMILES, or by enumerating the tautomeric forms of the same compound. LSTM language model trained with this approach generates achiral SMILES. Identical compounds can be generated with different writings of their SMILES. Tautomers of the same compound are generated as distinct molecules. After being trained on ChEMBL, the LSTM language model had a 94% SMILES chemical validity rate, which implies that the LSTM trained on ChEMBL database has learnt to generate molecules belonging to ChEMBL chemical space.

Crucially, generated molecules should stay near the chemical space of the lead optimization series. Thus, the previous LSTM model is retrained by teacher forcing on the lead optimization dataset. This second training allows to zoom in on the chemical space studied to generate molecules similar to the lead optimization chemical series. The simplest molecule optimization strategy that can be used along with a SMILES LSTM generator is the hill climbing strategy [94]. It is an iterative process where the LSTM generative model is fine-tuned in teacher forcing on an optimal set of SMILES that evolves over time as follow: step after step, this set of SMILES is updated by retaining only the top-scored generated compounds (10% for example) since the first step.

An MPO lead optimization dataset is a list of molecules with experimental bioactivity measurements on multiple biological assays. In order to score novel generated molecules, QSAR models are trained on each assay measurements. We recommend using binary classification models after binarizing the data using the desired thresholds of the lead optimization project. Indeed, binary classification models better handle unbalanced datasets and can better predict the minority class. The reward (fitness) score used for ranking molecules in hill climbing combines the predicted probabilities from QSAR models (p_i), a measure of similarity to the initial dataset, and any other physical or chemical properties of the project (Molecular Weight for example). An aggregation function that works pretty well is the geometric mean of

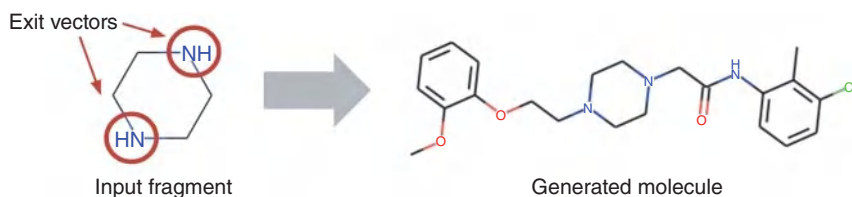


Figure 12.2 Growing from an initial fragment with defined exit vectors.

scaled scores (between 0 and 1), which allows us to transform our problem from multi-objective optimization to mono-objective optimization.

12.3.3 Fragment Growing

Generative models can be used to address different challenges in drug design, and one of them is fragment growing. The goal of fragment growing is to start from an initial molecular structure with a pre-defined exit vector, Figure 12.2, then choose new fragments to be plugged into each exit vector to obtain a new molecule. The fragments are selected from a dataset of commercial compounds and are added to the molecule through chemical reactions. One fragment and the associated reaction are predicted for each exit vector of the initial molecular structure.

This process can be learned by a deep learning model. The model relies on three architectures: a feed-forward neural network to predict the reactions, another feed-forward to choose the building blocks, and a recurrent neural network to capture the sequential aspect of the fragment growing process, Figure 12.3. With this architecture, it is possible to model the likelihood of the sequences with the model's parameters. Modeling the likelihood enables distribution learning and the use of reinforcement learning. With reinforcement learning, the selection of the building blocks plugged into the initial molecular structure is optimized for a given score such as docking score, quantitative estimate of druglikeness (QED), or similarity. The application of fragment growing is not limited to 2D approaches. It has been successfully implemented in structure-based drug design approaches, especially like fragment-based drug design wherein a hit fragment in the binding pocket of the target protein must be grown into a lead-like and (potentially) potent compound with good binding affinity in the binding pocket [96–99].

12.3.4 Novelty Generation

12.3.4.1 The Model

The novelty generator is a deep learning algorithm based on a transformer architecture [100], which was trained to jump from patent to patent creating, by design, highly novel molecules (Figure 12.4). The transformer model is used in a sequence-to-sequence way, very similar to machine translation tasks. The supervision dataset is composed of pairs of molecules (Molecule_A, Molecule_B) where:

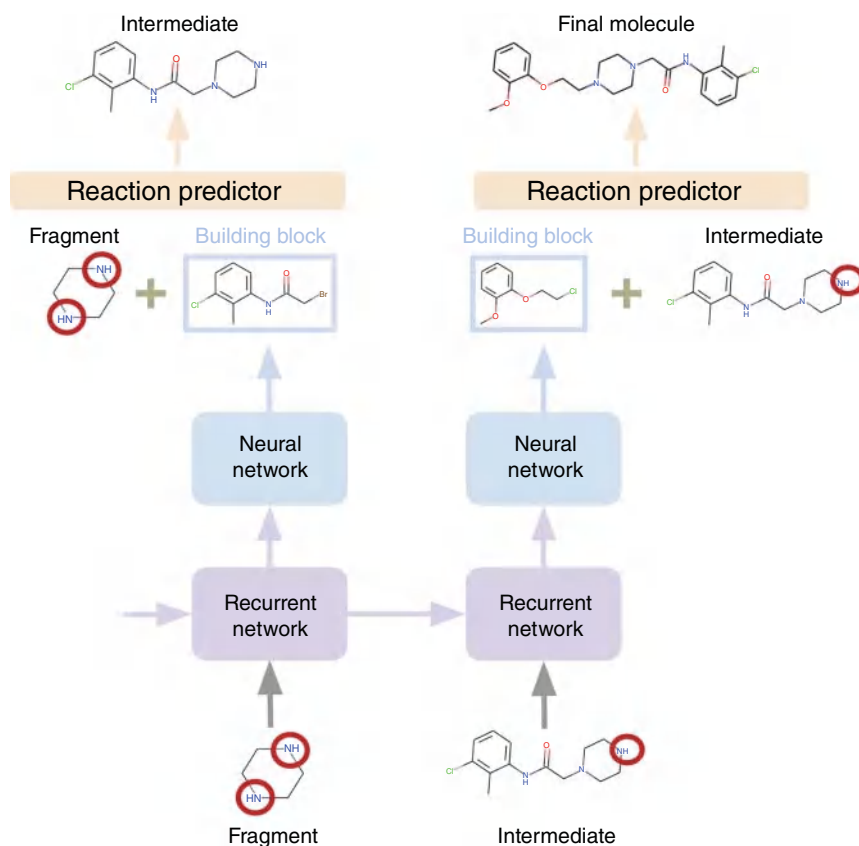


Figure 12.3 Model architecture for fragment growing. The model takes as input the fragment and its exit vectors. The model selects which building block among a dataset should react with the fragment provided.

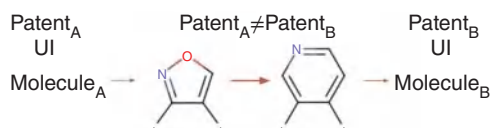


Figure 12.4 Description of how the input/output pairs are extracted to build the training dataset of the novelty generator.

- The molecules come from two different patents
- Molecule_B can be retrieved using an MMP (Matched Molecular Pair) rule from Molecule_A

The two challenges are first building this dataset, and then training the model to perform the desired transformation. To build the supervised dataset, the input dataset is a cleaned sample of SureChEMBL [101], this leads to $N=18$ million pairs. Extracting all the MMP rules on this dataset is fairly untractable – for k elements, there are k^2 pairs, hence $>10^{12}$ rules to compute. To reduce the complexity to $\frac{N^2}{k}$, the SureChEMBL dataset was clustered into $k=1000$ clusters. The entire pipeline is summed up in Figure 12.5.

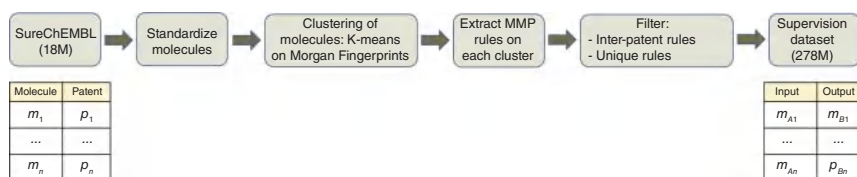


Figure 12.5 Pipeline to build the training dataset of the novelty generator.

As mentioned above, the model used for this task is a Transformer model, which is an encoder–decoder model with several attention mechanisms. The model was trained with the teacher forcing method [102], meaning that at each iteration the model samples a character, which is only used to contribute to the value of the loss function, but the truth character is added to the sequence. The model can be long and computationally expensive to train. As an estimate, training this model at Iktos took around 1 month on an 11 GB GPU (NVIDIA 2080 ti). The inference method is different from the training method. In inference, the model doesn't have access to the ground truth, so it samples character by character following the decoder probability distribution until an “end of sentence” character is sampled. If sampling occurs several times from the same input SMILES, the stochasticity of the decoding outputs different results with various probabilities. To further augment the solutions given by the model, we observed that enumerating the input SMILES and applying the inference to those SMILES increased the diversity of the generated molecules.

12.3.4.2 Optimization of the Novelty Generator

A reward function-based optimization method can be deployed for novelty generators. Given a defined reward function, a loop on the novelty generator can be created so that at every iteration the generator gets back the best-scored output molecules to be used in the next iteration, and therefore climb the reward function. In this optimization method, the model is considered as a black box, and its weights are not updated. The reward is optimized by selecting which molecules will be the input molecules of the next iteration. This optimization method is part of hill climbing optimization methods. An optimization method for the continuous latent space of the model have been tried [103]. A major limitation here is that the dimension of the encoded SMILES depends on the length of the SMILES, so the optimization is limited at each time step to the latent space corresponding to one SMILES length. The experiment to invalidate the [103] method was to compare the score for the same number of SMILES decoded with this method vs. decoding with random noise.

12.4 Importance of Synthetic Accessibility

12.4.1 Overview

In small molecule drug discovery projects, generative models can be used to design massive libraries of molecules with specific properties [29, 94]. The optimization

of an AI-molecular generator to explore a given chemical space and propose new well-scored molecules in an MPO project is mostly based on molecular properties and fingerprints [75, 94, 103–106]. However, one of the major challenges in any CADD project is that the molecules need to be synthesized. The question of whether or not the generated molecules can be synthesized is not systematically taken into account during generation, even though the synthesizability of the generated molecules is a fundamental requirement for such methods to be useful in practice. Generative models are known to sample lots of non accessible molecules [107, 108], and few synthesizability scores are known in the literature to be used in the pipeline of molecular generation [32, 85, 109, 110].

12.4.2 Synthetic Scores

No chemical rule is able to completely answer the question of whether a molecule with a valid SMILES can be synthesized or not. Moreover, the evaluation of such scores is challenging, particularly due to the difficulty in interpreting the values. A simple way to define synthesizability is with a binary score denoting “synthesizable” or “not synthesizable.” While a binary score is useful, it has limits, as it does not allow the prioritization of molecules of the same score. Also, a continuous score gives more signal when used as a reward for a *de novo* drug design algorithm. With the recent efforts of the community, some continuous scores were recently developed to describe synthetic accessibility [111–114]. These can be based on chemical substructures, domain expertise, or output of models fitting expert scores. However, as two very similar molecules may have different synthetic routes due to a single functional group, it may be difficult to find a proxy for a true retrosynthetic analysis. The RA score, for retrosynthetic accessibility score [113], is a predictor of the binary score given by the AiZynthFinder retrosynthesis tool [114]. Its value range is between 0 and 1, and according to the score, the higher the value the more optimistic the algorithm is about the synthesis of the molecule. The SC score, for synthetic complexity score [111], ranks the molecules and scores them from 1 to 5. Based on the criteria that products are more complex than their reactants, a neural network trained on a corpus of reactions was used to build the score. Molecules with lower values have a more optimistic synthesizability profile. Finally, the SA score for synthetic accessibility score [112], is based on a heuristics where molecular complexity and fragment contributions are used to evaluate synthetic tractability. Low scores indicate less complex molecules and therefore more feasible compounds. We believe that the features taken into account to compute these scores are not sufficient to encapsulate all of the information about synthesizability.

More recently, AI has been introduced to address some of these challenges and to aid synthetic, medicinal, and computational chemists. Iktos has developed Spaya [115], a template-based retrosynthesis AI software that computes synthetic routes and ranks them based on a synthesizability score. The Retro-Score (RScore) is a synthetic feasibility score derived from the output of a full Spaya retrosynthetic analysis for a given molecule. As highlighted below, conducting a full retrosynthetic analysis to determine synthesizability is essential. The RScore can be used:

- (1) to evaluate the synthesizability of molecules given by generative models,
- (2) to guide the molecular generator to an area of the chemical space where molecules are synthesizable.

Due to the high computational costs associated with the computation of a full retrosynthetic analysis needed to obtain the RScore, an easier way to compute score is RSPred, obtained by training a neural network on the output of the Spaya RScore and performs comparably to the RScore in a variety of tasks, but can be computed orders of magnitude faster. Further, in order to simplify the use of the algorithm on large batches of molecules (libraries), Spaya-API has been developed [115]. It is an API running on Spaya's algorithmic engine for library scoring purposes, which has been used herein to evaluate the synthetic accessibility of newly generated molecules. For a given molecule (m), the RScore is derived from routes proposed by Spaya, but handled in a high-throughput manner by Spaya-API. The worst RScore value is 0 (when no route is found in a given time interval), and the best score is 1 (when the route is a one-step retrosynthesis matching exactly a reaction in the literature). To score a molecule and obtain its RScore value, Spaya-API performs a retrosynthetic analysis with an early stopping process. The early stopping mode stops the Spaya run when a route with a score above the predefined threshold (set to 0.6 by default) is found, or after the defined timeout (set to one minute by default) has elapsed. The RScore of a molecule is defined as:

$$RScore(m) = \max_{\substack{\text{routes given by Spaya} \\ \text{with early stopping}}} (score(route(m))) \quad (12.2)$$

The score is rounded to 1 decimal, and hence can take 11 different values (from 0.0 to 1.0). Spaya-API also returns the number of steps for the best synthetic route found for each input molecule. The list of commercial compounds used for the retrosynthesis is a catalog of 60M commercially available starting materials provided by Mcule [116], Chemspace [117], eMolecules [118], and Key Organics [119].

Compute time is an essential attribute of a score as it may limit its usage on large-scale data sets. In Table 12.1 compute time estimates of the different synthetic scores are shown. The RScore, obtained through a full retrosynthesis (with a one minute timeout), is by far the most time-consuming score. Due to its scalability,

Table 12.1 Compute time per molecule for the different synthetic scores.

Synthetic score	Time per molecule (ms)
RA score	28
SC score	241
SA score	2
RScore	40 000
RSPred	1

Spaya-API accelerates RScore computation on large batch of molecules. The prediction of the latter, RSPred, is the fastest score to compute, only 1ms per molecule, 40 000 times faster than the RScore. The SA score closely follows with 2 ms per molecule, the RA score is one order of magnitude slower while the SC score is two orders of magnitude slower.

12.4.3 Integration of Synthetic Scores in Generative AI

In the context of generative AI, the generator's weights are optimized in order to maximize a reward function. The various synthetic scores described above can be integrated in the reward function, so that the generated molecules get easier to synthesize. Figure 12.6 illustrates the pipeline on the use case presented below.

12.4.3.1 An Example of a Lead Optimization Use Case

This task is a generation around a library of 463 structurally homogeneous PI3K and mTOR inhibitors. The details of this experiment can be found in reference [120]. The dataset used here serves as a simplified proxy for a real-life MPO in a lead optimization project with four objectives to be optimized : QED, Pi3K, mTOR, and similarity to initial dataset. Six generations were run based on this dataset — one without any synthetic score constraint, and five with synthetic score constraints (RA, SC, SA, RScore, and RSPred). RScore1min and RScore3min correspond to RScore calculated at one and three minute timeouts, respectively.

The main metric to evaluate the quality of a generation method is the number of generated molecules validating all the constraints, which also have a good RScore. The right graph in Figure 12.6 shows for each of the five generations how many molecules validate the thresholds, and their RScore range. To summarize the obtained results in the PI3K/mTOR experiment, RScore appeared to be the best synthetic feasibility score to use as a synthesizability constraint integrated in an MPO generation, since it outperformed other methods in generating a high number of compounds in the defined blueprint with a good synthetic feasibility score. It is worth noting that RSPred is a good proxy of the RScore metric with a much lower computational cost. The SA score has some correlation to the RScore, but the generation under SA constraint outputs less than half as many interesting molecules as the generation under RScore1min or RSPred constraint. The other synthetic constraints were not very useful in this experiment — the RA score has poor precision, meaning that among the molecules well scored by RA score, very few actually have a good RScore3min, and when included in the reward of a generation, almost all molecules get a high reward and the generator can't be optimized toward easier to make molecules; the SC score has no correlation to the RScore3min, so it comes as no surprise that the generation under SC score constraint fails to optimize the RScore3min during the generation, and gives poor results.

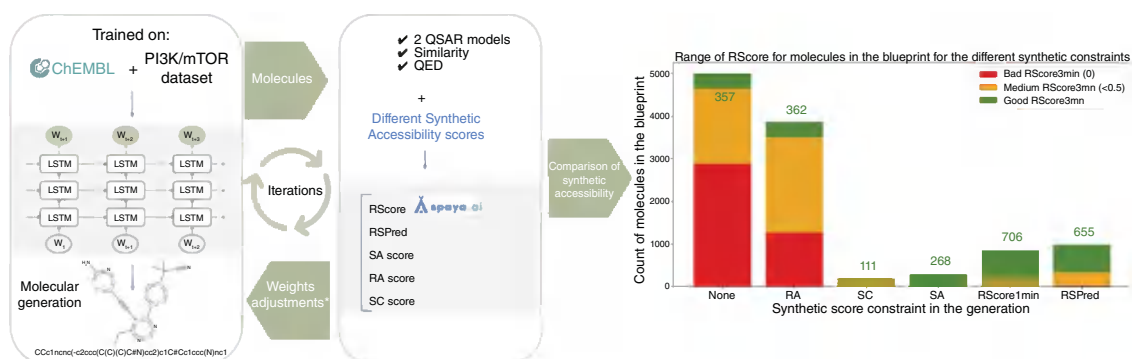


Figure 12.6 Pipeline of AI generative models with a synthetic accessibility constraint.

12.5 The Road Ahead

AI and machine learning-based tools are poised to become *must-haves* for every pharmaceutical company and drug discovery laboratory over the next decade. The methodologies and processes are maturing, and outstanding issues are being resolved. The innovation in both algorithmic and computational capacity is rapidly accelerating because of the desire to overcome longstanding issues in this field of study, and it is evident by the rapid increase in the scientific meetings and publications in this field of study. Caution is however warranted, and it is worth re-emphasizing that AI is not magic [121, 122]. While AI and machine learning algorithms are valuable in their ability to learn, it is the data generated by chemists in a myriad of experiments involving drug-like molecules that forms the backbone of this approach. Thus, both the quality of the AI framework (e.g. neural net architecture, deep learning algorithms) and the training data are of utmost importance [123]. It is also imperative to remember that the best AI system in drug discovery is the one that works alongside chemist's human intelligence — a chemist's domain and project knowledge cannot be replicated by AI.

Depending on the complexity of the problem, significant amounts of training data may be required, and it is unreasonable to expect one source to generate all or even most of it. Individual data sources (e.g. pharmaceutical companies) fiercely guard and seldom publicly disclose the data they generate in drug discovery pipelines for business confidentiality reasons. Data sharing remains a critical issue in the community, but crucial attempts are being made to use data from multiple sources to train model without actually making the data public (Federated Learning) [124, 125].

As the understanding of the benefits of using AI across the value chain of the drug discovery process continues to improve, some of the skepticism held by major stakeholders has been alleviated [126–128]. This understanding has been further augmented by examples of AI adding value to challenging projects [29]. Furthermore, there has been growing interest in developing tools for the explainability and interpretability of the models underneath the AI and machine learning systems used in the drug design and discovery space [129–131]. These methods can reveal the rationale behind the predictions made by models, offering insight in to the *black box*, thereby helping chemists devise useful strategies. Taken together, these methods can go a long way in further establishing the credentials of AI-based methods as a useful component in the drug discovery value chain.

References

- 1 Ha, J., Park, H., Park, J., and Park, S.B. (2021). Recent advances in identifying protein targets in drug discovery. *Cell Chemical Biology* 28 (3): 394–423.
- 2 Hughes, J.P., Rees, S., Kalindjian, S.B., and Philpott, K.L. (2011). Principles of early drug discovery. *British Journal of Pharmacology* 162 (6): 1239–1249.
- 3 Keserü, G.M. and Makara, G.M. (2006). Hit discovery and hit-to-lead approaches. *Drug Discovery Today* 11 (15–16): 741–748.

- 4 Mouchlis, V.D., Afantitis, A., Serra, A. et al. (2021). Advances in de novo drug design: from conventional to machine learning methods. *International Journal of Molecular Sciences* 22 (4): 1676.
- 5 Dang, C.V., Reddy, E.P., Shokat, K.M., and Soucek, L. (2017). Drugging the 'undruggable' cancer targets. *Nature Reviews Cancer* 17 (8): 502–508.
- 6 An, S. and Fu, L. (2018). Small-molecule PROTACs: an emerging and promising approach for the development of targeted therapy drugs. *eBioMedicine* 36: 553–562.
- 7 Müller, C.E., Hansen, F.K., Gütschow, M. et al. (2021). New drug modalities in medicinal chemistry, pharmacology, and translational science: joint virtual special issue by *Journal of Medicinal Chemistry*, *ACS Medicinal Chemistry Letters*, and *ACS Pharmacology & Translational Science*. *Journal of Medicinal Chemistry* 64 (19): 13935–13936.
- 8 Yang, W., Gadgil, P., Krishnamurthy, V.R. et al. (2020). The evolving druggability and developability space: chemically modified new modalities and emerging small molecules. *The AAPS Journal* 22 (2): 1–14.
- 9 Maurya, N.S., Kushwaha, S., and Mani, A. (2019). Recent advances and computational approaches in peptide drug discovery. *Current Pharmaceutical Design* 25 (31): 3358–3366.
- 10 Sliwoski, G., Kothiwale, S., Meiler, J., and Lowe, E.W. (2014). Computational methods in drug discovery. *Pharmacological Reviews* 66 (1): 334–395.
- 11 Lionta, E., Spyrou, G., Vassilatis, D.K., and Cournia, Z. (2014). Structure-based virtual screening for drug discovery: principles, applications and recent advances. *Current Topics in Medicinal Chemistry* 14 (16): 1923–1938.
- 12 Hamza, A., Wei, N.-N., and Zhan, C.-G. (2012). Ligand-based virtual screening approach using a new scoring function. *Journal of Chemical Information and Modeling* 52 (4): 963–974.
- 13 Hoffmann, T. and Gastreich, M. (2019). The next level in chemical space navigation: going far beyond enumerable compound libraries. *Drug Discovery Today* 24 (5): 1148–1156.
- 14 Walters, W.P. and Wang, R. (2020). New trends in virtual screening. *Journal of Chemical Information and Modeling* 60 (9): 4109–4111.
- 15 Fresnais, L. and Ballester, P.J. (2021). The impact of compound library size on the performance of scoring functions for structure-based virtual screening. *Briefings in Bioinformatics* 22 (3): bbaa095.
- 16 Gentile, F., Yaacoub, J.C., Gleave, J. et al. (2022). Artificial intelligence-enabled virtual screening of ultra-large chemical libraries with deep docking. *Nature Protocols* 17 (3): 672–697.
- 17 Raymond, J.-L. (2015). The chemical space project. *Accounts of Chemical Research* 48 (3): 722–730.
- 18 Furman, J. and Seamans, R. (2019). Ai and the economy. *Innovation Policy and the Economy* 19 (1): 161–191.
- 19 Woo, M. (2019). An ai boost for clinical trials. *Nature* 573 (7775): S100–S100.
- 20 Muehlematter, U.J., Daniore, P., and Vokinger, K.N. (2021). Approval of artificial intelligence and machine learning-based medical devices in the USA and

- EUROPE (2015–20): a comparative analysis. *The Lancet Digital Health* 3 (3): e195–e203.
- 21 Paul, D., Sanap, G., Shenoy, S. et al. (2021). Artificial intelligence in drug discovery and development. *Drug Discovery Today* 26 (1): 80.
- 22 Park, S.-W., Ko, J.-S., Huh, J.-H., and Kim, J.-C. (2021). Review on generative adversarial networks: focusing on computer vision and its applications. *Electronics* 10 (10): 1216.
- 23 Reed, S., Akata, Z., Yan, X. et al. (2016). Generative adversarial text to image synthesis. *International Conference on Machine Learning*, 1060–1069. PMLR.
- 24 Wang, L., Chen, W., Yang, W. et al. (2020). A state-of-the-art review on image synthesis with generative adversarial networks. *IEEE Access* 8: 63514–63537.
- 25 Vogt, M. (2022). Using deep neural networks to explore chemical space. *Expert Opinion on Drug Discovery* 17 (3): 297–304.
- 26 Wang, M., Wang, Z., Sun, H. et al. (2022). Deep learning approaches for de novo drug design: an overview. *Current Opinion in Structural Biology* 72: 135–144.
- 27 Schneider, G. and Clark, D.E. (2019). Automated de novo drug design: are we nearly there yet? *Angewandte Chemie International Edition* 58 (32): 10792–10803.
- 28 Blaschke, T., Arús-Pous, J., Chen, H. et al. (2020). REINVENT 2.0: an AI tool for de novo drug design. *Journal of Chemical Information and Modeling* 60 (12): 5918–5922.
- 29 Perron, Q., Mirguet, O., Tajmouati, H. et al. (2022). Deep generative models for ligand-based de novo design applied to multi-parametric optimization. *Journal of Computational Chemistry* 43 (10): 692–703.
- 30 Makara, G.M., Kovács, L., Szabó, I., and Pöcze, G. (2021). Derivatization design of synthetically accessible space for optimization: *in silico* synthesis vs deep generative design. *ACS Medicinal Chemistry Letters* 12 (2): 185–194.
- 31 Miljković, F., Rodríguez-Pérez, R., and Bajorath, J. (2021). Impact of artificial intelligence on compound discovery, design, and synthesis. *ACS Omega* 6 (49): 33293–33299.
- 32 Gao, W. and Coley, C.W. (2020). The synthesizability of molecules proposed by generative models. *Journal of Chemical Information and Modeling* 60 (12): 5714–5723.
- 33 Kar, S., Roy, K., and Leszczynski, J. (2018). Impact of pharmaceuticals on the environment: risk assessment using QSAR modeling approach. In: *Computational Toxicology. Methods in Molecular Biology*, vol. 1800 (ed. O. Nicolotti), 395–443. New York: Springer.
- 34 Zagidullin, B., Wang, Z., Guan, Y. et al. (2021). Comparative analysis of molecular fingerprints in prediction of drug combination effects. *Briefings in Bioinformatics* 22 (6): bbab291.
- 35 Wigh, D.S., Goodman, J.M., and Lapkin, A.A. (2022). A review of molecular representation in the age of machine learning. *WIREs Computational Molecular Science* 12 (5): e1603.

- 36 Capecchi, A., Probst, D., and Reymond, J.-L. (2020). One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *Journal of Cheminformatics* 12: 43.
- 37 Pattanaik, L. and Coley, C.W. (2020). Molecular representation: going long on fingerprints. *Chem* 6 (6): 1204–1207.
- 38 Orosz, Á., Héberger, K., and Rácz, A. (2022). Comparison of descriptor- and fingerprint sets in machine learning models for ADME-Tox targets. *Frontiers in Chemistry* 10: 852893.
- 39 Sandfort, F., Strieth-Kalthoff, F., Kühnemund, M. et al. (2019). A structure-based platform for predicting chemical reactivity. *ChemRxiv*.
- 40 Venkatraman, V., Gaiser, J., Roy, A., and Wheeler, T.J. (2022). Molecular fingerprints are not useful in large-scale search for similarly active compounds†. *bioRxiv*.
- 41 O’Boyle, N.M. and Sayle, R.A. (2016). Comparing structural fingerprints using a literature-based similarity benchmark. *Journal of Cheminformatics* 8: 36.
- 42 Rogers, D. and Hahn, M. (2010). Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling* 50 (5): 742–754.
- 43 Mittal, R.R., McKinnon, R.A., and Sorich, M.J. (2009). Comparison data sets for benchmarking QSAR methodologies in lead optimization. *Journal of Chemical Information and Modeling* 49 (7): 1810–1820.
- 44 Preuer, K., Renz, P., Unterthiner, T. et al. (2018). Fréchet ChemNet distance: a metric for generative models for molecules in drug discovery. *Journal of Chemical Information and Modeling* 58 (9): 1736–1741.
- 45 Yang, K., Swanson, K., Jin, W. et al. (2019). Are learned molecular representations ready for prime time? *ChemRxiv*.
- 46 Salentin, S., Schreiber, S., Haupt, V.J. et al. (2015). PLIP: fully automated protein-ligand interaction profiler. *Nucleic Acids Research* 43 (W1): W443–W447.
- 47 Laufkötter, O., Sturm, N., Bajorath, J. et al. (2019). Combining structural and bioactivity-based fingerprints improves prediction performance and scaffold hopping capability. *Journal of Cheminformatics* 11 (1): 54.
- 48 Duchowicz, P.R. (2018). Linear regression QSAR models for polo-like kinase-1 inhibitors. *Cells* 7 (2): 13.
- 49 Konovalov, D.A., Llewellyn, L.E., Heyden, Y.V., and Coomans, D. (2008). Robust cross-validation of linear regression QSAR models. *Journal of Chemical Information and Modeling* 48 (10): 2081–2094.
- 50 Svetnik, V., Liaw, A., Tong, C. et al. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Sciences* 43 (6): 1947–1958.
- 51 Lee, K., Lee, M., and Kim, D. (2017). Utilizing random forest QSAR models with optimized parameters for target identification and its application to target-fishing server. *BMC Bioinformatics* 18 (16): 567.
- 52 Trinh, T.X., Seo, M., Yoon, T.H., and Kim, J. (2022). Developing random forest based QSAR models for predicting the mixture toxicity of TiO₂ based nano-mixtures to *Daphnia magna*. *NanoImpact* 25: 100383.

- 53 Shi, Y. (2021). Support vector regression-based QSAR models for prediction of antioxidant activity of phenolic compounds. *Scientific Reports* 11: 8806.
- 54 Mei, H., Zhou, Y., Liang, G., and Li, Z. (2005). Support vector machine applied in QSAR modelling. *Chinese Science Bulletin* 50: 2291–2296.
- 55 Darnag, R., Minaoui, B., and Fakir, M. (2017). QSAR models for prediction study of HIV protease inhibitors using support vector machines, neural networks and multiple linear regression. *Arabian Journal of Chemistry* 10: S600–S608.
- 56 Wu, Z., Ramsundar, B., Feinberg, E.N. et al. (2018). MoleculeNet: a benchmark for molecular machine learning. *Chemical Science* 9: 513–530.
- 57 Kokabi, M., Donnelly, M., and Xu, G. (2020). Benchmarking small-dataset structure-activity-relationship models for prediction of wnt signaling inhibition. *IEEE Access* 8: 228831–228840.
- 58 Arshadi, A.K., Salem, M., Firouzbakht, A., and Yuan, J.S. (2022). MolData, a molecular benchmark for disease and target based machine learning. *Journal of Cheminformatics* 14 (1): 10.
- 59 Czub, N., Paclawski, A., Szlek, J., and Mendyk, A. (2021). Curated database and preliminary AutoML QSAR model for 5-HT1A receptor. *Pharmaceutics* 13 (10): 1711.
- 60 Norinder, U., Carlsson, L., Boyer, S., and Eklund, M. (2014). Introducing conformal prediction in predictive modeling. A transparent and flexible alternative to applicability domain determination. *Journal of Chemical Information and Modeling* 54 (6): 1596–1603.
- 61 Liu, R. and Wallqvist, A. (2019). Molecular similarity-based domain applicability metric efficiently identifies out-of-domain compounds. *Journal of Chemical Information and Modeling* 59 (1): 181–189.
- 62 Sahigara, F., Ballabio, D., Todeschini, R., and Consonni, V. Defining a novel k-nearest neighbours approach to assess the applicability domain of a QSAR model for reliable predictions. *Journal of Cheminformatics* 5 (1): 27.
- 63 Aniceto, N., Freitas, A.A., Bender, A., and Ghafourian, T. (2016). A novel applicability domain technique for mapping predictive reliability across the chemical space of a QSAR: reliability-density neighbourhood. *Journal of Cheminformatics* 8: 69.
- 64 McMahan, H.B., Moore, E., Ramage, D., and y Arcas, B.A. (2016). Federated learning of deep networks using model averaging. *arXiv*, 2, 2016.
- 65 Pejó, B. (2020). The good, the bad, and the ugly: quality inference in federated learning. *arXiv*, abs/2007.06236.
- 66 Davies, R., Fowkes, A., Williams, R., and Johnston, L. (2020). Consortium-led federated QSAR models for secondary pharmacology - preparing the data. Granary Wharf House, 2 Canal Wharf, Leeds, LS11 5PS.
- 67 Matveieva, M. and Polishchuk, P. (2021). Benchmarks for interpretation of QSAR models. *Journal of Cheminformatics* 13: 41.
- 68 Lundberg, S.M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems* 30 (NIPS 2017).

- 69 Rodríguez-Pérez, R. and Bajorath, J. (2019). Interpretation of compound activity predictions from complex machine learning models using local approximations and shapley values. *Journal of Medicinal Chemistry* 63 (16): 8761–8777.
- 70 Wojtuch, A., Jankowski, R., and Podlowska, S. (2021). How can SHAP values help to shape metabolic stability of chemical compounds? *Journal of Cheminformatics* 13: 74.
- 71 Dahl, G.E., Jaitly, N., and Salakhutdinov, R. (2014). Multi-task neural networks for QSAR predictions. *arXiv*.
- 72 Xu, Y., Dai, Z., Chen, F. et al. Deep learning for drug-induced liver injury. *Journal of Chemical Information and Modeling* 55: 2085–2093.
- 73 Gawehn, E., Hiss, J.A., and Schneider, G. (2016). Deep learning in drug discovery. *Molecular Informatics* 35: 3–14.
- 74 Zhang, L., Tan, J., Han, D., and Zhu, H. (2017). From machine learning to deep learning: progress in machine intelligence for rational drug discovery. *Drug Discovery Today* 22: 1680–1685.
- 75 Chen, H., Engkvist, O., Wang, Y. et al. (2018). The rise of deep learning in drug discovery. 23 (6): 1241–1250.
- 76 Li, X., Xu, Y., Lai, L., and Pei, J. Prediction of human cytochrome P450 inhibition using a multitask deep autoencoder neural network. *Molecular Pharmaceutics* 15: 4336–4345.
- 77 Bhattacharai, B., Walters, W.P., Hop, C.E.C.A. et al. (2019). Opportunities and challenges using artificial intelligence in ADME/Tox. *Nature Materials* 18: 418–422.
- 78 Sun, M., Zhao, S., Gilvary, C. et al. Graph convolutional networks for computational drug development and discovery. *Briefings in Bioinformatics* 21 (3): 919–935.
- 79 Ma, J., Sheridan, R.P., Liaw, A. et al. (2015). Deep neural nets as a method for quantitative structure-activity relationships. *Journal of Chemical Information and Modeling* 55: 263–274.
- 80 Jiang, D., Wu, Z., Hsieh, C.Y. et al. (2021). Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. *Journal of Cheminformatics* 13: 1–23.
- 81 Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences* 28 (1): 31–36.
- 82 Krenn, M., Häse, F., Nigam, A.K. et al. (2020). Self-referencing embedded strings (SELFIES): a 100 robust molecular string representation. *Machine Learning: Science and Technology* 1 (4): 045024.
- 83 Mercado, R., Rastemo, T., Lindelöf, E. et al. (2020). Practical notes on building molecular graph generative models. *Applied AI Letters* 1 (2): <https://doi.org/10.1002/ail2.18>.
- 84 Chen, B., Fu, X., Barzilay, R., and Jaakkola, T. (2021). Fragment-based sequential translation for molecular optimization.
- 85 Bradshaw, J., Paige, B., Kusner, M.J. et al. (2020). Barking up the right tree: an approach to search over molecule synthesis dags. *CoRR*, abs/2012.11522.

- 86 Zhou, Z., Kearnes, S., Li, L. et al. (2018). Optimization of molecules via deep reinforcement learning. *CoRR*, abs/1810.08678.
- 87 Gregor, K., Danihelka, I., Mnih, A. et al. (2014). Deep autoregressive networks. *Proceedings of Machine Learning Research* 32 (2): 1242–1250.
- 88 Bank, D., Koenigstein, N., and Giryas, R. (2020). Autoencoders. *CoRR*, abs/2003.05991
- 89 Kaelbling, L.P., Littman, M.L., and Moore, A.W. (1996). Reinforcement learning: a survey. *CoRR*, cs.AI/9605103.
- 90 Frazier, P.I. (2018). A tutorial on Bayesian optimization.
- 91 Bartz-Beielstein, T., Branke, J., Mehnen, J., and Mersmann, O. (2014). Evolutionary algorithms. *WIREs Data Mining and Knowledge Discovery* 4 (3): 178–195.
- 92 Nicolaou, C.A. and Brown, N. (2013). Multi-objective optimization methods in drug design. *Drug Discovery Today: Technologies* 10 (3): e427–e435.
- 93 Greff, K., Srivastava, R.K., Koutník, J. et al. (2016). LSTM: a search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems* 28 (10): 2222–2232.
- 94 Segler, M.H.S., Kogej, T., Tyrchan, C., and Waller, M.P. (2018). Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Central Science* 4 (1): 120–131.
- 95 Williams, R.J. and Zipser, D. (1989). A learning algorithm for continually running fully recurrent neural networks. *Neural Computation* 1 (2): 270–280.
- 96 de Souza Neto, L.R., Moreira-Filho, J.T., Neves, B.J. et al. (2020). *In silico* strategies to support fragment-to-lead optimization in drug discovery. *Frontiers in Chemistry* 8: 93.
- 97 Li, Q. (2020). Application of fragment-based drug discovery to versatile targets. *Frontiers in Molecular Biosciences* 7: 180.
- 98 Zhang, G., Zhang, J., Gao, Y. et al. (2022). Strategies for targeting undruggable targets. *Expert Opinion on Drug Discovery* 17 (1): 55–69.
- 99 Penner, P., Martiny, V., Gohier, A. et al. (2020). Shape-based descriptors for efficient structure-based fragment growing. *Journal of Chemical Information and Modeling* 60 (12): 6269–6281.
- 100 Vaswani, A., Shazeer, N., Parmar, N. et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems* 30 (NIPS 2017).
- 101 Papadatos, G., Davies, M., Dedman, N. et al. (2015). SureChEMBL: a large-scale, chemically annotated patent document database. *Nucleic Acids Research* 44 (D1): D1220–D1228.
- 102 Lamb, A.M., ALIAS PARTH GOYAL, A.G., Zhang, Y. et al. (2016). Professor forcing: a new algorithm for training recurrent networks. *Advances in Neural Information Processing Systems* 29 (NIPS 2016).
- 103 Winter, R., Montanari, F., Steffen, A. et al. (2019). Efficient multi-objective molecular optimization in a continuous latent space. *Chemical Science* 10: 8016–8024.

- 104** Gómez-Bombarelli, R., Wei, J.N., Duvenaud, D. et al. (2018). Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science* 4 (2): 268–276.
- 105** Sattarov, B., Baskin, I.I., Horvath, D. et al. (2019). De novo molecular design by combining deep autoencoder recurrent neural networks with generative topographic mapping. *Journal of Chemical Information and Modeling* 59 (3): 1182–1196.
- 106** Gao, K., Nguyen, D.D., Tu, M., and Wei, G.-W. (2020). Generative network complex for the automated generation of drug-like molecules. *Journal of Chemical Information and Modeling* 60 (12): 5682–5698.
- 107** Renz, P., Van Rompaey, D., Wegner, J.K. et al. (2019). On failure modes in molecule generation and optimization. *Drug Discovery Today: Technologies* 32–33: 55–63.
- 108** Brown, N., Fiscato, M., Segler, M.H.S., and Vaucher, A.C. (2019). GuacaMol: benchmarking models for de novo molecular design. *Journal of Chemical Information and Modeling* 59 (3): 1096–1108.
- 109** Bradshaw, J., Paige, B., Kusner, M.J. et al. (2019). A model to search for synthesizable molecules. *CoRR*, abs/1906.05221.
- 110** Liu, C.-H., Korablyov, M., Jastrzebski, S. et al. (2020). RetroGNN: approximating retrosynthesis by graph neural networks for de novo drug design. *CoRR*, abs/2011.13042.
- 111** Coley, C.W., Rogers, L., Green, W.H., and Jensen, K.F. (2018). SCScore: synthetic complexity learned from a reaction corpus. *Journal of Chemical Information and Modeling* 58 (2): 252–261.
- 112** Ertl, P. and Schuffenhauer, A. (2009). Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of Cheminformatics* 1 (1): 1–11.
- 113** Thakkar, A., Chadimová, V., Bjerrum, E.J. et al. (2021). Retrosynthetic accessibility score (RAscore)—rapid machine learned synthesizability classification from AI driven retrosynthetic planning. *Chemical Science* 12: 3339–3349.
- 114** Genheden, S., Thakkar, A., Chadimová, V. et al. (2020). AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning. *Journal of Cheminformatics* 12 (1): 1–9.
- 115** Spaya. <https://spaya.ai/> (accessed 26 August 2023).
- 116** Molecule database. <https://molecule.com/database/> (accessed 26 August 2023).
- 117** Chem-space. <https://chem-space.com/> (accessed 26 August 2023).
- 118** eMolecules. <https://www.emolecules.com/> (accessed 26 August 2023).
- 119** Key Organics. <https://www.keyorganics.net/> (accessed 26 August 2023).
- 120** Parrot, M., Tajmouati, H., da Silva, V.B.R. et al. (2021). Integrating synthetic accessibility with AI-based generative drug design. *ChemRxiv*.
- 121** Marcus, G. and Davis, E. (2019). *Rebooting AI: Building Artificial Intelligence We Can Trust*. Vintage.
- 122** Collins, H. (2021). The science of artificial intelligence and its critics. *Interdisciplinary Science Reviews* 46 (1–2): 53–70.

- 123 Turk, J.-A., Gendreau, P., Drizard, N., and Gaston-Mathé, Y. (2022). A molecular assays simulator to unravel predictors hacking in goal-directed molecular generations. *ChemRxiv*.
- 124 Wise, J., de Barron, A.G., Splendiani, A. et al. (2019). Implementation and relevance of fair data principles in biopharmaceutical r&d. *Drug Discovery Today* 24 (4): 933–938.
- 125 Lhuillier-Akakpo, M., Hoffmann, B., Huu, N.D. et al. (2021). Preparing a public dataset for drug discovery. <https://www.melloddy.eu/blog/preparing-public-dataset/> (accessed 26 August 2023).
- 126 Smalley, E. (2017). Ai-powered drug discovery captures pharma interest. *Nature Biotechnology* 35 (7): 604–606.
- 127 Jiménez-Luna, J., Grisoni, F., Weskamp, N., and Schneider, G. (2021). Artificial intelligence in drug discovery: recent advances and future perspectives. *Expert Opinion on Drug Discovery* 16 (9): 949–959.
- 128 Vijayan, R.S.K., Kihlberg, J., Cross, J.B., and Poongavanam, V. (2021). Enhancing preclinical drug discovery with artificial intelligence. *Drug Discovery Today* 27 (4): 967–984.
- 129 Jiménez-Luna, J., Grisoni, F., and Schneider, G. (2020). Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence* 2 (10): 573–584.
- 130 Preuer, K., Klambauer, G., Rippmann, F. et al. (2019). Interpretable deep learning in drug discovery. In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Lecture Notes in Computer Science, vol. 11700 (ed. W. Samek, G. Montavon, A. Vedaldi, et al.), 331–345. Cham: Springer.
- 131 Luo, Y., Peng, J., and Ma, J. (2022). Next Decade's AI-based drug development features tight integration of data and computation. *Health Data Science* 2022: 9816939.

13

Reliability and Applicability Assessment for Machine Learning Models

Fabio Urbina and Sean Ekins

Collaborations Pharmaceuticals, Inc., 840 Main Campus Drive, Raleigh, NC 27606, USA

13.1 Introduction

Techniques for using small molecule structures and related physicochemical property or bioactivity data to generate computational models of different types have existed for decades (and are outlined elsewhere in this book). Over the last 10 years, we have observed an increased use of machine learning and quantitative structure-activity relationship (QSAR) across the pharmaceutical industry for a range of property predictions and virtual screening for drug discovery, lead optimization, and toxicity prediction [1, 2], which can in turn accelerate the production of new hits and drug lead candidates [3]. At the same time, there is now a wide array of accessible databases containing thousands of structure-activity datasets available for physicochemical properties, molecules screened against drug targets, or phenotypic screens in public resources like ChEMBL [4], PubChem [5, 6], or others [7]. These provide the starting points for demonstrating the application of a diverse number of machine learning methods with many classic algorithms such as k-Nearest Neighbors (kNN) [8], naïve Bayesian [9–13], decision trees [14], support vector machines (SVMs) [15–21], and others [22, 23], as well as newer algorithms such as deep neural networks (DNNs) [24–32], long short term memory (LSTM) [33], and transformers [34]. These efforts have enabled several large-scale analyses of datasets with different machine learning methods and molecular descriptors [35–42]. Some of the largest comparisons of machine learning models have used over 1000 models [43–47]. Most recently, we have described extracting over 5000 datasets from ChEMBL (endpoints such as IC_{50} , K_i , and MIC) for use with the ECFP6 fingerprint descriptor and comparing random forest, k-Nearest Neighbors, support vector classification, naïve Bayesian, AdaBoosted decision trees, and deep neural networks. The model performance was assessed using fivefold cross-validation metrics that were generated, including area-under-the-curve, F1 score, Cohen's kappa, and Matthews correlation coefficient. We demonstrated using ranked normalized scores for the metrics for all methods that they appeared comparable, while the distance from the top metric suggested our implementation of the

Bayesian algorithm and support vector classification were essentially comparable. This work represents one of the largest-scale comparisons of machine learning algorithms [48]. ECFP6 represents just one fingerprint of many that could be potentially accessed in this manner [49]. Another large-scale evaluation performed by the Novartis Institute for Biomedical Research used 8558 proprietary Novartis assays to generate Random Forest Regressor models [50].

13.2 Challenges for Modeling

Such machine learning models are not perfect, they are “an approximation of reality.” The models may have errors inherent in the data, as biological data that are generated *in vitro* will have experimental errors resulting from dispensing, analytical or other errors [51]. Other issues may include small dataset size, lack of diversity, poor activity distribution of the data, or other biases [7]. These experimental errors will impact the dataset that is being modeled and could affect the resulting predictions. Also, the training set for the model may be of a limited size, which will impact the diversity of the chemical property space covered, the data distribution, and the utility of the model to make predictions outside of the training set space. An early concept was that more diverse models could be considered as “global,” whereas those with a very limited focus on a narrow SAR were likely to describe “local” properties [2]. Hence, a machine learning model’s prediction confidence will be impacted by its domain of applicability (also called domain extrapolation) outside of this training domain [52]. An early demonstration of the calculation of prediction confidence and domain extrapolation used the estrogen receptor binding activity datasets and decision forest models. In this case, the prediction accuracy was related to the ratio of correct predictions to the total number of molecules in the domain and showed that accuracy declined sharply with domain extrapolation [52] (Table 13.1). Other groups have suggested how the initial poor performance of ADME-Tox models could have been related to their applicability domain. Several methods were described early on that related to the calculation of the applicability domain using either the descriptor space (missing fragments approach) or methods based on the similarity of molecules in descriptor space. The latter uses different methods such as Euclidean, city block, Tanimoto, Mahalanobis, hotelling T2, and leverage to measure the distance from a training set to a test set [56, 65, 66] to compare the chemical space of the datasets [2]. The applicability domain was also proposed early on as an important problem for QSAR studies, and a review described several limitations of the methods employed [67]. Pharmaceutical companies have applied applicability domains early on. For example, Bayer and Schering used several approaches, including Bayesian Gaussian process models, distance-based methods, and ensembles for regression models of lipophilicity [57] and solubility [68] that showed that the mean absolute error decreased as compounds were binned with higher confidence. For a visual appreciation of the applicability domain concept, the reader is also referred to these published articles and others, such as Aniceto et al. [53].

Table 13.1 Selected methods for applicability domain, error, and confidence predictions. See also additional articles in the text Adapted from Aniceto et al., 2016; Rakhimbekova et al., 2020; Sushko 2011.

Method	Reference
Domain extrapolation	[52]
Prediction confidence	[52]
Euclidean distance	[56]
City block	[56]
Mahalanobis	[56]
Hotelling T^2	[56]
Leverage	[56]
Bayesian Gaussian process models, distance-based methods, and ensembles	[57]
Optimal assignment kernel, flexible optional assignment kernel, marginalized graph kernel	[58]
Number of fingerprint features	[59]
A reliability-density neighborhood	[53]
Sum of distance-weighted contributions	[60]
Conformal prediction	[61]
Test time dropout	[62]
Rivality index	[63]
Entropy, Monte Carlo dropout, Multi-Initial, FPsDist, and LatentDist	[64]

As SVMs have been widely used for QSAR, one study has proposed three applicability domain approaches for kernel-based QSAR relying on similarity: the optimal assignment kernel, the flexible optional assignment kernel, and the marginalized graph kernel. Using three different virtual screening examples, these showed the models performed better inside the domains [58]. As molecular fingerprints such as ECFP are widely used in machine learning, they have also been used to define applicability domains. One study used the nearest neighbor or random forests in combination to provide an applicability domain for an Ames mutagenicity model. The number of ECFP_4 or ECFP_2 features for a test compound that is not present in a training set was used as an indicator of the applicability domain [59]. Several different methods have also been used to assess machine learning model reliability with 20 regression model datasets. These reliability estimates included Mahalanobis distance to nearest neighbors, Mahalanobis distance to the data set center, sensitivity analysis scores, bootstrap variance, local cross-validation error, local prediction error modeling, and combination of bootstrap variance and local prediction error score. Error-based estimation methods outperformed or were on a par with similarity-based methods, while performance did not depend on global or local model or descriptor type [69]. A reliability-density neighborhood approach was used as an applicability domain for P-gp, Ames, and CYP450 models and was proposed to take into account sparse regions by mapping data density and local

precision and bias [53]. A new applicability domain metric that considered the contribution of every training sample weighted by its distance to the molecule being predicted (called sum of distance-weighted contributions) was demonstrated with several toxicities and physicochemical property datasets and correlated more strongly with prediction error than other methods like distance to model or ensemble variance measures [60]. This approach has also been used with a melting point dataset and showed that it outperformed the other methods utilized [70]. Several approaches have been proposed to compute the uncertainty of predictions [71], and one is conformal prediction (also see later section), which can provide confidence regions and was used with deep neural networks and benchmarked on 24 regression datasets from ChEMBL as well as against random forest-based conformal predictions. The confidence intervals for the deep confidence approach had a smaller spread than for the random forest approach [61]. The test time dropout and conformal prediction approaches have been used to reliably compute errors for neural network models created for the same 24 bioactivity datasets [62]. Conformal prediction has also been the subject of a minireview, which also applied the approach with three transporter models [72].

The rivalry index has been proposed as another method for assessing the reliability of predictions or applicability domains by generating a normalized distance measurement between each molecule and its nearest neighbor belonging to the same class and the nearest neighbor belonging to a different class. This approach was tested with four classification datasets across 12 algorithms [63]. Uncertainty estimation using five methods, including Entropy, Monte Carlo dropout, Multi-Initial, FPsDist, and LatentDist, was used with a BBB dataset and several different machine learning approaches. The combination of Entropy and Monte Carlo dropout to predict uncertainty was used for the GROVER BBB model [64].

While these represent just a snapshot of some of the many efforts to address the applicability domain or confidence in prediction, these areas are not always covered in exhaustive reviews describing machine learning or QSAR methods [73]. It is for this reason that they should perhaps be given more exposure. We now provide several examples from our own work to explore this further.

13.3 Example 1: BBB Applicability Domain Comparison

The blood brain barrier (BBB) represents a significant challenge for drug delivery to the central nervous system (CNS). Molecules (such as antidepressants and antipsychotics, for example) must cross the BBB to act within the CNS or the brain. We recently published a BBB model based on a binary dataset of 2358 published compounds that either crossed the brain barrier or did not [74, 75]. To illustrate the use of applicability domains, we have investigated a modified reliability-density neighborhood approach. We compare this approach to simple training set distance metrics as baselines.

Machine learning: Our software, Assay Central, uses multiple algorithms integrated into web-based software to build models, as described previously in

detail [76]. The machine learning model validation was performed using a nested fivefold cross-validation with an external test set. This optimized model is then used to predict the initial 20% hold-out set. The final nested fivefold cross-validation scores are an average of each of the holdout set metrics. We chose a random forest model to investigate as it was among the top-performing models, and a standard deviation can be extracted by aggregating the predictions from the individual trees.

Modified reliability-density (RDE) neighborhood estimation: We adopted the approach from Aniceto et al. [53]. As our model uses ECFP6 descriptors, we simplified the AD score to the following: After the random forest model is built on the training data, it is then used to predict on the same training set to extract $\text{bias} = \text{abs}(\hat{Y}_i - Y_i)$, as well as the standard deviation of the individual tree predictions. Upon inference with a new molecule, the applicability domain is applied using the following equation:

$$\text{AD score} = w_T * w_{AD}$$

where w_T is the average Tanimoto similarity between the top- n most similar molecule(s) in the training dataset and the molecule for inference and $w_{AD} = \left(\frac{\sum (1 - \text{STD}_{i,j})}{n} \right) * \left(\frac{\sum (1 - \text{abs}(\hat{y}_i - y_i))}{n} \right)$ for the top- n most similar molecules in the model. As the Tanimoto distance is performed on the same ECFP6 1024-bit vector that informs the model, the maximum Tanimoto similarities highlight how informed the model is of the input features while w_{ADi} penalizing the similarity score based on the bias and decision tree agreement of the model on the most similar molecules. We performed two iterations, using n of 1 and n of 5 for the modified RDE calculation. We compare this method by using the average Tanimoto similarity to the training data, the maximum Tanimoto similarity to the training data, and the top-5 average Tanimoto similarities to the dataset as AD score baselines. We perform this comparison using stratified fivefold cross-validation on the BBB dataset and fit a locally estimated scatterplot smoothing (LOESS) regression model.

Results: While there is no correlation between the average Tanimoto similarity to the dataset and the absolute error, both the maximum Tanimoto and the top-5 Tanimoto distance to the training set show a small but non-robust correlation (Figure 13.1). The modified RDE method of weighting the maximum or top-5 maximum Tanimoto similarity shows a stronger correlation to the absolute error that is more consistent between folds and closer to a strictly-decreasing function. Evaluating the top $X\%$ of AD scores shows that all methods enrich correct predictions while the average Tanimoto lags significantly (Figure 13.2). The RDE top-5 method shows the most consistent enrichment, suggesting the corrective weighting factor helps reduce the influence of model bias.

13.4 Example 2: Models for Uncertainty Estimation for Multitask Toxicity Predictions

Many applicability domain scores have restrictions, such as being only applicable to classification tasks or specific feature inputs. While applicability domains are usually

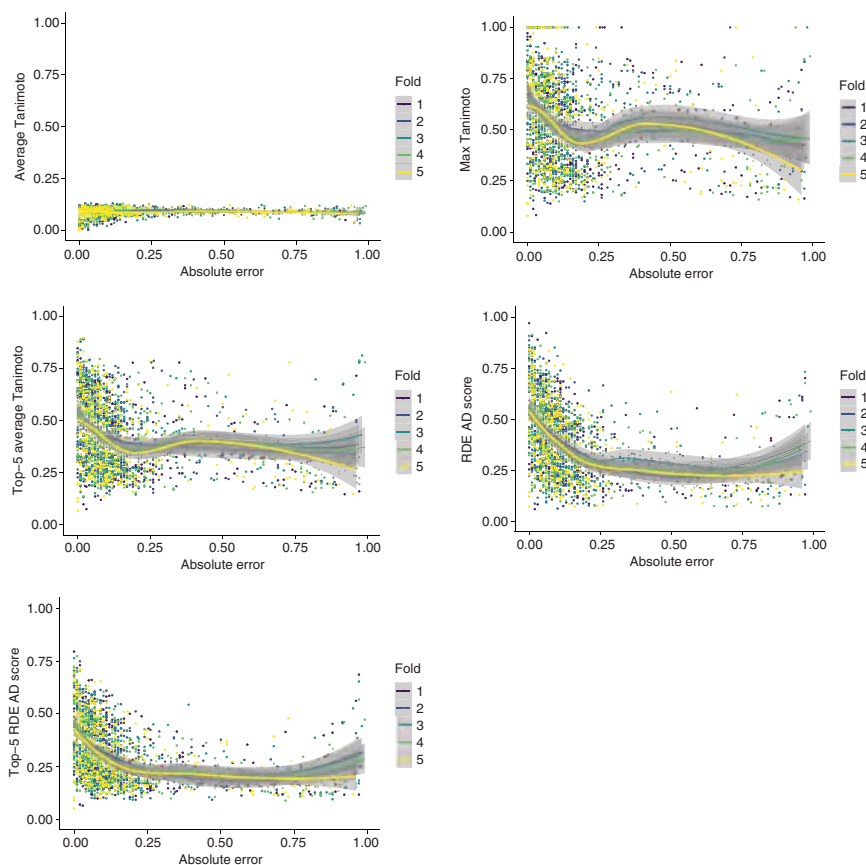


Figure 13.1 Applicability score vs. absolute error test set scores for stratified 5-cross-validation of a binary classification blood brain barrier random forest model (8850 actives/2580 inactives). Each fold's test set had 2286 molecules. Average Tanimoto: average Tanimoto against the training set, Max Tanimoto: Maximum Tanimoto against the training set. Top-5 Average: Top 5 Tanimoto distances against the training set averaged. RDE AD: Top-1 Reliability-density neighborhood estimation. Top-5 RDE: Top-5 Reliability-density neighborhood estimation.

investigated for single-endpoint models, the advent of larger and more complex deep learning models may require a new set of applicability domain algorithms to better represent their predictive performance. Here, we investigate the use of Monte Carlo (MC) dropout for uncertainty estimation in the predictions of a deep learning end-to-end multitask regression model.

Overview: The requirements for a candidate molecule to become a drug at a simplistic level include efficacy and specificity against a target of interest, as off-target interactions can lead to undesirable side effects and pose a potential safety hazard. Commercial *in vitro* safety profiling screens are often used to search for critical off-targets, which could lead to adverse drug reactions [77]. *In silico* approaches offer an appealing substitute, as they are comparably inexpensive and inference is

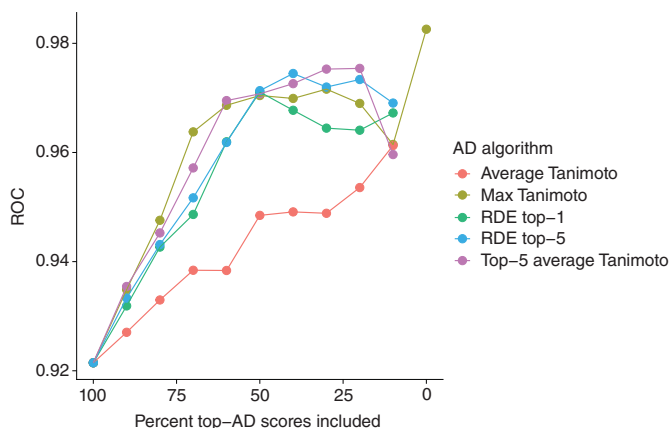


Figure 13.2 Comparison of ROC of applicability domain (AD) score inclusion. The top $X\%$ of molecules have AD scores and are evaluated at each interval.

significantly faster than *in vitro* assays. Thus, much research has been performed over several decades into predicting toxicity.

Especially crucial to toxicity models is how trustworthy predictions are, as false negatives are not tolerated in safety profiling. We recently built and described a multi-task neural network model to predict IC_{50} 's of 42 of the 44 endpoints used in the SafetyScreen44™, a commercial *in vitro* safety profiling assay for which we could find sufficient publicly available datasets [78]. The purpose of this machine learning model was to increase inference speed by utilizing SMILES as the only input, removing the temporal bottleneck of feature creation (e.g. generating Morgan fingerprints). As a further useful case study for applicability domains, we now revisit this model and investigate using MC dropout to approximate uncertainty estimation for the model.

Datasets: The data was curated as described in Ref. [78]. Briefly, IC_{50} -only target-activity data for 42 toxicity targets were downloaded from ChEMBL 30 and standardized (salts removed, charges neutralized, and canonical SMILES generated). The datasets were split randomly at 70%/15%/15% and stratified for each target. Seventy percent of the data was used for training, 15% for validation, and 15% for test results.

Machine Learning: We used a convolutional long-short term memory (ConvLSTM)-based model with an embedding layer (size 50), 1-D convolutional layer (size 256), a batch-norm layer, a bidirectional LSTM layer (size 1024), and three fully-connected layers with dropout (25%) followed by rectified linear unit (ReLU) layer of size 2048, 1024, and a final 42 for the output layer.

Results: During model training, dropout layers are used for model regularization [79]. These layers are generally turned off during inference so that the full model can be utilized in a deterministic manner. Dropout layers can be utilized during inference, however, to approximate Bayesian model uncertainty without alterations to the final model by using dropout layers during inference [80]. Running multiple predictions with different neurons due to dropout is equivalent to an ensemble of models performing inference. The predictions are then averaged for

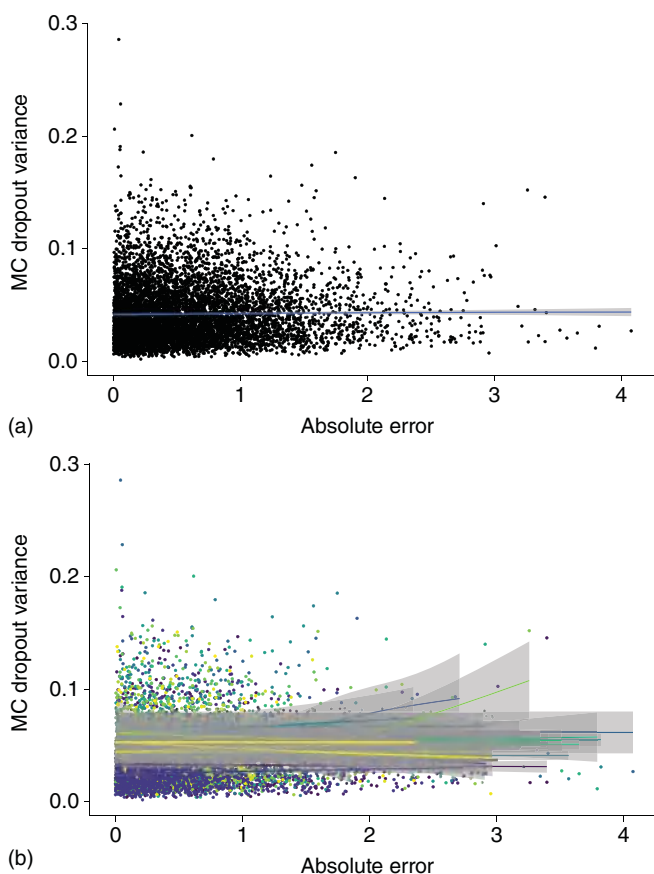


Figure 13.3 MC dropout variance vs. Absolute error for prediction on a test set using a multitask regression model (LSTM model using 42 toxicity models). Data fit using Generative additive models (GAM). (a) MC Dropout with a GAM fit on the entire test set. (b) MC Dropout with a GAM fit on each of the 42 individual target endpoints.

a final prediction. The variation of predictions gives an estimation of uncertainty: Predictions with high variation can be inferred as high uncertainty. Applying this technique, we performed MC dropout inference (25 predictions per test set datapoint) and calculated the standard deviation of each set of predictions (Figure 13.3). Surprisingly, while the variance correlated with several independent target predictions, the variance of predictions did not correlate with a significant number of endpoints or the predictive error of the entire test set (Figure 13.3). Altering the dropout rate (20–50%), number of neural network layers with dropout (1–3), or number of predictions (25–100) did not change the relationship between variance and predictive performance. This suggests that MC dropout may not be generally applicable to all scenarios, and care must be taken when selecting an applicability domain for more complex models.

13.5 Example 3: Class-Conditional Conformal Predictors

Conformal predictors are a framework that can be applied to any model that outputs prediction scores and have been used for various drug discovery and environmental QSAR applications [72, 81, 82]. A recent extensive review [83] describes how, in simple terms, the framework uses a calibration dataset to determine the optimal prediction score threshold for each class (1, 0) such that prediction scores that exceed the class-specific threshold have the following validity: $1 - \alpha \leq P\left(Y_{\text{test}} \in C(X_{\text{test}}) \leq 1 - \alpha + \frac{1}{n-1}\right)$ for a user-chosen value of $\alpha \in [0, 1]$.

More formally, let k be the number of classes predicted by a model $\hat{f}(x) \in [0, 1]^k$. Let $(x_i, y_i)_{i=1}^n$ be an independent calibration set, assumed I.I.D., and from the same distribution as the training data. Define a conformal score to be $S_i = 1 - \hat{f}(x_i)_{y_i}$ of the calibration set and $\hat{q}_k = \frac{[(n+1)(1-\alpha)]}{n}$ as the empirical quantile of the conformal scores $S_{k_1}, \dots, S_{k_i}$ for each k . For any new predictions, return the set of possible classes $C(X_{\text{test}}) = \{y : \hat{f}(X_{\text{test}})_k \geq 1 - \hat{q}_k\}$. In the binary case, return the prediction score for class 0 if $1 - \hat{f}(X_{\text{test}}) \geq 1 - \hat{q}_0$ and the prediction score for class 1 if $\hat{f}(X_{\text{test}}) \geq 1 - \hat{q}_1$. As a heuristic, we may treat the return of no prediction scores or prediction scores for both classes to be inconclusive and “trust” only single-class predictions that are returned.

Evaluation: Using a two-class biodegradation dataset as an example. The two classes are “not-readily biodegradable” and “readily biodegradable.” Following the concatenation, these datasets were cleaned (subjected to charge neutralization, salt removal, and standardization via custom software using open-source RDkit functions. Following these, the duplicates were then removed after a check for unambiguous classification) prior to model building. The final dataset contained 3428 unique compounds (inorganic compounds removed), with 962 classified as “readily biodegradable.” Using a class-stratified random 70/20/10% split for the training, test, and calibration sets, respectively, we then built a Random Forest model with 1000 trees. The initial model had a recall of 0.72 and a ROC of 0.86; however, it had a lower precision of 0.63 (Figure 13.4a). We next calculated the thresholds at different levels of α using the calibration dataset. Predictions that did not meet the required threshold of either class (no predictions returned) or were predicted to belong to both classes (both prediction scores returned) were labeled “inconclusive.” Only conclusive predictions were used for metric calculations. Stricter levels of alpha classified more of the test set as “inconclusive” (Figure 13.4b), but enriched correct predictions, up to an $\alpha = 0.2$, where the ROC rose to 0.9, Recall to 0.89, and Precision to 0.86, coinciding with the validity guarantee of $1 - \alpha$ accuracy for each class. As outlined in this example and as has been shown earlier by others, this methodology therefore allows for stricter confidence in returned predictions and rejection of prediction scores that do not meet the designated threshold. This therefore has utility as a method for applicability domain prediction.

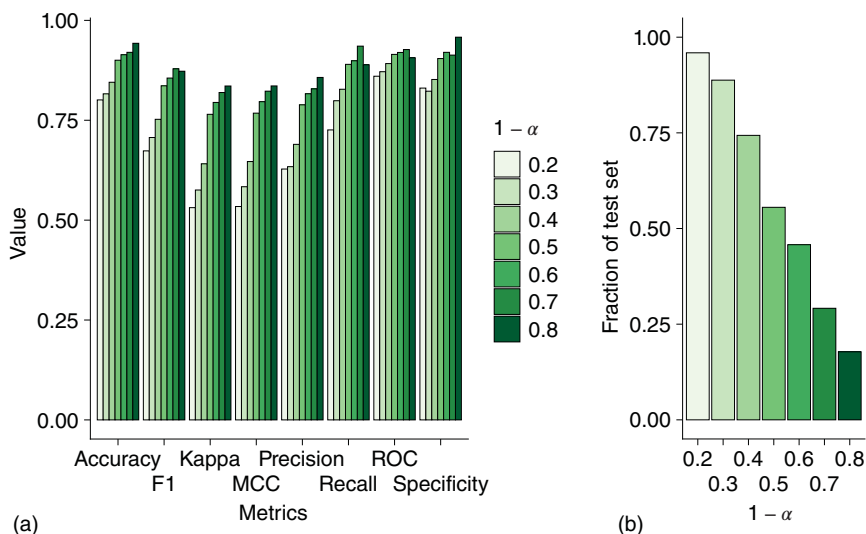


Figure 13.4 (a) Metrics of a trained random forest model on a test set using different α thresholds to classify test molecules as either not-readily biodegradable, readily biodegradable, or inconclusive. Metrics were calculated on non-inconclusive data points. (b) The fraction of the test set that was considered not inconclusive.

13.6 Conclusions

Applicability domains are becoming a de facto requirement before taking action on “black box” model predictions. Future challenges include that generative models that include property predictions to derive molecules with ideal properties may be limited by the applicability domain or model confidence in predictions [7, 84]. This is not an area that has been broadly addressed to date. Many applicability domains rely on a similarity measure to the training data, while generative models are often attempting to discover novel molecules, putting the use of such applicability domains at odds with molecule generation [84]. Another challenge is that while many applicability domains have been introduced and used effectively, their performance is not always significantly better than using the model’s prediction scores alone [85]. When coupled with the sometimes-strict requirements and complexity to calculate AD scores, a need for more effective AD algorithms with less strict requirements and more efficacy, is therefore highly desirable. Further, as we show, many AD algorithms are not readily applicable to more complex models such as multitask models or models with distinct inputs. There is clearly still considerable scope to both evaluate existing AD methods and develop new ones for the reliability and applicability assessment of machine learning models. Beyond applicability, the next important area will be in the interpreting or explainability of the model predictions, and there have already been recent discussions on this topic [86–88]. Perhaps in 20 years, model interpretability/explainability will be at a similar level of acceptance as a model applicability domain is today. Ultimately, the key will be

the transition of these concepts into commercial or widely used software products, by which point they will be regarded as standard.

Funding

We kindly acknowledge NIH funding from R44GM122196-02A1 from NIGMS and 2R44ES031038-02A1 from NIEHS. Research reported in this publication was supported by the National Institute of Environmental Health Sciences of the National Institutes of Health under Award Number 2R44ES031038-02A1. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.”

Competing Interests

S.E. is owner, and F.U. is an employee of Collaborations Pharmaceuticals, Inc.

References

- 1 Ekins, S., Puhl, A.C., Zorn, K.M. et al. (2019). Exploiting machine learning for end-to-end drug discovery and development. *Nat. Mater.* 18 (5): 435–441.
- 2 Cheng, F., Li, W., Liu, G., and Tang, Y. (2013). In silico ADMET prediction: recent advances, current challenges and future trends. *Curr. Top. Med. Chem.* 13 (11): 1273–1289.
- 3 Zhavoronkov, A., Ivanenkov, Y.A., Aliper, A. et al. (2019). Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.* 37 (9): 1038–1040.
- 4 Gaulton, A., Bellis, L.J., Bento, A.P. et al. (2012). ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 40 (Database issue): D1100–D1107.
- 5 Kim, S., Thiessen, P.A., Bolton, E.E. et al. (2016). PubChem substance and compound databases. *Nucleic Acids Res.* 44 (D1): D1202–D1213.
- 6 Anon *The PubChem Database*. <http://pubchem.ncbi.nlm.nih.gov>.
- 7 Nigam, A., Pollice, R., Hurley, M.F.D. et al. (2021). Assigning confidence to molecular property prediction. *Expert Opin. Drug Discovery* 16 (9): 1009–1023.
- 8 Shen, M., Xiao, Y., Golbraikh, A. et al. (2003). Development and validation of k-nearest neighbour QSPR models of metabolic stability of drug candidates. *J. Med. Chem.* 46: 3013–3020.
- 9 Wang, S., Sun, H., Liu, H. et al. (2016). ADMET evaluation in drug discovery. 16. Predicting hERG blockers by combining multiple pharmacophores and machine learning approaches. *Mol. Pharmaceutics* 13 (8): 2855–2866.
- 10 Li, D., Chen, L., Li, Y. et al. (2014). ADMET evaluation in drug discovery. 13. Development of in silico prediction models for P-glycoprotein substrates. *Mol. Pharmaceutics* 11 (3): 716–726.

- 11 Nidhi, Glick, M., Davies, J.W., and Jenkins, J.L. (2006). Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *J. Chem. Inf. Model.* 46 (3): 1124–1133.
- 12 Azzaoui, K., Hamon, J., Faller, B. et al. (2007). Modeling promiscuity based on in vitro safety pharmacology profiling data. *ChemMedChem* 2 (6): 874–880.
- 13 Bender, A., Scheiber, J., Glick, M. et al. (2007). Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure. *ChemMedChem* 2 (6): 861–873.
- 14 Susnow, R.G. and Dixon, S.L. (2003). Use of robust classification techniques for the prediction of human cytochrome P450 2D6 inhibition. *J. Chem. Inf. Comput. Sci.* 43 (4): 1308–1315.
- 15 Bennet, K.P. and Campbell, C. (2000). Support vector machines: hype or hallelujah? *SIGKDD Explor.* 2: 1–13.
- 16 Christianini, N. and Shawe-Taylor, J. (2000). *Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, MA: Cambridge University Press.
- 17 Chang, C.C. and Lin, C.J. (2011). LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* 2 (3): 1–27.
- 18 Lei, T., Chen, F., Liu, H. et al. (2017). ADMET evaluation in drug discovery. Part 17: development of quantitative and qualitative prediction models for chemical-induced respiratory toxicity. *Mol. Pharmaceutics* 14 (7): 2407–2421.
- 19 Kriegl, J.M., Arnhold, T., Beck, B., and Fox, T. (2005). A support vector machine approach to classify human cytochrome P450 3A4 inhibitors. *J. Comput.-Aided Mol. Des.* 19 (3): 189–201.
- 20 Guangli, M. and Yiyu, C. (2006). Predicting Caco-2 permeability using support vector machine and chemistry development kit. *J. Pharm. Pharm. Sci.* 9 (2): 210–221.
- 21 Kortagere, S., Chekmarev, D., Welsh, W.J., and Ekins, S. (2009). Hybrid scoring and classification approaches to predict human pregnane X receptor activators. *Pharm. Res.* 26 (4): 1001–1011.
- 22 Mitchell, J.B. (2014). Machine learning methods in chemoinformatics. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 4 (5): 468–481.
- 23 Wacker, S. and Noskov, S.Y. (2018). Performance of machine learning algorithms for qualitative and quantitative prediction drug blockade of hERG1 channel. *Comput. Toxicol.* 6: 55–63.
- 24 Schmidhuber, J. (2015). Deep learning in neural networks: an overview. *Neural Netw.* 61: 85–117.
- 25 Capuzzi, S.J., Politi, R., Isayev, O. et al. (2016). QSAR modeling of Tox21 challenge stress response and nuclear receptor signaling toxicity assays. *Front. Environ. Sci.* 4 (3).
- 26 Russakovsky, O., Deng, J., Su, H., et al. (2015) ImageNet Large Scale Visual Recognition Challenge. <https://arxiv.org/pdf/1409.0575.pdf>.
- 27 Zhu, H., Zhang, J., Kim, M.T. et al. (2014). Big data in chemical toxicity research: the use of high-throughput screening assays to identify potential toxicants. *Chem. Res. Toxicol.* 27 (10): 1643–1651.

- 28 Clark, A.M. and Ekins, S. (2015). Open source Bayesian models: 2. Mining a “big dataset” to create and validate models with ChEMBL. *J. Chem. Inf. Model.* 55: 1246–1260.
- 29 Ekins, S., Clark, A.M., Swamidass, S.J. et al. (2014). Bigger data, collaborative tools and the future of predictive drug discovery. *J. Comput.-Aided Mol. Des.* 28 (10): 997–1008.
- 30 Ekins, S., Freundlich, J.S., and Reynolds, R.C. (2014). Are bigger data sets better for machine learning? Fusing single-point and dual-event dose response data for *Mycobacterium tuberculosis*. *J. Chem. Inf. Model.* 54: 2157–2165.
- 31 Ekins, S. (2016). The next era: deep learning in pharmaceutical research. *Pharm. Res.* 33 (11): 2594–2603.
- 32 Baskin, I.I., Winkler, D., and Tetko, I.V. (2016). A renaissance of neural networks in drug discovery. *Expert Opin. Drug Discovery* 11: 785–795.
- 33 Greff, K., Srivastava, R.K., Koutník, J. et al. (2017). LSTM: a search space odyssey. *IEEE Trans. Neural Netw. Learn. Syst.* 28 (10): 2222–2232.
- 34 Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv, 1810.04805.
- 35 Wang, L., Ma, C., Wipf, P. et al. (2013). TargetHunter: an in silico target identification tool for predicting therapeutic potential of small organic molecules based on chemogenomic database. *AAPS J.* 15 (2): 395–406.
- 36 Koutsoukas, A., Lowe, R., Kalantarmotamedi, Y. et al. (2013). In silico target predictions: defining a benchmarking data set and comparison of performance of the multiclass Naive Bayes and Parzen-Rosenblatt window. *J. Chem. Inf. Model.* 53 (8): 1957–1966.
- 37 Cortes-Ciriano, I. (2016). Benchmarking the predictive power of ligand efficiency indices in QSAR. *J. Chem. Inf. Model.* 56 (8): 1576–1587.
- 38 Qureshi, A., Kaur, G., and Kumar, M. (2017). AVCpred: an integrated web server for prediction and design of antiviral compounds. *Chem. Biol. Drug Des.* 89 (1): 74–83.
- 39 Bieler, M., Reutlinger, M., Rodrigues, T. et al. (2016). Designing multi-target compound libraries with Gaussian process models. *Mol. Inf.* 35 (5): 192–198.
- 40 Huang, T., Mi, H., Lin, C.Y. et al., and for MZRW Group (2017). MOST: most-similar ligand based approach to target prediction. *BMC Bioinf.* 18 (1): 165.
- 41 Cortes-Ciriano, I., Firth, N.C., Bender, A., and Watson, O. (2018). Discovering highly potent molecules from an initial set of inactives using iterative screening. *J. Chem. Inf. Model.* 58 (9): 2000–2014.
- 42 Bosc, N., Atkinson, F., Felix, E. et al. (2019). Large scale comparison of QSAR and conformal prediction methods and their applications in drug discovery. *J. Cheminf.* 11 (1): 4.
- 43 Lenselink, E.B., Ten Dijke, N., Bongers, B. et al. (2017). Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. *J. Cheminf.* 9 (1): 45.

- 44 Mayr, A., Klambauer, G., Unterthiner, T. et al. (2018). Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chem. Sci.* 9 (24): 5441–5451.
- 45 Lee, K. and Kim, D. (2019). In-silico molecular binding prediction for human drug targets using deep neural multi-task learning. *Genes (Basel)* 10 (11): 906.
- 46 Awale, M. and Reymond, J.L. (2019). Polypharmacology browser PPB2: target prediction combining nearest neighbors with machine learning. *J. Chem. Inf. Model.* 59 (1): 10–17.
- 47 Škuta, C., Cortés-Ciriano, I., Dehaen, W. et al. (2020). QSAR-derived affinity fingerprints (part 1): fingerprint construction and modeling performance for similarity searching, bioactivity classification and scaffold hopping. *J. Cheminf.* 12 (1): 39.
- 48 Lane, T.R., Foil, D.H., Minerali, E. et al. (2021). Bioactivity comparison across multiple machine learning algorithms using over 5000 datasets for drug discovery. *Mol. Pharmaceutics* 18 (1): 403–415.
- 49 Clark, A.M., Dole, K., Coulon-Spektor, A. et al. (2015). Open source Bayesian models. 1. Application to ADME/Tox and drug discovery datasets. *J. Chem. Inf. Model.* 55 (6): 1231–1245.
- 50 Martin, E.J., Polyakov, V.R., Zhu, X.W. et al. (2019). All-assay-Max2 pQSAR: activity predictions as accurate as four-concentration IC50s for 8558 Novartis assays. *J. Chem. Inf. Model.* 59 (10): 4450–4459.
- 51 Ekins, S., Olechno, J., and Williams, A.J. (2013). Dispensing processes impact apparent biological activity as determined by computational and statistical analyses. *PLoS One* 8 (5): e62325.
- 52 Tong, W., Xie, Q., Hong, H. et al. (2004). Assessment of prediction confidence and domain extrapolation of two structure-activity relationship models for predicting estrogen receptor binding activity. *Environ. Health Perspect.* 112 (12): 1249–1254.
- 53 Aniceto, N., Freitas, A.A., Bender, A., and Ghafourian, T. (2016). A novel applicability domain technique for mapping predictive reliability across the chemical space of a QSAR: reliability-density neighbourhood. *J. Cheminf.* 8 (1): 69.
- 54 Rakhimbekova, A., Madzhidov, T.I., Nugmanov, R.I. et al. (2020). Comprehensive analysis of applicability domains of QSPR models for chemical reactions. *Int. J. Mol. Sci.* 21 (15): 5542.
- 55 Sushko, I. (2011). *Applicability Domain of QSAR Models*. Technische Universität München.
- 56 Tetko, I.V., Bruneau, P., Mewes, H.W. et al. (2006). Can we estimate the accuracy of ADME-Tox predictions? *Drug Discovery Today* 11 (15–16): 700–707.
- 57 Schroeter, T., Schwaighofer, A., Mika, S. et al. (2007). Machine learning models for lipophilicity and their domain of applicability. *Mol. Pharmaceutics* 4 (4): 524–538.
- 58 Fechner, N., Jahn, A., Hinselmann, G., and Zell, A. (2010). Estimation of the applicability domain of kernel-based machine learning models for virtual screening. *J. Cheminf.* 2 (1): 2.

- 59 Liu, R. and Wallqvist, A. (2014). Merging applicability domains for in silico assessment of chemical mutagenicity. *J. Chem. Inf. Model.* 54 (3): 793–800.
- 60 Liu, R., Glover, K.P., Feasel, M.G., and Wallqvist, A. (2018). General approach to estimate error bars for quantitative structure-activity relationship predictions of molecular activity. *J. Chem. Inf. Model.* 58 (8): 1561–1575.
- 61 Cortes-Ciriano, I. and Bender, A. (2019). Deep confidence: a computationally efficient framework for calculating reliable prediction errors for deep neural networks. *J. Chem. Inf. Model.* 59 (3): 1269–1281.
- 62 Cortes-Ciriano, I. and Bender, A. (2019). Reliable prediction errors for deep neural networks using test-time dropout. *J. Chem. Inf. Model.* 59 (7): 3330–3339.
- 63 Luque Ruiz, I. and Gomez-Nieto, M.A. (2019). Building of robust and interpretable QSAR classification models by means of the rivalry index. *J. Chem. Inf. Model.* 59 (6): 2785–2804.
- 64 Tong, X., Wang, D., Ding, X. et al. (2022). Blood-brain barrier penetration prediction enhanced by uncertainty estimation. *J. Cheminf.* 14 (1): 44.
- 65 Nikolova-Jeliazkova, N. and Jaworska, J. (2005). An approach to determining applicability domains for QSAR group contribution models: an analysis of SRC KOWWIN. *Altern. Lab Anim.* 33 (5): 461–470.
- 66 Jaworska, J., Nikolova-Jeliazkova, N., and Aldenberg, T. (2005). QSAR applicability domain estimation by projection of the training set descriptor space: a review. *Altern. Lab Anim.* 33 (5): 445–459.
- 67 Tropsha, A. and Golbraikh, A. (2007). Predictive QSAR modeling workflow, model applicability domains, and virtual screening. *Curr. Pharm. Des.* 13 (34): 3494–3504.
- 68 Schroeter, T.S., Schwaighofer, A., Mika, S. et al. (2007). Estimating the domain of applicability for machine learning QSAR models: a study on aqueous solubility of drug discovery molecules. *J. Comput.-Aided Mol. Des.* 21 (12): 651–664.
- 69 Toplak, M., Mocnik, R., Polajnar, M. et al. (2014). Assessment of machine learning reliability methods for quantifying the applicability domain of QSAR regression models. *J. Chem. Inf. Model.* 54 (2): 431–441.
- 70 Liu, R. and Wallqvist, A. (2019). Molecular similarity-based domain applicability metric efficiently identifies out-of-domain compounds. *J. Chem. Inf. Model.* 59 (1): 181–189.
- 71 Mervin, L.H., Johansson, S., Semenova, E. et al. (2021). Uncertainty quantification in drug design. *Drug Discovery Today* 26 (2): 474–489.
- 72 Alvarsson, J., Arvidsson McShane, S., Norinder, U., and Spjuth, O. (2021). Predicting with confidence: using conformal prediction in drug discovery. *J. Pharm. Sci.* 110 (1): 42–49.
- 73 Mao, J., Akhtar, J., Zhang, X. et al. (2021). Comprehensive strategies of machine-learning-based quantitative structure-activity relationship models. *iScience* 24 (9): 103052.
- 74 Wang, Z., Yang, H., Wu, Z. et al. (2018). In silico prediction of blood-brain barrier permeability of compounds by machine learning and resampling methods. *ChemMedChem* 13 (20): 2189–2201.

- 75 Urbina, F., Zorn, K.M., Brunner, D., and Ekins, S. (2021). Comparing the Pfizer central nervous system multiparameter optimization calculator and a BBB machine learning model. *ACS Chem. Neurosci.* 12 (12): 2247–2253.
- 76 Lane, T., Russo, D.P., Zorn, K.M. et al. (2018). Comparing and validating machine learning models for *Mycobacterium tuberculosis* drug discovery. *Mol. Pharmaceutics* 15 (10): 4346–4360.
- 77 Bowes, J., Brown, A.J., Hamon, J. et al. (2012). Reducing safety-related drug attrition: the use of in vitro pharmacological profiling. *Nat. Rev. Drug Discovery* 11 (12): 909–922.
- 78 Blay, V., Li, X., Gerlach, J. et al. (2022). Combining DELs and machine learning for toxicology prediction. *Drug Discovery Today* 27 (11): 103351.
- 79 Srivastava, N., Hinton, G., Krizhevsky, A. et al. (2014). Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15: 1929–1958.
- 80 Gal, Y. and Ghahramani, Z. (2015). Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning.
- 81 Norinder, U. and Boyer, S. (2016). Conformal prediction classification of a large data set of environmental chemicals from ToxCast and Tox21 estrogen receptor assays. *Chem. Res. Toxicol.* 29 (6): 1003–1010.
- 82 Fagerholm, U., Hellberg, S., Alvarsson, J. et al. (2021). In silico prediction of volume of distribution of drugs in man using conformal prediction performs on par with animal data-based models. *Xenobiotica* 51 (12): 1366–1371.
- 83 Angelopoulou, A.N. and Bates, S. (2021). *A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification*. arXiv:2107.07511.
- 84 Langevin, M., Grebner, C., Guessregen, S. et al. (2022). Impact of applicability domains to generative artificial intelligence. *ChemRxiv*.
- 85 Klingspohn, W., Mathea, M., Ter Laak, A. et al. (2017). Efficiency of different measures for defining the applicability domain of classification models. *J. Chem-inf.* 9 (1): 44.
- 86 Lundberg, S.M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems*.
- 87 Murdoch, W.J., Singh, C., Kumbier, K. et al. (2019). Definitions, methods, and applications in interpretable machine learning. *Proc. Natl. Acad. Sci. U. S. A.* 116 (44): 22071–22080.
- 88 Jiménez-Luna, J., Grisoni, F., and Schneider, G. (2020). Drug discovery with explainable artificial intelligence. *Nat. Mach. Intell.* 2 (10): 573–584.

Computational Drug Discovery

Computational Drug Discovery

Methods and Applications

*Edited by Vasanthanathan Poongavanam and
Vijayan Ramaswamy*

Volume 2

Editors

Dr. Vasanthanathan Poongavanam

Uppsala University
Department of Chemistry-BMC
751 05 Uppsala
Sweden

Dr. Vijayan Ramaswamy

University of Texas MD Anderson
Cancer Center
Institute for Applied Cancer Science
TX
United States

Cover: © Vasanthanathan Poongavanam

■ All books published by **WILEY-VCH** are carefully produced. Nevertheless, authors, editors, and publisher do not warrant the information contained in these books, including this book, to be free of errors. Readers are advised to keep in mind that statements, data, illustrations, procedural details or other items may inadvertently be inaccurate.

Library of Congress Card No.: applied for

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library.

Bibliographic information published by the Deutsche Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at <<http://dnb.d-nb.de>>.

© 2024 WILEY-VCH GmbH, Boschstraße 12, 69469 Weinheim, Germany

All rights reserved (including those of translation into other languages). No part of this book may be reproduced in any form – by photoprinting, microfilm, or any other means – nor transmitted or translated into a machine language without written permission from the publishers. Registered names, trademarks, etc. used in this book, even when not specifically marked as such, are not to be considered unprotected by law.

Print ISBN: 978-3-527-35375-0

ePDF ISBN: 978-3-527-84072-4

ePub ISBN: 978-3-527-84073-1

oBook ISBN: 978-3-527-84074-8

Typesetting Straive, Chennai, India

Contents

Volume 1

Preface xv

Acknowledgments xix

About the Editors xxi

Part I Molecular Dynamics and Related Methods in Drug Discovery 1

- 1 Binding Free Energy Calculations in Drug Discovery 3**
Anita de Ruiter and Chris Oostenbrink
- 2 Gaussian Accelerated Molecular Dynamics in Drug
Discovery 21**
Hung N. Do, Jinan Wang, Keya Joshi, Kushal Koirala, and Yinglong Miao
- 3 MD Simulations for Drug-Target (Un)binding Kinetics 45**
Steffen Wolf
- 4 Solvation Thermodynamics and its Applications in
Drug Discovery 65**
Kuzhanthaivelan Saravanan and Ramesh K. Sistla
- 5 Site-Identification by Ligand Competitive Saturation as
a Paradigm of Co-solvent MD Methods 83**
Asuka A. Orr and Alexander D. MacKerell Jr

Part II Quantum Mechanics Application for Drug Discovery 119

- 6 QM/MM for Structure-Based Drug Design: Techniques
and Applications 121**
Marc W. van der Kamp and Jaida Begum

7	Recent Advances in Practical Quantum Mechanics and Mixed-QM/MM-Driven X-Ray Crystallography and Cryogenic Electron Microscopy (Cryo-EM) and Their Impact on Structure-Based Drug Discovery 157 <i>Oleg Borbulevych and Lance M. Westerhoff</i>
8	Quantum-Chemical Analyses of Interactions for Biochemical Applications 183 <i>Dmitri G. Fedorov</i>
	Part III Artificial Intelligence in Pre-clinical Drug Discovery 211
9	The Role of Computer-Aided Drug Design in Drug Discovery 213 <i>Storm van der Voort, Andreas Bender, and Bart A. Westerman</i>
10	AI-Based Protein Structure Predictions and Their Implications in Drug Discovery 227 <i>Tahsin F. Kellici, Dimitar Hristozov, and Inaki Morao</i>
11	Deep Learning for the Structure-Based Binding Free Energy Prediction of Small Molecule Ligands 255 <i>Venkatesh Mysore, Nilkanth Patel, and Adegoke Ojewole</i>
12	Using Artificial Intelligence for <i>de novo</i> Drug Design and Retrosynthesis 275 <i>Rohit Arora, Nicolas Brosse, Clarisse Descamps, Nicolas Devaux, Nicolas Do Huu, Philippe Gendreau, Yann Gaston-Mathé, Maud Parrot, Quentin Perron, and Hamza Tajmouati</i>
13	Reliability and Applicability Assessment for Machine Learning Models 299 <i>Fabio Urbina and Sean Ekins</i>
	Volume 2
	Preface xv
	Acknowledgments xix
	About the Editors xxi
	Part IV Chemical Space and Knowledge-Based Drug Discovery 315
14	Enumerable Libraries and Accessible Chemical Space in Drug Discovery 317 <i>Tim Knehan, Nicholas A. Boyles, and Pieter H. Bos</i>
14.1	Chemical Space and Its Generation 317

14.1.1	Simple SMILES Enumeration and String Operations	317
14.1.2	R-Group and Scaffold Enumeration	319
14.1.3	Bioisosteres and Matched Molecular Pairs	320
14.1.4	Enumeration Within De Novo Design	321
14.2	Public and Commercial Chemical Libraries	323
14.2.1	Vendor Libraries	323
14.2.2	Reaction-Based Chemical Spaces	326
14.2.2.1	Reaction Sources	327
14.2.2.2	Available Building Blocks	328
14.3	How to Effectively Explore Chemical Space?	329
	References	331
15	Navigating Chemical Space	337
	<i>Ákos Tarcsay, András Volford, Jonathan Buttrick, Jan-Constantin Christopherson, Máte Erdős, and Zoltán B. Szabó</i>	
15.1	Introduction to Chemical Space	337
15.2	Workflows on the Chemical Space and Related Challenges	340
15.2.1	Project Data Exploration	341
15.2.2	Corporate Data Warehouses	341
15.2.3	Novel Compound Design	342
15.2.4	Registration Systems	343
15.2.5	SAR-by Catalog	343
15.3	Challenges in Chemical Data Search	344
15.4	Technologies	345
15.4.1	Representation Used for Similarity Search	345
15.4.1.1	Substructure-Preserving Fingerprints	345
15.4.1.2	Feature Fingerprints	346
15.4.2	Additional Representations	349
15.4.3	Commonly Used Similarity Definitions	350
15.5	Similarity Search Applications	352
15.6	Substructure Search	353
15.6.1	Atom by Atom Search	353
15.6.2	Ullmann Algorithm	353
15.6.3	VF2 Algorithm	354
15.7	Application Example	355
15.8	Summary and Outlook	357
	Acknowledgments	358
	References	358
16	Visualization, Exploration, and Screening of Chemical Space in Drug Discovery	365
	<i>José J. Naveja, Fernanda I. Saldivar-González, Diana L. Prado-Romero, Angel J. Ruiz-Moreno, Marco Velasco-Velázquez, Ramón Alain Miranda-Quintana, and José L. Medina-Franco</i>	
16.1	Introduction	365
16.2	Exploiting Bioactivity Data in the Artificial Intelligence Era	367

16.2.1	Databases Annotated with Biological Activity	368
16.2.2	Opportunities for AI in Ligand-Based Drug Design	368
16.2.3	Opportunities in Structure-Based Drug Design	369
16.3	Chemical Space and Chemical Multiverse	369
16.3.1	Recent Progress on Chemical Space	370
16.3.2	Chemical Multiverse and Constellation Plots	370
16.4	Hit Identification, Optimization, and Development of Bioactive Compounds	373
16.4.1	Virtual Screening	373
16.4.2	VISAS: General Approach that Expands Bioactive Molecules	373
16.4.2.1	ViSAS on an Antituberculosis Chemical Dataset	376
16.4.3	De Novo Design Libraries	377
16.4.3.1	Case Study: DNMT-Focused Libraries	379
16.5	Extended Similarity Methods	380
16.5.1	The Extended Similarity Framework	381
16.5.2	Global Description: Chemical Diversity, Chemical Library Networks, and Clustering	382
16.5.3	Local Description: Diversity Selection and Medoid Calculation	383
16.6	Conclusion and Outlook	383
	Acknowledgments	384
	Abbreviations	385
	References	385
17	SAR Knowledge Bases for Driving Drug Discovery	395
	<i>Nishanth Kandepedu, Anil Kumar Manchala, and Norman Azoulay</i>	
17.1	Introduction	395
17.2	The Origins of SAR Knowledge Bases	396
17.3	SAR Knowledge Base Landscape	398
17.3.1	Open-Source SAR Databases	398
17.3.1.1	PubChem	398
17.3.1.2	ChEMBL	399
17.3.1.3	DrugBank	399
17.3.1.4	BindingDB	400
17.3.1.5	Other Known Open-Source Databases	400
17.3.2	Commercial SAR Databases	401
17.3.2.1	GOSTAR	401
17.3.2.2	Reaxys Medicinal Chemistry	402
17.3.2.3	Other Known Commercial Databases	403
17.4	Comparison and Complementarity of SAR Databases	403
17.5	Applications of SAR Knowledge Base in Modern Drug Discovery	407
17.6	Future Direction	410
	Acknowledgment	410
	List of Abbreviations	411
	Disclaimer	412
	References	412

18	Cambridge Structural Database (CSD) – Drug Discovery Through Data Mining & Knowledge-Based Tools	419
	<i>Francesca Stanzione, Rupesh Chikhale, and Laura Friggeri</i>	
18.1	Introduction	419
18.2	The Cambridge Structural Database (CSD) and CSD-Based Tools	420
18.3	How CSD and CSD Knowledge-Based Tools Can Aid Drug Discovery	422
18.3.1	Target Identification and Target Validation	422
18.3.2	Hit Identification	425
18.4	Hit-to-Lead	428
18.4.1	Lead Optimization	430
18.5	Challenges in Drug Development	432
18.5.1	How CSD Can Further Impact Drug Discovery	433
	References	434
	 Part V Structure-Based Virtual Screening Using Docking	 441
19	Structure-Based Ultra-Large Virtual Screenings	443
	<i>Christoph Gorgulla</i>	
19.1	Introduction	443
19.2	Fundamentals	444
19.2.1	Receptor Structures and Preparation	444
19.2.2	Ligand Preparation and Ligand Libraries	444
19.2.3	Molecular Docking	446
19.2.4	Virtual Screenings	446
19.3	Ultra-Large Ligand Libraries	447
19.3.1	Commercial Libraries	447
19.3.1.1	REAL Database and REAL Space	447
19.3.1.2	CHEMriya	448
19.3.1.3	GalaXi	448
19.3.1.4	eXplore	449
19.3.1.5	Freedom Space	449
19.3.1.6	ZINC Libraries	449
19.3.1.7	VirtualFlow Libraries	449
19.3.2	Public Virtual Libraries	450
19.3.2.1	Generated Databases (GDBs)	450
19.3.2.2	KnowledgeSpace	450
19.4	Docking-Based Ultra-Large Virtual Screenings	451
19.4.1	Success Stories	451
19.4.1.1	Gorgulla (2018)	451
19.4.1.2	Lyu et al. (2019)	452
19.4.1.3	Stein et al. (2020)	453
19.4.1.4	Gorgulla et al. (2020)	453

19.4.1.5	Alon et al. (2021)	453
19.4.1.6	Other Success Stories	453
19.4.2	Available Software for ULVSS	454
19.4.2.1	UCSF DOCK	454
19.4.2.2	VirtualFlow	455
19.5	Synthon-Based Virtual Screenings	455
19.6	Machine Learning-Based Virtual Screenings	457
19.6.1	Deep Docking	458
19.6.2	MolPAL	459
19.7	Other Acceleration Techniques	460
19.7.1	Deep Learning Approaches to Molecular Docking	461
19.7.2	GPU Acceleration of Molecular Dockings	461
19.8	Quality of Ultra-Large Virtual Screening Results	462
19.9	Conclusion and Outlook	462
	References	463

20 Community Benchmarking Exercises for Docking and Scoring 471

Bharti Devi, Anurag TK Baidya, and Rajnish Kumar

20.1	Introduction	471
20.1.1	Overview of Molecular Docking	472
20.2	Need for Benchmarking	474
20.2.1	Benchmarking Datasets	475
20.2.1.1	Benchmarking Sets for Pose Prediction	475
20.2.1.2	Benchmarking Sets for Virtual Screening	477
20.2.2	Evaluation Metrics	478
20.2.2.1	Enrichment Factor	478
20.2.2.2	Docking Enrichment (DE)	479
20.2.2.3	BEDROC Score	479
20.2.2.4	Receiver Operating Characteristic Curve	480
20.2.2.5	Root Mean Square Deviation (RMSD)	480
20.2.2.6	Real-Space R Values	480
20.2.2.7	Root Mean Squared Error (RMSE)	481
20.2.2.8	Spearman's Rank-Order Correlation	481
20.2.2.9	Kendall Rank Correlation	481
20.3	Community Benchmarking Exercises	482
20.3.1	Statistical Assessment of Proteins and Ligands (SAMPL) Challenge	482
20.3.2	Drug Design Data Resource (D3R) Challenges	483
20.3.3	Critical Assessment of Computational Hit-Finding Experiments	483
20.3.4	Continuous Evaluation of Ligand Protein Predictions (CELPP)	485
20.4	Lessons Learned from the Benchmarking Exercises	486
20.4.1	Quality of Crystal Structures	486
20.4.2	Need of Sufficient Metrics to Assess Docking and Scoring	487
20.4.3	Usefulness of Statistics, Error Bars, and Confidence Intervals (CI)	487
20.4.4	Requirement of Good Data Set	488

20.5	Summary	489
	References	489

Part VI In Silico ADMET Modeling 495

21	Advances in the Application of In Silico ADMET Models – An Industry Perspective	497
	<i>Wenyi Wang, Fjodor Melnikov, Joe Napoli, and Prashant Desai</i>	
21.1	Introduction	497
21.2	QSAR Models	498
21.2.1	Conventional QSAR Model Development	499
21.2.1.1	Data Curation	499
21.2.1.2	Chemical Descriptors	503
21.2.1.3	Algorithms	504
21.2.1.4	Model Validation and Evaluation Metrics	506
21.2.1.5	Utilizing Relevant Properties to Improve QSAR Model Performance	507
21.2.1.6	Applicability Domain and Reliability of Individual Prediction	507
21.2.1.7	Model Interpretability vs. Predictivity	509
21.2.1.8	Model Deployment and Accessibility	509
21.2.2	In Silico ADMET Models and Their Influence During Drug Discovery	510
21.2.2.1	In Silico ADME Models	510
21.2.2.2	In Silico Toxicity Models	512
21.2.2.3	Typical QSAR Models for ADMET in the Industry	516
21.2.3	Emerging QSAR Technologies and Algorithms	516
21.3	Extended Scope of In Silico ADMET	519
21.3.1	Exploring New Chemical Space With Generative Models Considering ADMET	519
21.3.2	Mechanism-Based Models	520
21.3.3	Predictive MetID (Metabolites Identification From Chemical Structures)	521
21.4	Conclusion	522
	References	523

Part VII Computational Approaches for New Therapeutic Modalities 537

22	Modeling the Structures of Ternary Complexes Mediated by Molecular Glues	539
	<i>Michael L. Drummond</i>	
22.1	Introduction	539
22.2	Methodology	541

22.3	Results and Discussion	543
22.3.1	Approach 1: Treating MGs as Whole, Indivisible Molecules	543
22.3.2	Approach 2: Treating MGs as “linkerless PROTACs”	548
22.4	Conclusions	556
	References	557
23	Free Energy Calculations in Covalent Drug Design	561
	<i>Levente M. Mihalovits, György G. Ferenczy, and György M. Keserű</i>	
23.1	Introduction	561
23.2	Mechanism of Covalent Inhibition	561
23.3	Computational Characterization of Reversible Covalent Binding	562
23.4	Computational Characterization of Irreversible Covalent Binding	564
23.4.1	Computation of the Dissociation Constant of the Noncovalent Step	566
23.4.2	Computation of the Rate Constant of the Covalent Step	567
23.5	Case Studies of Reversible Covalent Inhibition	569
23.6	Case Studies of Irreversible Covalent Inhibition	571
23.7	Summary	574
	References	575

Part VIII Computing Technologies Driving Drug Discovery 579

24	Orion® A Cloud-Native Molecular Design Platform	581
	<i>Jesper Sørensen, Caitlin C. Bannan, Gaetano Calabrò, Varsha Jain, Grigory Ovanesyan, Addison Smith, She Zhang, Christopher I. Bayly, Tom A. Darden, Matthew T. Geballe, David N. LeBard, Mark McGann, Joseph B. Moon, Hari S. Muddana, Andrew Shewmaker, Jharrod LaFon, Robert W. Tolbert, A. Geoffrey Skillman, and Anthony Nicholls</i>	
24.1	Introduction	581
24.2	The Platform	582
24.3	Target Preparation and Structural Data Organization	585
24.4	Virtual Screening	591
24.5	Predicting Small-Molecule Binding Affinity	595
24.6	ADMET Prediction and Permeability in Drug Discovery	600
24.7	Predicting Drug Crystal Forms	604
24.8	Summary	608
	References	609
25	Cloud-Native Rendering Platform and GPUs Aid Drug Discovery	617
	<i>Mark Ross, Michael Drummond, Lance Westerhoff, Xavier Barbeau, Essam Metwally, Sasha Banks-Louie, Kevin Jorissen, Anup Ojah, and Ruzhu Chen</i>	
25.1	Introduction	617

25.2	Complex Molecular Mechanics at Scale	619
25.3	Modeling Billions of Molecules in a Day	621
25.4	Faster Free Energy	622
25.5	Vision for the Future	624
25.6	Concluding Remarks	625
	Disclaimer	625
	References	625
26	The Quantum Computing Paradigm	627
	<i>Thomas Ehmer, Gopal Karemore, and Hans Melo</i>	
26.1	What to Expect	627
26.1.1	Motivation	627
26.1.2	Structure of the Chapter	628
26.2	Another New Paradigm	628
26.2.1	Digital	629
26.2.2	Quantum	630
26.2.2.1	Refresher – Quantum Mechanics and Its Features	630
26.2.2.2	$ WTFUQC\rangle$ – The $\hbar(y\rangle + o\rangle)$ pe	631
26.2.3	Challenges	631
26.2.3.1	Cognitive Challenge – Quantum Literacy	631
26.2.3.2	Cultural Challenge	635
26.3	Quantum Computing Overview	635
26.3.1	Quantum Simulators	635
26.3.2	Embedding Quantum into Computing	636
26.3.3	What’s the New Idea?	636
26.3.4	Introducing the Concept of the Qubit	636
26.3.4.1	Superposition	638
26.3.4.2	The Bloch Sphere	638
26.3.4.3	Interference	638
26.3.4.4	Nondeterminism	639
26.3.4.5	Entanglement	640
26.3.4.6	Multi-particle Registers	640
26.3.5	Quantum Computing Stack	641
26.3.6	Major Applications of Quantum Computing	642
26.3.6.1	Classes of Applications	643
26.3.6.2	Famous Quantum Algorithms (Gate)	643
26.3.7	Gate Quantum Computing	644
26.3.7.1	Example: The Hadamard Gate	645
26.3.7.2	Example: The CNOT Gate	646
26.3.7.3	Development Kits	647
26.3.8	Adiabatic/Annealing	647
26.3.8.1	Intuitive Explanation	647
26.3.8.2	A Brief Explanation	648
26.3.8.3	In Summary	648
26.3.8.4	Digital Annealer	648

26.3.9	Hardware	648
26.3.10	Technological Challenges	650
26.3.10.1	Errors	650
26.3.10.2	Scalability	651
26.3.10.3	Connctivity	651
26.3.11	Quantum Computer for Molecular Biology	651
26.3.11.1	Local Quantum Effects in Biochemical Processes	653
26.3.11.2	Global Quantum Effects in Functional Bio-Molecules	654
26.3.11.3	Quantum Computing for Structural Biology	654
26.3.11.4	Quantum Computing for Data-Driven Approaches to Molecular Biology	655
26.4	Quantum Machine Learning	655
26.4.1	Introduction	655
26.4.2	The Shortcoming of Classical ML Models	656
26.4.3	Types of QML	657
26.4.3.1	CQ Algorithm Types	658
26.4.3.2	Quantum Regression Model or Quantum Algorithm for the Method of Least Squares	658
26.4.3.3	Quantum Clustering Model	658
26.4.3.4	Quantum Kernel Model or Quantum Support Vector Machines	659
26.4.3.5	Quantum Neural Networks	659
26.4.3.6	Quantum Generative Models	660
26.4.4	Limitations of QML	661
26.5	Designing Peptides and Proteins	662
26.5.1	Background	662
26.5.2	<i>De Novo</i> Rational Design	662
26.5.3	Peptide and Protein Design	663
26.5.4	Protein Design as an Optimization Problem	663
26.5.5	Quantum Optimization for Peptide and Protein Design	664
26.5.6	Quantum Annealing Approach	664
26.5.7	The qPacker Algorithm	665
26.5.8	Gate-Based Approaches	667
26.6	Conclusion	667
26.7	Further Reading	668
	References	669
	Index	679

Preface

Computer-aided drug design (CADD) techniques are used in almost every stage of the drug discovery continuum, given the need to shorten discovery timelines, reduce costs, and improve the odds of clinical success. CADD integrates modeling, simulation, informatics, and artificial intelligence (AI) to design molecules with desired properties. Briefly, the application of CADD methodologies in drug discovery dates back to the 1960s, tracing its origin to the development of quantitative structure–activity relationship (QSAR) approaches. Between the 1970s and 1980s, computer graphics programs to visualize macromolecules began to take off together with advancements in computational power. This coincided with the emergence of more sophisticated techniques, including mapping energetically favorable binding sites on proteins, molecular docking, pharmacophore modeling, and modeling the dynamics of biomolecules. Since then, CADD has evolved as a powerful technique opening new possibilities, leading to increased adoption within the pharmaceutical industry and contributing to the discovery of several approved drugs.

Recent developments in CADD have been propelled by advancements in computing, breakthroughs in related fields such as structural biology, and the emergence of new therapeutic modalities. Notably, the advent of highly parallelizable GPUs and cloud computing have significantly increased computing power, while quantum computing holds promise to simulate complex systems at an unprecedented scale and speed. Advances in AI technologies, particularly generative AI for molecule design, are reducing cycle times during lead optimization. Meanwhile, the resolution revolution in cryo-electron microscopy (cryo-EM) and AI-powered structure biology are shedding light on the three-dimensional structure of many therapeutically relevant drug targets, thereby expanding our ability to carry out structure-based drug design against these targets. Other exciting breakthroughs that offer new opportunities include the explosion in the size of "make-on-demand" chemical libraries that enable ultra-large-scale virtual screening for hit identification, the big data phenomena in medicinal chemistry with the advent of bioactivity databases like ChEMBL and GOSTAR that provide access to millions of SAR data points useful for building predictive models and for knowledge-based compound, the emergence of new therapeutic modalities like targeted protein degradation like PROTACs and molecular glues, and viable approaches for targeting various reactive amino acid side chains beyond cysteine for developing covalent inhibitors. These

developments are also now enabling drug discovery scientists to tackle high-value drug targets previously considered undruggable.

The changing paradigm in drug discovery, complemented by technological advancements, has significantly expanded the toolbox available for computational chemists to enable drug discovery in recent years. Against this backdrop, we felt a need for a book that offers up-to-date information on the most important developments in the field of CADD. This book, titled “Computational Drug Discovery,” is meant to be a valuable resource for readers seeking a comprehensive account of the latest developments in CADD methods and technologies that are transforming small-molecule drug discovery. The intended target audience for this book is medicinal chemists, computational chemists, and drug discovery professionals from industry and academia.

The book is organized into eight thematic sections, each dedicated to a cutting-edge computational method, or a technology utilized in computational drug discovery. In total, it comprises 26 chapters authored by renowned experts from academia, pharma, and major drug discovery software providers, offering a broad overview of the latest advances in computational drug discovery.

Part I explores the role of molecular dynamics simulation and related approaches in drug discovery. It encompasses various topics such as the utilization of physics-based methods for binding free energy estimation, the theory and application of enhanced sampling methods like Gaussian Accelerated MD to facilitate efficient sampling of the conformational space, understanding binding and unbinding kinetics of compound binding through molecular dynamics simulation, the application of computational approaches like WaterMap and 3D-RISM framework to understand the location and thermodynamic properties of solvents that solvate the binding pocket which offers rich physical insights compound design, and the use of mixed solvent MD simulations for mapping binding hotspots on protein surfaces based on the SILCS technology.

Part II focuses on the role of quantum mechanical approaches in drug discovery, covering topics such as the use of hybrid QM/MM method for modeling reaction mechanisms and covalent inhibitor design, refinement of X-ray and cryo-EM structures integrating QM and QM/MM approaches for accurate assignment of tautomer, protomers, and amide flip rotamers for downstream structure-based design, and quantifying protein–ligand interaction energies using QM methods at a reduced computational cost like the fragment molecular orbital (FMO) framework

Part III focuses on the application of AI in preclinical drug discovery, highlighting its growing importance across different stages of the drug discovery process. Given the recent advancements in AI and related technologies, we have chapters that outline advancements in deep learning for protein structure prediction, in particular the significant breakthrough achieved by AlphaFold2, the use of deep learning architectures such as Convolutional Neural Networks (CNNs), Graph Neural Networks (GNNs), and physics-inspired neural networks for predicting protein–ligand binding affinity, the emergence of generative modeling techniques for *de novo* design of synthetically tractable drug-like molecules that satisfy a defined set of constraints. In order to offer readers guidance on effectively applying

machine learning (ML) models and ensuring their validity and usefulness, this section includes a chapter that discusses different approaches for evaluating the reliability and domain applicability of ML models.

Part IV of this book focuses on how the concept of chemical space and the big data phenomenon are driving drug discovery. It includes chapters describing innovative approaches in reaction-based enumerations that enable the generation of virtual libraries containing tangible compounds, followed by computational solutions for visualizing and navigating this vast chemical space. Additionally, this section also highlights the use of SAR *knowledge bases like GOSATR* for extracting valuable insights and generating robust design ideas based on medicinal chemistry precedence. Wrapping up the section is a chapter highlighting how the wealth of knowledge gained by mining the data in CSD is proving valuable in various stages of drug discovery.

The ever-expanding size of compound libraries and the advent of make-on-demand compound libraries have elevated virtual screening to a whole new level.

Part V focuses on ultra-large-scale virtual screening using approaches that scale virtual screening methods to match the size of these massively large compound libraries. Although virtual screening using docking is a well-established approach for hit finding in drug discovery, the ability of docking programs to generate the correct binding mode and accurately estimate binding affinity is still a challenge. Hence, we have a chapter that reviews collaborative efforts within the scientific community for evaluating and comparing the performance of docking methods, establishing standardized metrics for assessing the efficiency of virtual screening techniques through rigorous competitive evaluations.

Early profiling of absorption, distribution, metabolism, excretion, and toxicity (ADMET) endpoints in early drug discovery is essential for designing and selecting compounds with superior ADMET properties. Consequently, major pharmaceutical companies have developed and implemented predictive models within their organizations for predicting multiple endpoints to enhance compound design. **Part VI** of the book chapter offers an overview of in silico ADMET methods and their practical applications in facilitating compound design within an industrial context. **Part VII** explores the role of computational techniques in accelerating the design of cutting-edge therapeutic modalities. This section provides a comprehensive focus on two key areas: the design of molecular glues and the design of covalent inhibitors.

In addition to the aforementioned methods and approaches that revolutionize the drug discovery process, computing technologies are further accelerating drug discovery with enhanced speed and accuracy.

Part VIII is dedicated to exploring how cloud computing and quantum computing significantly expand the range of drug discovery opportunities. Particularly, there is great hope and excitement surrounding the potential applications of quantum computing in drug discovery. *“The Quantum Computing Paradigm”* provides a comprehensive review on quantum computing from the perspective of drug discovery. In addition to discussing several drug discovery applications, including peptide design, the chapter also addresses challenges associated with this emerging drug discovery technology.

In conclusion, we believe that this book provides a thorough overview of the recent advancements in computational drug discovery, making it an engaging and captivating read. We would like to express our deepest appreciation to all the authors for their invaluable contributions to this book. Their expertise, insights, and unwavering commitment have greatly enriched its content and overall significance.

14 September 2023

Vasanthanathan Poongavanam, Uppsala, Sweden
Vijayan Ramaswamy, Texas, USA

Acknowledgments

First and foremost, we would like to extend our sincerest gratitude and profound appreciation to all the contributing authors. Their unwavering commitment, tireless efforts, and remarkable enthusiasm have been instrumental in bringing this book to fruition. It is their willingness to share their knowledge and experience that has greatly enriched its content, resulting in a truly valuable and comprehensive book that provides an account of the latest advancements in the field of computer-aided drug design.

We also extend our sincere gratitude to the external reviewers for their timely feedback and insightful suggestions that helped improve the quality of the book and shape the final outcome. Our special thanks to the following individuals for their invaluable contributions in reviewing the book chapters: Dr. Andreas Tosstorff (F. Hoffmann-La Roche, Switzerland), Dr. Sagar Gore, Dr. Suneel Kumar BVS (Molecular Forecaster, Canada), Dr. Pandian Sokkar, Dr. Ono Satoshi (Mitsubishi Tanabe Pharma, Japan), Dr. Octav Caldararu (Zealand Pharma, Denmark), Dr. Sundarapandian Thangapandian (HotSpot Therapeutics, Inc, USA), Dr. Vigneshwaran Namasivayam (Dewpoint Therapeutics, Germany), Dr. Yinglong Miao (University of Kansas, USA), Nanjie Deng (Pace University, USA), and Dr. Ansuman Biswas (Ernst & Young, India).

In conclusion, we would like to express our gratitude to the publisher Wiley for entrusting us with an opportunity to edit this book and for the fruitful collaboration. Especially, we convey our appreciation to Katherine Wong (Senior Managing Editor) and Dr. Lifan Yang (Program Manager) at Wiley for their unwavering support, encouragement throughout the editing process, and their commitment to ensuring the quality and excellence of this book. The editors also extend their thanks to Prof. Jan Kihlberg and Dr. Jason B. Cross for their continuous support, which made this project possible.

About the Editors

Vasanthanathan Poongavanam is a senior scientist in the Department of Chemistry-BMC, Uppsala University, Sweden. Before starting at Uppsala University in 2016, he was a postdoctoral fellow at the University of Vienna, Austria, and at the University of Southern Denmark. He obtained his PhD degree in Computational Medicinal Chemistry as a Drug Research Academy (DRA) Fellow at the University of Copenhagen, Denmark, on computational modeling of cytochrome P450. He has published more than 65 scientific articles, including reviews and book chapters. His scientific interests focus on in silico ADMET modeling, including cell permeability and solubility, and he has worked extensively on understanding the molecular properties that govern the pharmacokinetic profile of molecules in bRo5 property space, including macrocycles and PROTACs.

Vijayan Ramaswamy (R.S.K. Vijayan) is a senior research scientist affiliated with the Structural Chemistry division at the Institute for Applied Cancer Science, The University of Texas MD Anderson Cancer Center, TX, USA. In 2016, he joined MD Anderson Cancer after a brief tenure as a scientist in computational chemistry at PMC Advanced Technologies, New Jersey, USA. He undertook postdoctoral training at Rutgers University in New Jersey, USA, and Temple University in Pennsylvania, USA. He received his PhD as a CSIR senior research fellow from the Indian Institute of Chemical Biology, Kolkata, India. He is a named co-inventor on seven issued US patents, including an ATR kinase inhibitor that has advanced to clinical trials. He has published more than 20 scientific articles and authored one book chapter. His research focuses on applying computational chemistry methods to drive small-molecule drug discovery programs, particularly in oncology and neurodegenerative diseases.

Part IV

Chemical Space and Knowledge-Based Drug Discovery

14

Enumerable Libraries and Accessible Chemical Space in Drug Discovery

Tim Knehans, Nicholas A. Boyles, and Pieter H. Bos

Schrödinger Inc., 1540 Broadway, New York, NY 10036, USA

14.1 Chemical Space and Its Generation

In Cheminformatics, chemical space refers to all molecular instances within the confines of a series of properties, e.g. molecular weight (MW), number of carbons, ring type, or even similarity to a reference/tool compound. The most popular chemical space for pharmaceutically relevant molecules is confined within the standard upper boundaries of Lipinski's Rule of Five (Molecular Weight ≤ 500 Da, $\log P \leq 5$, Hydrogen Bond Acceptors ≤ 10 , Hydrogen Bond Donors ≤ 5) [1]. The size of chemical space within the constraints of Lipinski's Rule of Five is estimated to be at least 10^{60} molecules, which puts it in the neighborhood of the estimated number of atoms within the Milky Way 10^{68} [2–5]. However, several drugs already defy these rules, and it is speculated that the chemical space relevant to drug discovery is closer to the 10^{100} molecules estimated in early combinatorial exercises, where simple R-group analyses of single compounds already yielded libraries the size of 10^9 molecules [6].

Driven by the need to escape existing chemical matter to address novel and existing targets, combined with the commercial requirements to break out of an ever-increasing number of patents claiming more and more of the relevant chemical space, efficient access and exploration of previously unutilized regions of the potential chemical space, are key requirements in order to maintain success in modern drug discovery efforts. Currently, the largest documented chemical space generated for in silico drug discovery amounts to 10^{26} , highlighting that there is much chemical space within the Lipinski Rules that remains to be explored and utilized [7].

14.1.1 Simple SMILES Enumeration and String Operations

In order to explore chemical space in a suitable way for modern structure- and ligand-based modeling methods, molecules have to be generated in a suitable format so that their physicochemical properties, and respective target binding affinity, can be calculated. Considering the aforementioned amount of potentially relevant molecules within the desired chemical space, a representation that is efficient,

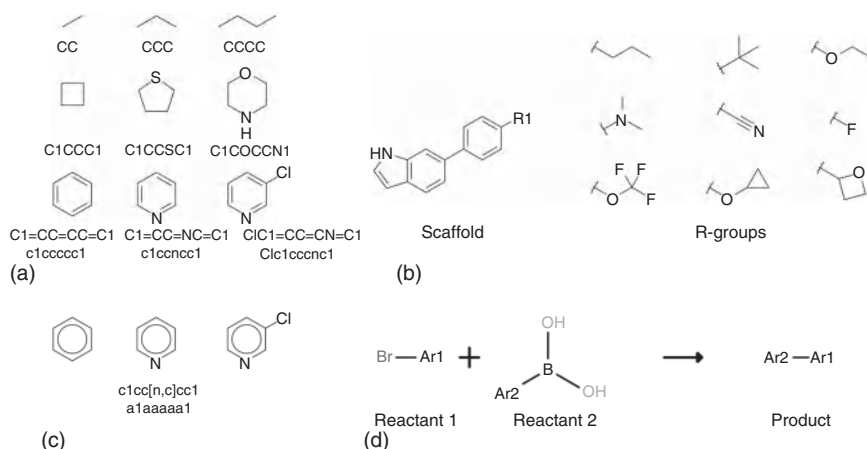


Figure 14.1 (a) 2D molecular and SMILES examples of simple molecules. The molecules serve to highlight how simple expansions of the SMILES string can be utilized to generate molecules in a brute-force combinatorial manner. While the first row shows a simple extension of an aliphatic chain by adding a single aliphatic carbon at a time, the second row shows increasing ring sizes with heteroatoms simply denominated by their elemental symbol. The integer indicates the atom where the ring starts and subsequently closes. The third-row shows (hetero) aromatic systems. These compounds can be represented in the Kekulé structure, captured by alternating single and double bonds or by using a lowercase representation indicating an aromatic atom type. (b) R-group enumeration. The left side shows an example scaffold molecule with a single R1 attachment point. The right side shows an example library of R-groups that can be attached. The squiggly line denotes the connection to the R1-connected atom, thus forming a single bond in this example. (c) Example of SMARTS patterns that match all three structures shown via either the application of square brackets and a comma (logical OR operator), or the use of “a” representing an aromatic atom. (d) Example reactions used in reaction-based enumerations. Molecules containing the shown substructure for reactant 1 and reactant 2 would be classified as such. Upon reacting both molecules, all atoms excluding Ar1 and Ar2 would be deleted and a single bond is formed between the two atoms.

primarily in terms of storage space, is highly desired. In the 1980s, David Weininger developed the **Simplified Molecular-Input Line-Entry System (SMILES)** standard, which is the representation of any molecule in the form of a one-dimensional string of characters (Figure 14.1a) [8, 9]. To this day, SMILES represents one of the most popular molecular formats, and most chemical databases are distributed as such. Other than the reduction of storage space, the SMILES standard allowed for the quick generation of chemical space, simply by adding a set of characters to an existing SMILES string in order to generate novel molecules [10, 11].

The SMILES standard was later expanded to contain the ability to incorporate primitives for atoms and bonds as well as to utilize a derived property like aromaticity, connectivity, or ring affiliation [12]. Additionally, advanced functionality like logical operators was also included, further improving the ability to efficiently store molecules as well as enabling functionalities like substructure searches in large libraries. Most modern modeling packages, commercial or free of charge, provide an interface to incorporate SMILES representations of molecules and the ability to use SMARTS patterns, both for substructure search as well as for the purpose of

transforming one molecule into another. Thus, in theory, the chemical space referenced above is accessible to everyone by simply generating all possible SMILES permutations and verifying molecular and valence integrity via an open-source modeling package [13]. In practice, the number of combinatorial solutions is restricted by reducing the number of fragments/building blocks to a set that can realistically be synthesized, restricting the manner in which fragments are combined by confining it toward synthetically accessible rules/reactions, and finally, by applying fitness functions (e.g. spatial restrictions, MPO scores).

The generated database (GDB) by Reymond et al. currently (GDB 17) represents the largest combinatorial library publicly available and consists of 166 billion explicitly enumerated molecules (Figure 14.2) [17]. To generate combinatorial solutions, molecules are abstracted initially to the graph level allowing nodes to have up to four connections to represent quaternary carbons and a total number of nodes equivalent to the version of the GDB. The graphs are then instantiated into saturated hydrocarbons containing only single bonds and carbons. The hydrocarbons are then “unsaturated” by substituting single bonds with double and triple bonds. Following this, the carbons are exchanged for nitrogens and oxygens. At each of these steps, specific rule sets were applied, such as excluding knotted graph topologies, unsaturation filters, and a variety of functional group filters in order to ensure that the remaining molecules have high chemical veracity. Finally, in a postprocessing step the incorporation and handling of aromatic heterocycles, oximes, nitro groups as well as halogens and sulfur is governed [18–21].

It is noteworthy that the GDB represents the largest fully enumerated space, containing 166 billion combinations, and is generated from molecules containing only 17 heavy atoms. This large quantity when dealing with combinatorial chemistry is referred to as combinatorial explosion, a fact well recognized, representing one of the biggest challenges to dealing with chemical space [22]. To highlight the practical implications, even generating 10^{15} SMILES is rather prohibitive in computational cost as well as hardware storage. Taking the GDB subset containing 50 million compounds at a size of 314 MB, and storage costs of 1 cent per GB/month would result in an annual storage cost of ~\$750 000 for the 10^{15} space. In a combined database encapsulating several vendor libraries that contain larger molecules, the average byte size of a SMILES string is 53 bytes, resulting in annual costs of around \$6 million for a set of 10^{15} molecules, excluding any additional stored information like molecular properties. Thus, methodologies that enumerate smaller, but more focused chemical subspaces, are generally preferred.

14.1.2 R-Group and Scaffold Enumeration

A standard way to quickly enumerate a reference molecule, or scaffold, and explore the immediate chemical environment around the scaffold is through the use of R-group enumeration. In R-group enumerations, a fixed attachment point is chosen on a given molecule, and a pregenerated library of R-groups is attached computationally to generate the enumerated set of output molecules (Figure 14.1b). Depending on the size of the R-group libraries, the resulting chemical space can vary from small to extremely large [6]. Although generally, the size of the attached

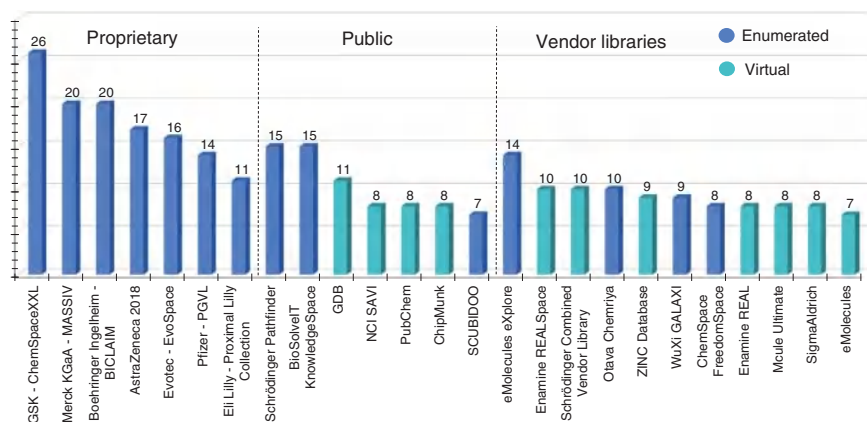


Figure 14.2 Overview of chemical spaces and their size in log units. Reference values are Adapted from Refs. [14–16] and updated from their latest reported sizes. Virtual spaces (blue) are spaces that are not-enumerated but are countable by the reactions that make up the space and all building blocks suitable for the respective reactions realizations of these spaces are effectuated by directed enumerations. Enumerated (cyan) spaces refer to libraries that are fully enumerated. Currently, the largest virtual space is generated by GSK, while the largest enumerated space is the GDB. Proprietary spaces are spaces where either reactions, building blocks, or both, stem from the in-house collections of the respective pharmaceutical company, while public spaces are generated from reactions and building blocks that are broadly available. Vendor library spaces like GalaXi and Enamine REAL Space exist in both virtual and enumerated forms, although newer iterations featuring larger sizes, will likely be entirely virtual.

R-groups is smaller than that of the scaffold in order to not venture too far from the properties/activities the original molecule provided. As an alternative to R-group enumerations, the inverse methodology of keeping existing R-groups in place and exchanging the scaffolds in between, also known as corehopping, is an equally widely applied practice to explore the chemical space surrounding interesting chemical matter [23]. Other than starting from a pre-identified scaffold with a desired property profile, scaffolds can be obtained by iteratively stripping molecules bond-by-bond for a certain number of steps, by splitting molecules along retrosynthetic rulesets, or by removing a selected set of R-groups from a library of larger molecules [24, 25]. The sources for scaffold enumeration can range from proprietary libraries to public libraries, for which varied structures with experimental data for targets exist [26]. Combining both methodologies, the size of the enumerated libraries involving scaffold and R-group enumeration methodologies can easily grow into trillions of molecules and beyond. However, for practical purposes, the number of molecules explored in each enumeration using these methodologies generally ranges from tens to hundreds of thousands of ligands.

14.1.3 Bioisosteres and Matched Molecular Pairs

One specific category of R-group enumeration is called bioisostere replacement, where bioisosteres are defined as moieties that are chemically different but

recognized in a similar fashion on a biological level. Bioisosteres are categorized as either classical or nonclassical, with the former encompassing mono-, bi-, and trivalent atoms as well as simple ring system substitutions, exemplified by the widely popular exchange of hydrogen with fluorine [27]. While the latter describes more structurally complex interchanges, for example, a carboxylic acid into a tetrazole [28]. In practical terms, Bioisostere replacement utilizes corehopping and R-group enumerations in which chemical moieties are exchanged with one another in order to tackle problems of compound potency, selectivity, unfavorable properties, reducing or redirecting metabolism, eliminating or modifying toxicophores, as well as escaping pre-claimed/patented chemical entities [29, 30]. It is important to be aware that bioisostere replacements are very specific to the problem they are trying to address, and the transferability of bioisostere libraries between projects is by no means a certainty.

For methodologies like R-groups, corehopping, and bioisosteres, replacements can be obtained from preexisting libraries, generated combinatorially, or derived from a series of molecules via matched molecular pair analysis (MMPA) [31]. A matched molecular pair transformation is derived from a pair of molecules that differ by a defined structural element while the remainder of the molecules are identical [32]. Using this analysis, differences in properties can be easily interpreted from the change in structure. The output of an MMPA is a series of structural transforms, which in combination with their effect on the property of the molecule, can then be deployed to address the previously mentioned problems. Transforms resulting in no changes in overall properties can be utilized for the purpose of generating novel intellectual property while maintaining the overall profile of the compounds, whereas changes in properties can be utilized to improve the compounds. Analysis of libraries of molecules and extraction of the matched molecular pair transforms are popular tools used to guide compound optimization in drug discovery programs, and the number of transforms ranges from hundreds to hundreds of millions [33, 34].

14.1.4 Enumeration Within De Novo Design

While the previously shown methodologies kept at least one component identical between molecules, from a chemical space point of view, de novo design describes a strategy that specifically aims at generating previously unexplored regions of chemical space by assembling together entire molecules from smaller fragments [35, 36]. However, to derisk the methodology from the disadvantages of combinatorial chemistry, mostly the sheer amount of molecules possible, de novo design methodologies are conceptualized in such a manner as to steer the generation of new molecules to be synthetically tractable, provide the desired activity with a pharmaceutical target, and satisfy the overall property profiles of the compound. This is achieved through the application of a fitness function, which at every generative iteration filters out undesired solutions to reduce the number of molecules for the subsequent round of calculations.

To increase the likelihood of desired biological activity of the de novo-generated molecules, the decomposition of known active molecules can be utilized to generate

building block libraries. One popular early methodology for generating fragments from larger molecules is called the retrosynthetic combinatorial analysis procedure (RECAP) [25]. This method relies on the identification of bonds in a library of active molecules that can be split into building blocks according to a defined number of retrosynthetic rules. The resulting building blocks can then be reassembled into new molecules, satisfying the synthesizability constraints of the generated molecules. Within the constraints of classical Hansch analysis, the underlying assumption is that the reconstituted molecules provide preferred motifs, resulting in similarly active molecules while simultaneously being able to escape previously claimed chemical space [37]. In addition to the previously shown MMP analysis and RECAP, other methodologies for the generation of fragments have been developed in the past, like the scaffold tree decomposition method by Schuffenhauer et al. [24].

Combining fragments within the confines of tractable chemistry is a common way to reduce the combinatorial expansion of undesired solutions. Exemplified via the software SYNOPSIS, which starts from building blocks and expands them through applicable reactions, evaluates the resulting product for a given fitness function (dipole moment in the original article), and then adds the molecule to the result set, at which point the cycle is repeated for its next iteration [38]. It is important to keep in mind that most *de novo* design software differentiates at the following junctions: input fragments, applicable reactions, and fitness function/selection criteria to determine which solutions to keep.

Expanding upon the SYNOPSIS concept, Hartenfeller et al. published DOGS, which incorporated additional reactions in order to better access chemical space through more ring-closure reactions, as well as employing a similarity-based fitness function compared to the physics-based fitness function in SYNOPSIS [39].

Several other methodologies have been developed in the past that utilize libraries of fragments in order to combine them into novel molecules, and although fragment-based *de novo* design can be done entirely without protein, incorporating target information helps reduce enumerations by imposing a spatial constraint. These methods involve the enumeration of defined fragments in a semi-combinatorial fashion, either by starting from single fragments and extending them into larger druglike molecules or by linking multiple fragments in three-dimensional space [40–42].

Exchanging certain fragments from one molecule to another in order to improve properties and/or activity is a standard process in medicinal chemistry. BREED generates novel chemical space by interchanging fragments of a series of active molecules by aligning them beforehand and subsequently exchanging fragments of the molecules, thus creating novel molecules with the intention of retaining the favorable properties of the constituting fragments [42].

Similarly, the GROW algorithm starts by placing molecules in sub-pockets and growing into larger molecules that iteratively fill an increasing portion of the active site with the aim to increase specificity and activity one step at a time [41]. Conversely, linking starts from multiple fragments placed at different positions in the active site and tries to place linkers between the fragments to generate novel ligands that retain the original fragment's positions and combine the individual

fragment contributions to the overall binding affinity against the target [40]. Both the linking and the growing approaches can be confined within the rules of existing chemistry or simply by forming single bonds between the fragments, albeit at the risk of reducing the synthetic feasibility of the resulting molecule and increasing the likelihood of the aforementioned combinatorial explosions. Additionally, these methodologies are closely linked to structure-based design tools like docking, as the starting fragments for *de novo* design strategies are generally identified either in virtual or classical high-throughput fragment screens and stem from physically existing chemical stock.

14.2 Public and Commercial Chemical Libraries

The most common risk of enumerations of chemical libraries in drug discovery is the significant uncertainty of synthetic tractability of the molecules generated *in silico*. These concerns are partly addressed by operating computer-assisted drug discovery in chemical space that has been previously synthesized and can simply be purchased from a variety of vendors. For these purposes, annotated chemical databases comprising all available vendors provide the most significant starting point for the enumerations highlighted above.

The ChEMBL database is a publicly available collection of organic compounds and annotations for their activity and respective targets. At the time of writing, ChEMBL [43] contained 2.3 million distinct molecules derived from 85 000 publications with 19 million data points for activity against 15 072 biological targets (Figure 14.2) [43–45]. Generating and expanding subsets of molecules with activities against certain targets is a general practice, and multiple dedicated databases for drug targets have been generated over the years [46–48]. Currently, there are ongoing efforts to generate a database that hosts at least one molecule with activity against every possible drug target [49].

14.2.1 Vendor Libraries

The most popular aggregator vendor database is the ZINC database. Released in 2005 by Shoichet and Irwin, containing 727 842 molecules for download in a variety of formats and a web interface for searching a molecule's presence and purchasability (Figure 14.2) [50]. Initially, the database included compounds from 10 vendors that kept their inventories up-to-date in order to minimize the chances of hits in a virtual screening project being unavailable for purchase. Over time, the ZINC database grew to a size of ~900 million molecules in 2022 and hundreds of vendors [51, 52]. In addition to providing the molecules for search and download, the proprietors of the database also provide precalculated properties, and the database is confined to the boundaries of Lipinski's rules. Subsets based on desired property ranges can be generated, in order to only screen molecules with properties that satisfy the requirements of the drug discovery project, e.g. MW, TPSA, and *clogP*. In parallel to aggregators, the chemicals vendors contributing to the ZINC database are continuously increasing the number of molecules that are available [14].

Chemical vendors provide a reliable source of chemical matter for *in silico* drug discovery. The main advantage is the near-immediate availability of compounds, something that otherwise can cause a significant delay to drug discovery projects, i.e. when an ideal compound has been identified but the waiting time to confirm all predictions would take months. To reduce the overhead of maintaining a distributed list of vendors, several other aggregators have developed over time like Molport and eMolecules. These aggregators provide cataloged information on molecules' (multi-)vendor availability, shipping time, available quantity, price, and molecular properties. Classically, the number of molecules in the most reliable stock ranges from thousands to millions, with an ever-increasing degree of uncertainty in terms of available quantities and lead times.

With the reinvigorated interest in reaction-based enumeration technology and the incorporation of virtual molecules, the number of molecules that are available has increased dramatically [14]. In addition to existing libraries of organic compounds, virtual chemical vendor libraries are available that consist of molecules that have not yet been synthesized, but the availability of the corresponding building blocks and the chemistry that can be performed on them has led to a level of confidence for their synthesis so that they can be listed as on-demand stock [53].

As a result, the incorporation of virtual molecules has increased the chemical space available for purchase from millions to billions of molecules. The most prominent proprietors for virtual compounds are Enamine's REAL Space, WuXi's Galaxi, Otava's CHEMriya, and eMolecule's eXplore [54–57]. These databases are generated from a set of robust reactions that have a broad substrate scope in combination with libraries of relevant building blocks. These databases have seen increases in both scope and popularity in recent years, primarily because the synthetic tractability of the virtual compounds has increased, which increases the chances that these compounds can be synthesized in a timely fashion. The two components that confine the space of reaction-based libraries are the availability of building blocks together with the corresponding reactions that can be applied to these building blocks. As an example, the PathFinder space of Schrödinger currently encompasses 10^{15} molecules (Figure 14.2) that are not explicitly stored as the enumerated products but can be generated by combining all available reactions with their corresponding classes of building blocks. Consequently, the expansion of these spaces can be driven both on the side of the reactions as well as the corresponding building blocks.

By combining several large vendor libraries, we have assembled a virtual screening database of more than 28 billion unique molecules (Figure 14.3) for which we calculated a series of physicochemical properties and investigated their respective distributions. The distribution of hydrogen bond acceptors (HBA) follows a classical Gaussian distribution with most of the compounds containing four to six acceptors, while for hydrogen bond donors (HBD) the majority of molecules feature zero to three donors. Similar to the HBA distribution, the polar surface area (PSA) and the number of rotatable bonds (RB) follow a similar Gaussian distribution with the maximum values residing between 75 and $100 \text{ \AA}^2$ for the PSA and 4–7 for RB, respectively. The molecular hydrophobicity follows a similar distribution with most compounds featuring an AlogP value between 2 and 2.5, while more than

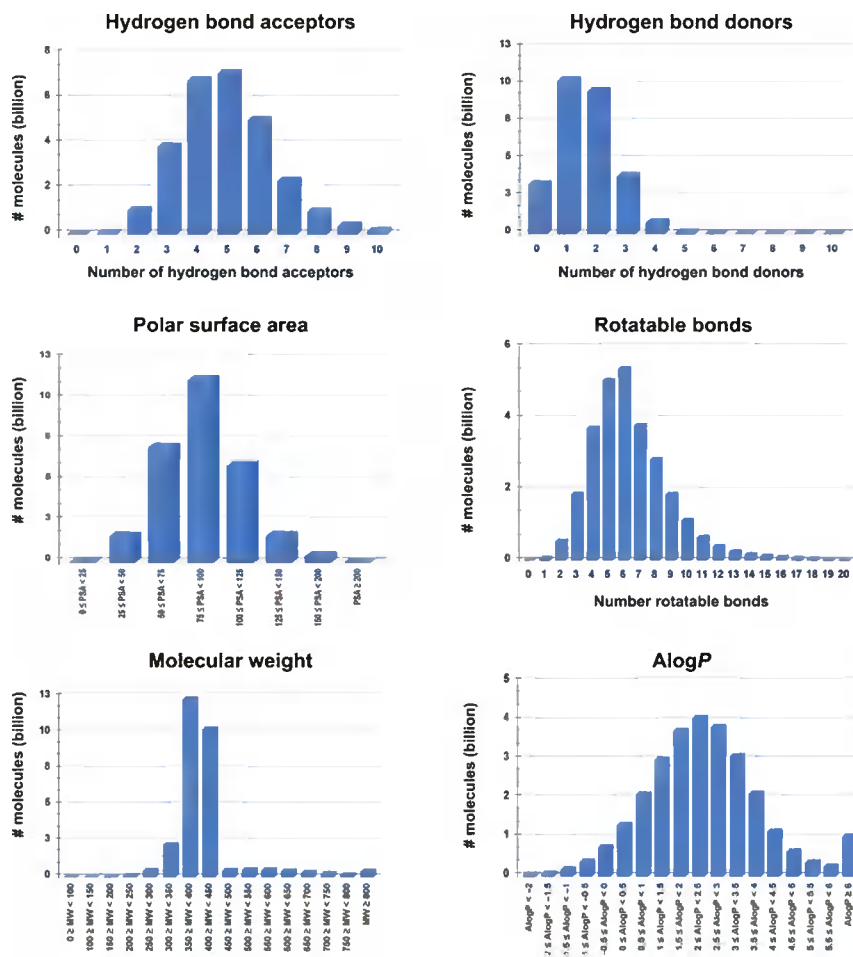


Figure 14.3 Overview of the Lipinski properties (HBA, HBD, rotatable bonds, MW, AlogP) and the polar surface area (PSA) of the combined vendor libraries from Mcule, WuXi, Ottawa, EnamineREAL, Enamine REAL Space, eMolecules, Chemspace, Molport, and Aldrich Market Select, exceeding 28 billion molecules.

a billion molecules feature an AlogP value larger than 6, indicating significant chemical matter beyond the Lipinski boundary for that property. The MW distribution shows that nearly the entire database consists of molecules weighing between 300 and 500 Da, indicating that comparatively few small molecular fragments are present in the database. Chemical space below 300 Da is simply more limited due to the finite combination of atoms.

By applying the following cut-off criteria for $\text{AlogP} \leq 3$, $\text{MW} > 110$, $\text{PSA} \leq 110$, $\text{HBD} \leq 3$, $\text{HBA} \leq 5$, $\text{RB} \leq 3$, chiral centers ≥ 1 , and heavy atoms ≤ 18 , the resulting number of fragments in this database was determined to be 39 527 646. Setting the cut-off criteria to $0 \geq \text{AlogP} \leq 3$, $250 \geq \text{MW} \leq 375$, $\text{PSA} < 110$, $\text{HBD} \leq 2$, $\text{HBA} \leq 5$, $\text{RB} \leq 3$, and chiral centers ≤ 1 , the number of lead like molecules is 581 290 435.

Conversely, with the following cut-off criteria $-1 \geq \text{AlogP} \leq 4$, $250 \geq \text{MW} \leq 500$, $50 \geq \text{PSA} \leq 130$, $\text{HBD} \leq 5$, $\text{HBA} \leq 10$, $\text{RB} \leq 10$, and chiral centers ≤ 3 , the number of druglike molecules in this database is 14 654 960 293.

14.2.2 Reaction-Based Chemical Spaces

In addition to commercial offerings, pharmaceutical companies have utilized their in-house building block collections and available chemistry to generate virtual chemical spaces for screening purposes in order to enhance drug discovery efforts.

In 2016, Nicolaou and colleagues presented the Proximal Lilly Collection (PLC), a virtual synthesis engine combining reactions belonging to 10 (recently 25) broader reaction classes and compatible building blocks in the range of ten to hundreds of thousands [7, 58]. The rapid enumeration engine utilizing these reactions is able to provide 10^6 – 10^9 molecules for screening out of an estimated space size of 10^{11} (Figure 14.2), which certainly could be expanded even further by increasing the reaction classes and/or the size of the building block libraries. Additionally, searches of molecules from Drugbank (6059), Pubchem (43 629 726), and an in-house Lilly collection (100 000) in the PLC resulted in retrieval rates of 3.2%, 20.0%, and 23.2%, respectively, illustrating the PLC's ability to generate molecules that are pharmaceutically relevant.

AstraZeneca has its own virtual inventory derived from 4100 in-house synthesis protocols and complementary building blocks estimated between 10^{15} originally published and 10^{17} with the addition of Enamine Building Blocks [59, 60]. Similarly, the Pfizer Virtual Global Library, an in-house developed reaction-based enumeration engine, relies on 1244 virtual reactions, consisting of 436, 725, and 83 2-, 3-, and 4-component reactions respectively. Taking the Pfizer building blocks, the space consists of an estimated 10^{14} molecules but could reach 10^{18} when the Available Chemicals Directory was incorporated [61].

BioSolveIT has generated the Knowledge Space (Figure 14.2), which utilizes more than 100 published reactions to generate a virtual space containing 10^{14} molecules [7, 62]. Based on the Feature Trees methodology, the Merck MASSIV space was generated along similar lines, involving a series of some hundred robust reactions and the full catalog of Sigma Aldrich compounds, as well as Merck KGaA Healthcare's in-house building block inventory, to produce a virtual space for screening in the range of 10^{20} molecules [63]. In a similar fashion, GlaxoSmithKline created a virtual space originally encompassing 10^{15} molecules, containing 296 individual chemical sub-spaces from 120 individual reactions, and is slated to be 10^{26} molecules, currently representing the most sizable virtual chemical space [7].

In contrast, companies like Evotec and Boehringer Ingelheim created virtual chemical spaces on comparatively smaller scales: 10^{16} (EvoSpace) and 10^{11} (BICLAIM 2008 deploying in-house validated combinatorial libraries contrary to reactions). Although smaller, in both cases the virtual chemical space is still large enough to make explicit storage a very costly undertaking [14].

Contrary to the assumption that there may be significant overlap between these spaces, an analysis by Lessel and Lemmen has shown that the overlap between

the Enamine REAL Space, BioSolveIT's KnowledgeSpace, and a recent version of Boehringer Ingelheim's BICLAIM spaces is only 45, 89, and 1523 molecules after having enumerated a subset of 1 million molecules by similarity to 100 reference molecular probes [64]. The overlap between all three spaces amounted to merely three molecules, which stands to reason that comparing 10^6 molecules of spaces sized from 10^9 to 10^{20} may not reveal the full picture. The authors point out that with the potential chemical space, beyond Lipinski's rules, containing 10^{180} molecules, the subsets are simply too small to compare the overlap. It may also indicate differences between either the building blocks used for this reaction-based enumeration technique or the reactions that combine these fragments. However, it is likely that with more molecules enumerated from these spaces, an increase in the overlap is expected.

In a different comparison performed by Bellmann et al. on Enamine's REAL Space, WuXi's GalaXi Space, and Otava's CHEMriya, it was shown that only 76 524 molecules are identical in all three virtual spaces of sizes ranging between 10^9 and 10^{10} . However, the overlap between the two spaces respectively was much larger, up to 10^7 [15].

14.2.2.1 Reaction Sources

Apart from the textbook knowledge, experience, and chemical intuition each synthetic chemist picks up over the course of their career, approaching a (retro)synthetic problem algorithmically significantly reduces the number of possible reactions a chemist would have to recall. Interestingly, despite a plethora of chemical reactions being available in the form of publications and textbooks, there are surprisingly few publicly available machine-readable generalized reaction examples that can be found in the literature.

While not only proposing the concepts of retrosynthetic analysis, E. J. Corey also engaged in computer-aided retrosynthesis with the development of a computer program called LHASA (Logic and Heuristics Applied to Synthetic Analysis) [65]. Within LHASA, reactions were coded in the Pattern Transformation/Chemical Transformation (PATRAN/CHMTRN) language, in which not only the transformation rule from retron into synthons was defined but also the probability of success of the reaction, functional group alerts, and even protective group chemistry measures. The database at its end consisted of several thousand proprietary reactions compiled by synthetic chemists. Most recently, the reactions have been revived to be used in the CACTVS software interface, and a subset of these reactions have been employed in the generation of the publicly available reaction-based database synthetically accessible virtual inventory (SAVI) [66–68]. This database currently contains 1.7 billion molecules (Figure 14.2). The advantage of the database is that on top of the binary classification of the molecules as synthesizable by them simply being present in the database, there is also the incorporation of the prediction of synthesizability annotated for each reaction, resulting in subsets indicating which molecules have the highest confidence of being synthetically feasible.

Additionally, refining generalized reactions by functional group (in-)compatibilities and protective chemistry requirements has not stopped with LHASA but was

continued within the academic CHEMATICA retro-synthetic software, which in its current commercial iteration, SYNTHIA, comprises more than 100 000 reactions, which could also serve as templates for enumeration [69, 70]. The commercial REAXYS and SPRESI databases contain 57 and 4.6 million reactions, respectively, which could be generalized for reaction-based enumerations and could provide substantial input generating chemical space, which currently has yet to be assessed for the potential virtual spaces they would enable [71, 72].

One example of publicly available reactions for enumerations is the small collection of reactions by Hartenfeller et al. from 2011 incorporated within the previously mentioned DOGS [73]. The dataset consists of 58 of the most robust and widely used chemical reactions, and half of the reactions represent ring-forming reactions with the intention to increase the generation of novel chemotypes. The set of reactions can be utilized to enumerate large libraries by simply enumerating all building blocks against all other building blocks using the respective reactions, or to derivatize a base scaffold by keeping one reactant fixed and varying the other reactant(s). The most prominent database constructed from these reactions is called SCUBIDOO (Figure 14.2) and contains 21 million virtual molecules [74].

Another available source for reactions is the RetrotransformDB by Avramova et al. [75]. Similar to the Hartenfeller reactions, the authors took a series of popular reactions and annotated them in the SMIRKS format for public use in order to perform a simple retrosynthetic analysis. Reversed, these reactions could then be used to run enumerations on available building blocks in order to generate chemically feasible libraries for virtual screening.

In addition to publicly available reactions, pharmaceutical companies also have a plethora of reactions present in their in-house chemical databases. Despite a lack of generalized transforms published by the industry for public use, statistical analysis of the patented reaction space shows that the number of reactions used within the pharmaceutical industry remained comparatively small, thus indicating that the publicly available transforms by Hartenfeller and Avramova provide a solid basis for reaction-based enumeration that only requires limited expansion to be comparable to industry efforts. By performing a generalized reaction analysis, Boström and Brown have shown that the majority of synthetic chemistry published in patents between 1984 and 2014 has not changed between reactions [76]. Similarly, in 2016, a reaction-fingerprint-based analysis of the largest database of patented reactions showed that most reactions can be classified as belonging to the same 10–400 categories that have remained consistent over the last 40 years [77].

14.2.2.2 Available Building Blocks

Once a series of reactions has been assembled, the next step is to identify available building blocks that are suitable for these reactions. The sources of building blocks are usually the same as for the larger screening compounds, namely the aforementioned aggregators/vendors of organic compounds. Specific subsets are generally created by applying cut-offs to MW and/or other physicochemical properties, or by subsetting based on the presence of certain functional groups that would enable a specific reaction (e.g. Boronic acids for a Suzuki cross-coupling reaction).

Commercial vendors, as highlighted above, that provide larger molecules for screening have adapted to this general request and provide specifically tailored subsets of building blocks. The definition of a building block is usually derived from its size (below 300 Da) and properties (or fragment properties), or explicit presence of functional groups enabling specific reactions [78].

A recent analysis by Varnek and colleagues analyzed the overall properties of eMolecules building blocks by classifying them into synthons for several reactant classes and subsequently comparing their make-up, presence in eMolecules and ChEMBL building blocks in terms of supported reactions and overlap [79]. Similarly, scientists at AbbVie recently analyzed their internal building block collections and discovered that the majority of reactions belong to the class of amide coupling reactions, with carboxylic acids and amines being the most represented building block classes, echoing the results of the analysis by Varnek et al. [80]. Following this work, we analyzed the compounds from a combined building block library made up of 30 vendor catalogs (Table 14.1). In total, 476 million molecules were classified via SMARTS-based substructure searches into 45 reactant classes suitable for the 152 reactions that are currently part of Schrödinger's PathFinder reaction-based enumeration tool. Similar to the results found by Varnek and AbbVie, amines represent the largest group of reactants with 271 493 071 entries, followed by alcohols and halides on the same order of magnitude. These results are consistent with a small number of reliable reactions providing most of the chemistry deployed for novel drugs [82]. Carboxylic acids represent a reactant group operating on the order of 10^7 . The smallest amounts of reactants contributing 3, 16, and 283 respectively are cyanoacetamides, isocyanoacetates, and silanes. In a simplified combinatorial estimate, one can calculate that the chemical space produced by a Suzuki cross-coupling reaction of aryl bromides and aryl boronates already surpasses 10^{14} molecules.

14.3 How to Effectively Explore Chemical Space?

Recently, we published on the development of a novel approach toward chemical space exploration titled AutoDesigner [34]. AutoDesigner is a fully integrated design algorithm for the ultra-large-scale exploration of chemical space that combines multiple compound enumeration strategies with an advanced filtering cascade on a scalable cloud-based platform. This *de novo* design algorithm aims to solve two of the major limitations of traditional medicinal chemistry approaches and/or structure-based drug design (SBDD): (i) difficulty in rapidly designing potent molecules that adhere to a wide range of project criteria, also known as the multiparameter optimization (MPO) problem, and (ii) the relatively small number of molecules explored compared to the vast size of chemical space. Millions to billions of virtual molecules are explored and optimized while adhering to project criteria such as physicochemical properties and potency. The algorithm only requires a single ligand with measurable affinity and a putative binding model as a starting point, so it can be employed in the early stages of an SBDD project where

Table 14.1 Overview of the number of building blocks and corresponding reactant classes according to PathFinder definitions (Schrödinger Release 2022-3: Maestro, Schrödinger LLC, New York, NY, 2021.) definitions for a set of 476 million building blocks originating from 30 sources (Ambinter, AOB Chemicals, ChemBridge, ChemDiv, Chemical Block, ChemSpace, CNH Technologies, ComInnex, eMolecules, Enamine, FCH Chemicals, InnovaPharm, InterBioScreen, Key Organics BIONET, Life Chemicals, Liverpool, ChiroChem, Maybridge, Mcule, Molport, Otava, Princeton BioMolecular Research, SAVI, Specs, SpiroChem, TimTec, Vitas, WuXi, X-Chem, ZINC).

Reactant class	# building blocks	Reactant class	# building blocks
total BBs	476 059 457	esters	10 638 194
5het_H1	6 058 351	friedel_crafts_substrates	35 326 493
acetophenones_2OH	14 757	glycines_N_acyl	8640
acid_chlorides	53 398	halides ^{a)}	104 559 584
acrylates_amino	22 259	heterocycles_NH	12 277 445
alcohols	102 234 858	hydrazides	668 389
aldehydes	2 503 652	hydrazines	6 347 789
alkynes_monosub	7 127 303	isocyanates	83 239
amides_prim	5 417 166	isocyanoacetates	16
amidines	1 228 018	ketones ^{a)}	13 137 586
amines ^{a)}	271 493 071	nitriles	11 490 795
anilines_prim	18 868 733	nitro ^{a)}	5 674 743
aryls_vilsmeier	18 872 601	organostannanes	8862
azides	284 012	phenols	7 779 897
azoles_nH	9 841 620	pyrazolones	23 002
benzoates_2amino	535 853	quinazolinones	61 333
benzoicacids_2acyl	17 253	silanes ^{a)}	282
boronates_aryl	4 119 980	styrenes	766 205
bromides_aryl	43 146 889	sulfonamides_primsec	8 668 226
carboxylates	52 394 503	sulfonyl_chlorides	287 783
cyanoacetamides	3	tetrazoles ^{a)}	341 752
cyanohydrins	14 332	thio ^{a)}	4 160 567
cyclohexanones	756 293	trifluoroborates	328 634

a) indicates a combined reaction class based on a common structural moiety (i.e. thio refers to thioethers and thioureas).

Source: Adapted from Konze et al. [81].

limited data are available. AutoDesigner utilizes a combination of three different enumeration methods: (i) matched molecular pair transforms; (ii) reaction-based enumeration; and (iii) multiple rounds of R-group decoration. After each round of enumeration and chemical space expansion, the enumerated compounds are funneled through a filtering cascade comprising physicochemical property filters, structure-based filters, and docking-based filters.

To assess the effectiveness of AutoDesigner, its application in the design of novel inhibitors of D-amino acid oxidase (DAO), a target for the treatment of schizophrenia, was reported [34, 83]. AutoDesigner was able to generate and efficiently explore over 1 billion molecules per design cycle to successfully address a variety of project goals. The reported data demonstrated that AutoDesigner can play a key role in accelerating the discovery of novel, potent chemical matter within the constraints of a given drug discovery lead optimization campaign.

With the ever-expanding size of chemical space, technologies, such as AutoDesigner, that are able to explore a highly targeted area of chemical space in a very efficient and thorough manner, are of increasing importance in SBDD.

References

- 1 Lipinski, C.A., Lombardo, F., Dominy, B.W., and Feeney, P.J. (1997). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* 23 (1): 3–25.
- 2 Carlesi, E., Hoffman, Y., and Libeskind, N.I. (2022). Estimation of the masses in the local group by gradient boosted decision trees. *Mon. Not. R. Astron. Soc.* 513 (2): 2385–2393.
- 3 *Sun Fact Sheet*. [Internet]. [cited 2022 Nov 10]. Available from: <https://nssdc.gsfc.nasa.gov/planetary/factsheet/sunfact.html>.
- 4 Bohacek, R.S., McMartin, C., and Guida, W.C. (1996). The art and practice of structure-based drug design: a molecular modeling perspective. *Med. Res. Rev.* 16 (1): 3–50.
- 5 Ertl, P. (2002). *Cheminformatics Analysis of Organic Substituents: Identification of the Most Common Substituents, Calculation of Substituent Properties, and Automatic Identification of Drug-Like Bioisosteric Groups*. ACS Publications. American Chemical Society [Internet]. [cited 2022 Nov 10]. Available from: <https://pubs.acs.org/doi/pdf/10.1021/ci0255782>.
- 6 Walters, W.P., Stahl, M.T., and Murcko, M.A. (1998). Virtual screening—an overview. *Drug Discovery Today* 3 (4): 160–178.
- 7 Warr W. *Report on an NIH Workshop on Ultralarge Chemistry Databases*. 2021 [cited 2022 Nov 10]. Available from: <https://chemrxiv.org/engage/chemrxiv/article-details/60c75883bdbb89984ea3ada5>
- 8 Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* 28 (1): 31–36.
- 9 Weininger, D., Weininger, A., and Weininger, J.L. (1989). SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.* 29 (2): 97–101.
- 10 Schüller, A., Schneider, G., and Byvatov, E. (2003). SMILIB: rapid assembly of combinatorial libraries in SMILES notation. *QSAR Comb. Sci.* 22 (7): 719–721.
- 11 Schüller, A., Hähnke, V., and Schneider, G. (2007). SmlLib v2.0: a Java-based tool for rapid combinatorial library enumeration. *QSAR Comb. Sci.* 26 (3): 407–410.

- 12 *Daylight Theory: SMARTS – A Language for Describing Molecular Patterns*. [Internet]. [cited 2022 Oct 14]. Available from: <https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html>.
- 13 *RDKit: Open-Source Cheminformatics*. [Internet]. [cited 2022 Oct 14]. Available from: <http://www.rdkit.org>.
- 14 Warr, W.A., Nicklaus, M.C., Nicolaou, C.A., and Rarey, M. (2022). Exploration of ultralarge compound collections for drug discovery. *J. Chem. Inf. Model.* 62 (9): 2021–2034.
- 15 Bellmann, L., Penner, P., Gastreich, M., and Rarey, M. (2022). Comparison of combinatorial fragment spaces and its application to ultralarge make-on-demand compound catalogs. *J. Chem. Inf. Model.* 62 (3): 553–566.
- 16 *infiniSee*. [Internet]. BioSolveIT. [cited 2022 Nov 10]. Available from: <https://www.biosolveit.de/products/infinisee>.
- 17 Ruddigkeit, L., van Deursen, R., Blum, L.C., and Reymond, J.L. (2012). Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J. Chem. Inf. Model.* 52 (11): 2864–2875.
- 18 Blum, L.C., van Deursen, R., and Reymond, J.L. (2011). Visualisation and subsets of the chemical universe database GDB-13 for virtual screening. *J. Comput.-Aided Mol. Des.* 25 (7): 637–647.
- 19 Ruddigkeit, L., Awale, M., and Reymond, J.L. (2014). Expanding the fragrance chemical space for virtual screening. *J. Cheminf.* 6 (1): 27.
- 20 Fink, T. and Reymond, J.L. (2007). Virtual exploration of the chemical universe up to 11 atoms of C, N, O, F: assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes, and drug discovery. *J. Chem. Inf. Model.* 47 (2): 342–353.
- 21 Blum, L.C. and Reymond, J.L. (2009). 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J. Am. Chem. Soc.* 131 (25): 8732–8733.
- 22 Krippendorff, Klaus. *A Dictionary of Cybernetics*. 1986.
- 23 Zhang, L.S., Wang, S.Q., Xu, W.R. et al. (2012). Scaffold-based Pan-agonist design for the PPAR α , PPAR β and PPAR γ receptors. *PLoS One* 7 (10): e48453.
- 24 Schuffenhauer, A., Ertl, P., Roggo, S. et al. (2007). The scaffold tree–visualization of the scaffold universe by hierarchical scaffold classification. *J. Chem. Inf. Model.* 47 (1): 47–58.
- 25 Lewell, X.Q., Judd, D.B., Watson, S.P., and Hann, M.M. (1998). RECAP–retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* 38 (3): 511–522.
- 26 Langdon, S.R., Brown, N., and Blagg, J. (2011). Scaffold diversity of exemplified medicinal chemistry space. *J. Chem. Inf. Model.* 51 (9): 2174–2185.
- 27 Patani, G.A. and LaVoie, E.J. (1996). Bioisosterism: a rational approach in drug design. *Chem. Rev.* 96 (8): 3147–3176.
- 28 Meanwell, N.A. (2011). Synopsis of some recent tactical application of bioisosteres in drug design. *J. Med. Chem.* 54 (8): 2529–2591.

- 29 Wagener, M. and Lommerse, J.P.M. (2006). The quest for bioisosteric replacements. *J. Chem. Inf. Model.* 46 (2): 677–685.
- 30 Hamada, Y. and Kiso, Y. (2012). The application of bioisosteres in drug design for novel drug discovery: focusing on acid protease inhibitors. *Expert Opin. Drug Discovery* 7 (10): 903–922.
- 31 Kenny, P.W. and Sadowski, J. (2005). Structure modification in chemical databases. In: *Chemoinformatics in Drug Discovery*, 271–285. John Wiley & Sons, Ltd [Internet]. [cited 2022 Oct 14]. Available from: <https://onlinelibrary.wiley.com/doi/abs/10.1002/3527603743.ch11>.
- 32 Tyrchan, C. and Evertsson, E. (2017). Matched molecular pair analysis in short: algorithms, applications and limitations. *Comput. Struct. Biotechnol. J.* 15: 86–90.
- 33 Dalke, A., Hert, J., and Kramer, C. (2018). Mmpdb: an open-source matched molecular pair platform for large multiproperty data sets. *J. Chem. Inf. Model.* 58 (5): 902–910.
- 34 Bos, P.H., Houang, E.M., Ranalli, F. et al. (2022). AutoDesigner, a de novo design algorithm for rapidly exploring large chemical space for lead optimization: application to the design and synthesis of d-amino acid oxidase inhibitors. *J. Chem. Inf. Model.* 62 (8): 1905–1915.
- 35 Schneider, G. and Fechner, U. (2005). Computer-based de novo design of drug-like molecules. *Nat. Rev. Drug Discovery* 4 (8): 649–663.
- 36 Hartenfeller, M. and Schneider, G. (2011). Enabling future drug discovery by de novo design. *WIREs Comput. Mol. Sci.* 1 (5): 742–759.
- 37 Corwin, H. and Toshio, F. (1964). p- σ - π analysis. A method for the correlation of biological activity and chemical structure. *J. Am. Chem. Soc.* 86 (8): 1616–1626.
- 38 Vinkers, H.M., de Jonge, M.R., Daeyaert, F.F.D. et al. (2003). SYNOPSIS: SYNthesize and OPTimize System in Silico. *J. Med. Chem.* 46 (13): 2765–2773.
- 39 Hartenfeller, M., Zettl, H., Walter, M. et al. (2012). DOGS: reaction-driven de novo design of bioactive compounds. *PLoS Comput. Biol.* 8 (2): e1002380.
- 40 Dey, F. and Caflisch, A. (2008). Fragment-based de novo ligand design by multi-objective evolutionary optimization. *J. Chem. Inf. Model.* 48 (3): 679–690.
- 41 Moon, J.B. and Howe, W.J. (1991). Computer design of bioactive molecules: a method for receptor-based de novo ligand design. *Proteins Struct. Funct. Bioinf.* 11 (4): 314–328.
- 42 Pierce, A.C., Rao, G., and Bemis, G.W. (2004). BREED: generating novel inhibitors through hybridization of known ligands. Application to CDK2, P38, and HIV protease. *J. Med. Chem.* 47 (11): 2768–2775.
- 43 ChEMBL Database. [Internet]. [cited 2022 Nov 10]. Available from: www.ebi.ac.uk/chembl.
- 44 Davies, M., Nowotka, M., Papadatos, G. et al. (2015). ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic Acids Res.* 43 (Web Server issue): W612–W620.
- 45 Mendez, D., Gaulton, A., Bento, A.P. et al. (2019). ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* 47 (D1): D930–D940.

- 46 Carles, F., Bourg, S., Meyer, C., and Bonnet, P. (2018). PKIDB: a curated, annotated and updated database of protein kinase inhibitors in clinical trials. *Mol. J. Synth. Chem. Nat. Prod. Chem.* 23 (4): 908.
- 47 Qi, Y., Wang, D., Wang, D. et al. (2016). HEDD: the human epigenetic drug database. *Database J. Biol. Databases Curation* 2016: baw159.
- 48 Torchet, R., Druart, K., Ruano, L.C. et al. (2021). The iPPI-DB initiative: a community-centered database of protein–protein interaction modulators. *Bioinformatics* 37 (1): 89–96.
- 49 Ackloo, S., Al-awar, R., Amaro, R.E. et al. (2022). CACHE (Critical Assessment of Computational Hit-finding Experiments): a public–private partnership benchmarking initiative to enable the development of computational methods for hit-finding. *Nat. Rev. Chem.* 6 (4): 287–295.
- 50 Irwin, J.J. and Shoichet, B.K. (2004). ZINC – a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* 45 (1): 177–182.
- 51 Irwin, J.J., Tang, K.G., Young, J. et al. (2020). ZINC20—a free ultralarge-scale chemical database for ligand discovery. *J. Chem. Inf. Model.* 60 (12): 6065–6073.
- 52 Tingle B, Tang K, Castanon J, Gutierrez J, Khurelbaatar M, Dandarchuluun C, et al. ZINC-22 – A Free Multi-Billion-Scale Database of Tangible Compounds for Ligand Discovery. 2022 [cited 2022 Nov 10]. Available from: <https://chemrxiv.org/engage/chemrxiv/article-details/634f2185dfbd2bbe525b876a>
- 53 Grygorenko, O.O., Radchenko, D.S., Dziuba, I. et al. (2020). Generating multi-billion chemical space of readily accessible screening compounds. *iScience* 23 (11): 101681.
- 54 *REAL Space – Enamine*. [Internet]. [cited 2022 Nov 10]. Available from: <https://enamine.net/compound-collections/real-compounds/real-space-navigator>.
- 55 *12 Billion Novel Molecules: CHEMriya – OTAVA’s On-Demand Chemical Space*. [Internet]. [cited 2022 Nov 10]. Available from: <https://www.otavachemicals.com/products/chemriya>.
- 56 *WuXi AppTec Research Service Division and BioSolveIT Introduce GalaXi®, a Vast New Chemical Space of Tangible Molecules – WuXi XPress: for WuXi News and R&D Insights*. [Internet]. [cited 2022 Nov 10]. Available from: <https://wxpress.wuxiapptec.com/wuxi-apptec-research-service-division-and-biosolveit-introduce-galaxi-a-vast-new-chemical-space-of-tangible-molecules>.
- 57 *eXplore*. [Internet]. [cited 2022 Nov 10]. Available from: <https://marketing.emolecules.com/explore>.
- 58 Nicolaou, C.A., Watson, I.A., Hu, H., and Wang, J. (2016). The proximal Lilly collection: mapping, exploring and exploiting feasible chemical space. *J. Chem. Inf. Model.* 56 (7): 1253–1266.
- 59 Vainio, M.J., Kogej, T., and Raubacher, F. (2012). Automated recycling of chemistry for virtual screening and library design. *J. Chem. Inf. Model.* 52 (7): 1777–1786.
- 60 Grebner, C., Malmerberg, E., Shewmaker, A. et al. (2019). Virtual screening in the cloud: how big is big enough? *J. Chem. Inf. Model.* [Internet]. [cited 2022 Nov 11]. Available from: <https://pubs.acs.org/doi/pdf/10.1021/acs.jcim.9b00779>.

- 61 Hu, Q., Peng, Z., Sutton, S.C. et al. (2012). *Pfizer Global Virtual Library (PGVL): A Chemistry Design Tool Powered by Experimentally Validated Parallel Synthesis Information*. ACS Publications. American Chemical Society [Internet]. [cited 2022 Nov 10]. Available from: <https://pubs.acs.org/doi/pdf/10.1021/co300096q>.
- 62 Detering, C., Claussen, H., Gastreich, M., and Lemmen, C. (2010). KnowledgeSpace – a publicly available virtual chemistry space. *J. Cheminf.* 2 (1): O9.
- 63 Hoffmann, T. and Gastreich, M. (2019). The next level in chemical space navigation: going far beyond enumerable compound libraries. *Drug Discovery Today* 24 (5): 1148–1156.
- 64 Lessel, U. and Lemmen, C. (2019). Comparison of large chemical spaces. *ACS Med. Chem. Lett.* [Internet]. [cited 2022 Nov 10]. Available from: <https://pubs.acs.org/doi/pdf/10.1021/acsmchemlett.9b00331>.
- 65 Pensak, D.A. and Corey, E.J. (1977). LHASA—logic and heuristics applied to synthetic analysis. In: *Computer-Assisted Organic Synthesis*, ACS Symposium Series, vol. 61, 1–32. American Chemical Society [Internet]. [cited 2022 Nov 11]. Available from: doi: <https://doi.org/10.1021/bk-1977-0061.ch001>.
- 66 Patel H, Ihlenfeldt W, Judson P, Moroz YS, Pevzner Y, Peach M, et al. *Synthetically Accessible Virtual Inventory (SAVI)*. 2020 [cited 2022 Nov 11]. Available from: <https://chemrxiv.org/engage/chemrxiv/article-details/60c74a539abda204cdf8ce1e>
- 67 Patel, H., Ihlenfeldt, W.D., Judson, P.N. et al. (2020). SAVI, in silico generation of billions of easily synthesizable compounds through expert-system type rules. *Sci. Data* 7 (1): 384.
- 68 Judson, P.N., Ihlenfeldt, W.D., Patel, H. et al. (2020). Adapting CHMTRN (ChEMistry TRaNslator) for a new use. *J. Chem. Inf. Model.* [Internet]. [cited 2022 Nov 11]. Available from: <https://pubs.acs.org/doi/pdf/10.1021/acs.jcim.0c00448>.
- 69 Grzybowski, B.A., Szymkuć, S., Gajewska, E.P. et al. (2018). Chematica: a story of computer code that started to think like a chemist. *Chem* 4 (3): 390–398.
- 70 Molga, K., Szymkuć, S., Gołębiowska, P. et al. (2022). A computer algorithm to discover iterative sequences of organic reactions. *Nat. Synth.* 1 (1): 49–58.
- 71 *Reaxys*. [Internet]. [cited 2022 Nov 11]. Available from: <https://www.reaxys.com/#/search/quick>.
- 72 Roth, D.L. (2005). SPRESIweb 2.1, a selective chemical synthesis and reaction database. *J. Chem. Inf. Model.* 45 (5): 1470–1473.
- 73 Hartenfeller, M., Eberle, M., Meier, P. et al. (2011). A collection of robust organic synthesis reactions for in silico molecule design. *J. Chem. Inf. Model.* 51 (12): 3093–3098.
- 74 Chevillard, F. and Kolb, P. (2015). SCUBIDOO: a large yet screenable and easily searchable database of computationally created chemical compounds optimized toward high likelihood of synthetic tractability. *J. Chem. Inf. Model.* 55 (9): 1824–1835.
- 75 Avramova, S., Kochev, N., and Angelov, P. (2018). RetroTransformDB: a dataset of generic transforms for retrosynthetic analysis. *Data* 3 (2): 14.

- 76** Brown, D.G. and Boström, J. (2016). Analysis of past and present synthetic methodologies on medicinal chemistry: where have all the new reactions gone? *J. Med. Chem.* 59 (10): 4443–4458.
- 77** Schneider, N., Lowe, D.M., Sayle, R.A. et al. (2016). Big data from pharmaceutical patents: a computational analysis of medicinal Chemists' bread and butter. *J. Med. Chem.* 59 (9): 4385–4402.
- 78** Congreve, M., Carr, R., Murray, C., and Jhoti, H. (2003). A 'Rule of Three' for fragment-based lead discovery? *Drug Discovery Today* 8 (19): 876–877.
- 79** Zabolotna, Y., Volochnyuk, D.M., Ryabukhin, S.V. et al. (2022). A close-up look at the chemical space of commercially available building blocks for medicinal chemistry. *J. Chem. Inf. Model.* 62 (9): 2171–2185.
- 80** Wang, Y., Haight, I., Gupta, R., and Vasudevan, A. (2021). What is in our kit? An analysis of building blocks used in medicinal chemistry parallel libraries. *J. Med. Chem.* [Internet]. [cited 2022 Nov 11]. Available from: <https://pubs.acs.org/doi/pdf/10.1021/acs.jmedchem.1c01139>.
- 81** Konze, K.D., Bos, P.H., Dahlgren, M.K. et al. (2019). Reaction-based enumeration, active learning, and free energy calculations to rapidly explore synthetically tractable chemical space and optimize potency of cyclin-dependent kinase 2 inhibitors. *J. Chem. Inf. Model.* 59 (9): 3782–3793.
- 82** Boström, J., Brown, D.G., Young, R.J., and Keserü, G.M. (2018). Expanding the medicinal chemistry synthetic toolbox. *Nat. Rev. Drug Discovery* 17 (10): 709–727.
- 83** Tang, H., Jensen, K., Houang, E. et al. (2022). Discovery of a novel class of d-amino acid oxidase inhibitors using the Schrödinger computational platform. *J. Med. Chem.* 65 (9): 6775–6802.

15

Navigating Chemical Space

Ákos Tarcsay, András Volford, Jonathan Buttrick, Jan-Constantin Christopherson, Máté Erdős, and Zoltán B. Szabó

Chemaxon Kft., Váci út 133, 1138, Budapest, Hungary

15.1 Introduction to Chemical Space

Drug discovery workflows iteratively construct scientific rationale and hypotheses based on available evidence to ideate and prioritize the next round of compounds to be tested experimentally. New data points are stored, analyzed, and evaluated in the context of chemical structural modifications or assessed against existing compounds to prove or disprove the hypothesis. This cycle utilizes different types of data depending on its current step.

Discovery projects generate multi-dimensional data with structural and nonstructural information, including registration metadata, and measured and/or calculated physicochemical properties, and extend these with project-related information. An array of experimental data is linked to the structures with increasing density of information as a compound progresses in the screening cascade, especially in the case of extensively studied reference and front-runner compounds. On-target, off-target, pharmacokinetics, and toxicological properties are gradually collected for these chemical entities. The corporate databases have frequent and multi-record transactions, such as uploading batches of assay data, bulk registering, or updating records. Project data is generally at the scale of 100 s to 1000s of compounds, and in-house collections typically range from 10 k–2 M compounds, with the associated data an order of magnitude higher. ChEMBL, for example, contains 2.3 M distinct compounds, ~20 M activity data points [1], and a wealth of additional information that is stored in 70+ database tables [2]. For this use case, it is critical to convert chemical structures into a standard representation, accurately capture stereochemical information, and explore possible tautomeric conflicts (matching tautomers) [3]. From a search viewpoint, the query structures can be complex, including features such as list atoms or position variation bonds. All hits are to be precisely retrieved, and the result set is joined to data fulfilling additional query conditions.

In contrast, idea generation during the abovementioned cycle should theoretically be free from accessibility bias and utilize the nearly infinite synthesizable chemical

space. Ciba-Geigy researchers estimated the chemical space of small drug-like molecules could easily scale up to 10^{63} compounds [4], simply through the rational combination of up to 30 C, H, N, S, O, or halogen atoms, while the number of “licensed” small molecule drugs is currently on the scale of 10^4 . Search strategies first screen internal compound collections, in-stock vendor libraries, public data sets, and patented compounds. Enumerated libraries of synthesizable compounds are on the order of billions; these can be valuable sources for hit expansion or lead optimization. Theoretically, enumerable sets, that are not stored as exact enumerated structures, represent the next sphere of compound space to navigate in order to find candidate molecules fitting the data-driven scientific hypothesis. These collections might represent the product matrix of in-house reactions and accessible building blocks or abstract structures of synthons that may be combined via functionalization points represented with Markush structures. These spaces are characterized by markedly different attributes, and their size range is astronomical (10^6 – 10^{26}) (Figure 15.1. and Table 15.1). Available chemical structure data sets are not necessarily detailed (stereochemistry, tautomer duplicates), are typically relatively static (only updated periodically), and often have limited dimensions, containing little more than the chemical structure and an identifier. Searching such large spaces is often done by similarity or matching abstracted pharmacophores, resulting in large hit sets (often >1 M), which can be challenging to process. Thus, ranking methods, which can reduce hit sets to more limited, relevant subsets, are highly valuable. These varying characteristics of discovery project space and idea generation space represent different, and to some extent orthogonal, challenges to navigating these landscapes.

How can the community of chemists reach out into the uncharted territory of viable compounds? This chapter will discuss the problems that the creators, maintainers, and users of large chemical databases face and the technical solutions

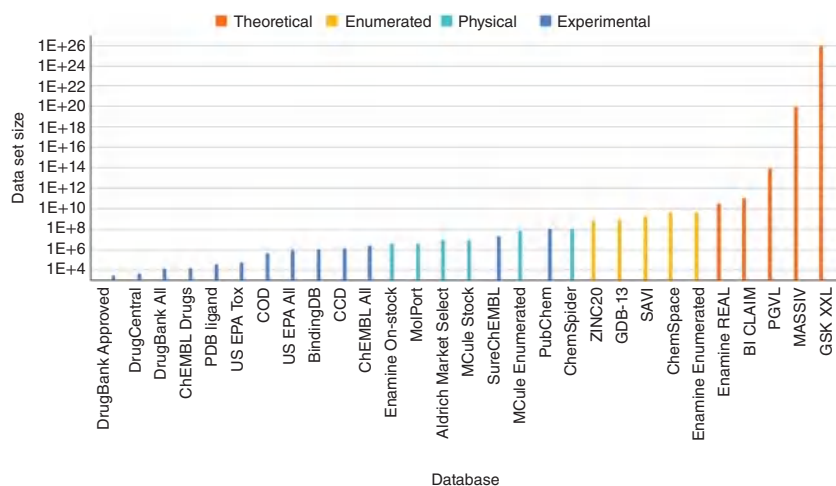


Figure 15.1 Data set sizes of example compound collections. Color coding represents the complexity and confidence of the data. Source: Ákos Tarcay.

Table 15.1 Selected examples of databases.

Database	Size	Data type	Reference
DrugBank Approved	2721	Approved drugs	[5]
DrugCentral	4714	Active ingredients chemical entities	[6]
DrugBank All	11 993	All drugs	[5]
ChEMBL Drugs	14 293	Drugs subset	[7]
PDB ligand	37 962	Experimental X-ray structure with large molecule	[8]
US EPA Tox	57 919	Structures for the 772 721 toxicity assay data points	[9]
COD	491 597	Crystal structures of organic, inorganic, metal–organic compounds, and minerals, excluding biopolymers	[10]
US EPA All	848 945	US Environmental Protection Agency	[11]
BindingDB	1.1 M	Public molecular recognition database	[12]
CCD	1.2 M	Small molecule X-ray data	[13]
ChEMBL All	2.3 M	Compound collection with 19 780 369 activity data points	[1]
Enamine On-stock	3.8 M	Stock screening compounds	[14]
MolPort	4 M	Stock screening compounds	[15]
Aldrich Market Select	8 M	Stock compounds	[16]
MCule Stock	9.3 M	Stock compounds	[17]
SureChEMBL	19 M	Patent extracted chemicals	[18]
MCule Enumerated	76 M	Purchasable and enumerated collections	[17]
PubChem	112 M	Large chemical and assay data with 296 M Bioactivities	[19]
ChemSpider	114 M	Theoretical and stock collections	[20]
ZINC20	230 M ready to dock 750 M purchasable	Enumerated	[21]
GDB-13	977 M	Theoretical enumeration	[22]
SAVI	175 B	Reaction-based enumeration	[23]
ChemSpace	4.5 B	Theoretical and stock collections	[24]
Enamine Enumerated	4.5 B	Enumerated database	[25]
Enamine REAL	29 B	Theoretical database	[25]
BI CLAIM	10 ¹¹	Theoretical database	[26]
PGVL	10 ¹⁴	Theoretical database	[27]
MASSIV	10 ²⁰	Theoretical database	[28]
GSK XXL	10 ²⁶	Theoretical database	[29]

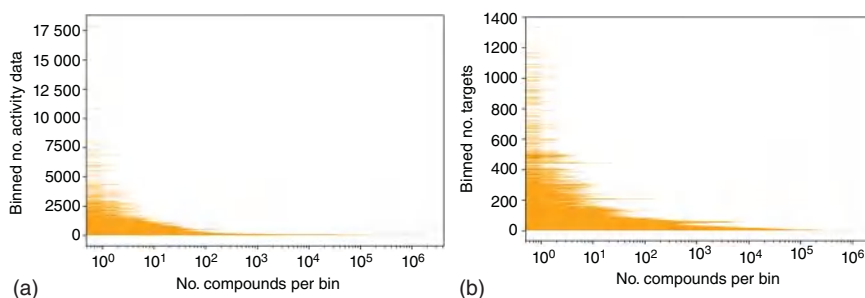


Figure 15.2 (a) Binned number of ChEMBL (version 31) activity values per compound, reflecting the volume of data per compound and (b) binned number of targets per compound, reflecting the dimensionality of the dataset. Source: Ákos Tarcsey.

currently in place to solve them. In order to highlight the data complexity of the public medicinal chemistry data, Figure 15.2a shows the (binned) number of assay measurements per compound in the ChEMBL 31 data set. A few compounds have a great many reported assay values, while a large number of compounds have only one or a few activity values. Ciprofloxacin (compound ID: ChEMBL8), a fluoroquinolone antibiotic used to treat different types of bacterial infections, is the most studied compound with more than 18 k assay data points in ChEMBL (version 31) [1]. Imatinib, a chemotherapy medication used to treat cancer (ChEMBL941), has the highest number of reported targets, counting more than 1300 target records.

Analyzing Figure 15.2a,b reveals that the assay data count (x-axis) shows a hyperbolic relationship between the number of compounds being involved in a particular biological assay. Relatively few compounds are available with a higher degree of experimental characterization. These assays came alive for more specific problems, and fewer compounds were tested against them. This is the section of hit-to-lead and lead optimization assays – an advanced phase compared to HTS – where only a few thousand compounds occur in a project. Moving along with the hyperbolic line, we reach the realm of those compounds that have only limited or no experimental data (Figures 15.1 and 15.2). This trend holds for the ultralarge datasets where the associated metadata or calculated properties are limited and no experimental data is available.

15.2 Workflows on the Chemical Space and Related Challenges

Before diving into the typical workflows associated with cheminformatics, we should define “spaces,” “libraries,” and “databases.” Spaces are theoretical ensembles of all molecules defined by a rule set (e.g. drug-like space). The largest virtual collections are combinatorially constructed. Typically, not all of the chemical structures are enumerated, and therefore the coverage of the space is larger than the precisely defined chemical structures. Libraries, which are subsets of spaces, contain enumerated collections of full structures and range from small numbers

to 10^9 compounds (like all products of a reaction scheme and a reagent library). Chemical databases are a way to store chemical information (e.g. libraries) mostly in relation to other data (in a relational database format, like ChEMBL [1]); these constitute the collections on which researchers perform hit-finding (virtual screening), exploration of SAR and project data analyses, substructure and similarity searches, overlap analyses, novelty checks, and clustering. Altogether, these activities constitute the most commonly occurring tasks that researchers perform when they are involved in one or more phases of drug discovery (hit identification/validation, hit-to-lead, lead-optimization, etc.).

15.2.1 Project Data Exploration

Efficient exploration of data is necessary in order to drive a project forward. Obtaining such data exhaustively may prove difficult, since historical data silos within organizations can make the retrieval and combination of relevant information difficult. Additionally, public data sources often observe their own data architecture norms, and significant effort may be required to access and reorganize the information within them. The burden of wrangling, querying, and dissemination of project data that is associated with modern Data Science has been a core part of the drug discovery process since before the term “Data Science” was popularized [30]. The development and adoption of the FAIR principles (Findable, Accessible, Interoperable, and Reusable) to ameliorate the situation is gaining traction and being employed both at private as well as public research organizations [31].

Once the data has been collected and organized, querying is performed by scientists throughout the different stages of the preclinical workflow. Medicinal chemists tend to use dedicated search and visualization interfaces (such as Instant JChem and Design Hub or Tableau extension from Chemaxon, Spotfire from Tibco, D360 from Certara, Live Design from Schrödinger, or Browser from Dotmatics), which can query a myriad of data sources. Often, a specialized chemical search engine, such as a cartridge is integrated into these tools to facilitate fast and chemically accurate searching. Many workflow steps require combined query capabilities so that both structure and other data fields (such as predicted physicochemical or ADMET properties) can be used as filters.

Chemical query features may be employed by the searcher to more quickly obtain a controlled diversity of results. These features allow for variability in the submitted structure, such as atom lists at specific positions, substitution patterns (e.g. R groups used in Markush structures), ring size, and more.

15.2.2 Corporate Data Warehouses

In the early days of cheminformatics, the incorporation of chemical data into databases was very limited. In 1979, MDL released Molecular ACCess System (MACCS) as one of the first solutions, which provided a way to store and search databases of 2D structures. In 1985, ISIS Host incorporated MACCS into the ISIS package, providing an integrated solution for storing and searching chemical

Table 15.2 Major database cartridges.

Name	Source	Database type	Reference
RDKit cartridge	RDKit, Open source, BSD 3	PostgreSQL	[33]
Bingo	EPAM, Open source, Apache License, Version 2.	Oracle/Postgres/SQL	[34]
Direct	BIOVIA/Accelrys/MDL	Oracle	[35]
MolCart	ICM	MySQL	[36]
MolSQL	Scilligence	MS SQL	[37]
JChem Oracle cartridge	Chemaxon	Oracle	[38]
JChem Choral cartridge	Chemaxon	Oracle	[39]
JChem Postgres cartridge	Chemaxon	PostgreSQL	[40]
SaChem	Institute of Organic Chemistry and Biochemistry of the CAS, Open source	PostgreSQL, Lucy	[41]

structure data in one database paired with nonstructural information in a separate Oracle database [32].

In 1998, chemical data storage took off with the release of Oracle 8i. This release introduced the concept of cartridges, which allow for the extension of the database functionality. The addition of chemical structure search and store functions to relational databases has grown to include many of the most common relational databases (e.g. PostgreSQL, MySQL, and MSSQL) and is still in widespread use today (Table 15.2) [42]. Chemaxon was founded in 1998, offering Java applications to store and search chemical data [43] in relational databases, and became the industry leader in chemical data storage for relational database management systems. The JChem product line includes JChem Base Java API to index chemical data in various databases (Oracle, MySQL, IBM DB2, Microsoft SQL Server, HSQLBD, Microsoft Access, Derby, and PostgreSQL) [44] and chemical cartridges for Oracle and PostgreSQL databases [39, 40].

In modern drug discovery informatics ecosystems, data warehouses involve intricately related schemas consisting of both structural and nonstructural data (e.g. compound registration, bioassay, inventory, analytical, etc.), often utilizing a single database to serve as the unified source of truth. In addition to the technical aspects of storing data, the adoption of the FAIR guiding principles has led to major advances in the usability of chemical data. Improving the quality of data storage has greatly reduced the amount of data silos and opened the door to applications such as advanced data visualization and machine learning.

15.2.3 Novel Compound Design

Ensuring the uniqueness of a structure vs. a certain comparison set of compounds is an important step in a number of medicinal chemistry workflows. Molecular

design is a prime example; designers need to make sure that their compounds are novel and do not infringe on previously patented work (Freedom to Operate), and are not redundant with previous (failed) in-house designs. Modern design systems can make such feedback instantly available to a chemist if they are integrated to appropriate data sources, assuming an appropriate search technology is also available.

The requirements of “appropriate search technology” can vary based on the size and nature of the data source. While conventional approaches can be used for many libraries, ultralarge libraries such as DNA Encoded Libraries [45] require specialized technology. This is not just due to the large size of the database but also the typical non-enumerated representation this information is stored in.

Even for more conventionally sized libraries, special technologies may need to be used when the searches are performed in bulk. For example, a library overlap analysis comparing, e.g. 10 k structures to a larger library of 1 M structures would otherwise face performance issues associated with the resulting combinatorial explosion.

While not strictly determining novelty, uniqueness checking is important in other cheminformatic workflows. Typically, an output of a library enumeration is filtered to only include unique compounds, and compound registry systems rely on uniqueness to properly control ID assignment. Chemical transformations such as tautomerization may also need to be taken into account in such cases.

15.2.4 Registration Systems

Compound Registration systems serve to group compounds based on a set of rules and typically generate unique identifiers accordingly. These systems consist of a database to store structural information and an evaluation mechanism that compares incoming compounds to those already in the database using structural search algorithms. Compound registration systems are often hierarchical, with distinct tiers to help group similar compounds (e.g. matching different salt/solvate forms under the same neutral parent). A robust system will feature extensive configurability, due to the range of institutional business rules regarding the standardization of incoming structures to improve homogeneity and data quality. More advanced systems allow for additional chemical intelligence, such as the matching of tautomers.

15.2.5 SAR-by Catalog

As part of hit validation and hit-to-lead workflows, scientists often attempt to establish a systematic structure-activity relationship (SAR) in order to determine the effect chemical modifications have on binding or activity. This approach is prevalent in fragment-based drug discovery, a bottom-up approach where small, low complexity compounds are grown into drug-like compounds [46]. A relatively small number of readily accessible similar compounds are identified and obtained, often by purchase, and the activity value is determined, leading to the term “SAR-by-catalog.”

A number of vendors exist offering access to such libraries, and chemists run searches using a number of conventional methods such as chemical fingerprint-based similarity, substructure, or pharmacophore feature searches to navigate the libraries prior to purchase.

More specialized tools are necessary when dealing with the largest of these libraries, such as the Enamine REAL database (29 billion compounds) [47] and WuXi AppTec's DEL Selection Package (50 billion compounds) [48]. In-house libraries also present a convenient point of access, and provide the advantage that additional institutional knowledge about the compound may be available. Representing such knowledge presents another challenge, as information about the properties, common transformations (such as Matched-Molecular Pairs [49]), or structural features may all be desirable. Recently, graph database representations have become a popular means to represent such classifications, to increase performance, and provide a more logical representation to human scientists [50].

15.3 Challenges in Chemical Data Search

Chemical matter is stored in various forms, depending on the workflow to be executed. In order to define the unique set, the input structures are normalized and converted into a standardized form. This process includes transformation to a drawing-independent chemical format, like the representation of the nitro groups or handling salts, solvates, tautomers, isotopes, and stereocenters. As an example, Chemaxon offers a widely used and comprehensive Standardizer module to convert structures to a canonical tautomer representation [51]. A thorough comparison of tautomer canonicalization methods found that the Chemaxon Tautomers node outperformed the other tested approaches [3]. A recent study investigated 40 databases embracing altogether more than 210 M structures for tautomer conflict using 119 detailed tautomeric transform rules [52]. The conflict rates typically fall in the range between 0.1% and 1%, with the extrema being 11.59% at the high end and 0.0% at the low end, and a median of 0.23%. The average conflict rate was about 0.77% [52].

The second most important challenge, after standardization, is the computational resources required to navigate in the chemical space. As noted in the introduction, the complexity of the search is further influenced by the variety of additional data stored, searched, and retrieved and the frequency of data updates (static or dynamic). Memory consumption and CPU utilization necessary to identify most similar structures scale nearly linearly with the library size. The dataset size increases logarithmically [4] with the number of atoms and enumerated library sizes have also expanded logarithmically during the past decade [29]. Storing molecular fingerprints in memory requires approximately 100–250 GB of RAM per 1B compounds, depending on the fingerprint representation used. These data storage requirements represent one of the key challenges and can set an upper limit on the size of enumerated sets that can be handled. Larger theoretical sets that can be potentially enumerated call for other paradigms for searching. Other approaches include using the reactants and reactions to search and enumerate possible products on the fly [53] or using the

Markush representation of the chemical space to search [54]. Horizontal resource scaling by sharding the database and using clusters of worker nodes offers a solution using cloud technologies [55].

15.4 Technologies

Search approaches to navigate in the chemical space can be categorized into three major types.

- (1) Filtering for exact matches between chemical structures (also referred to as duplicate searching) is a fundamental step in determining uniqueness when registering compounds.
- (2) Substructure search and superstructure search, depending on the subgraph isomorphism relation between the query and the target. Substructure search identifies subgraph matches of the query in the target structure, while superstructure search identifies the target molecules that are substructure matches of the query molecule.
- (3) Searching for structures that represent structural similarity without requiring exact substructure matches. While duplicate and substructure or superstructure searches are unambiguous, the expression of structural similarity between two chemical structures depends on the representation of the structure and the similarity metric that quantifies the degree of similarity [56]. Therefore, similarity is subjective, and the representation and similarity metric are to be defined to align with the goal of the comparison.

15.4.1 Representation Used for Similarity Search

In order to efficiently calculate chemical similarity, structures are represented with fixed-dimension vectors to reduce the complexity of calculating the subgraph similarity problem [57, 58]. The similarity is quantified using a distance metric between the query and target vectors. As a simple example, vectors that encode chemical property dimensions like molecular weight, aromatic rings, lipophilicity, number of rotatable bonds, etc. [56] can be compared by calculating their Euclidean distance. In practice, the structural similarity is calculated based on the chemical graph in such a way that the dimensions reflect the presence or absence of subgraph patterns. The number of such subgraph features yields an integer vector. The most frequently used representation is a binary (0,1) vector, which offers more efficient computational comparison; these vectors are called chemical fingerprints. Various types of fingerprint generation methods are available; the major types are discussed below.

15.4.1.1 Substructure-Preserving Fingerprints

This type of fingerprint fulfills the substructure relationship that all features (*on bits*) contained within a query must also be contained within the target. This is a necessary condition for the substructure relationship; therefore, it allows rapid elimination of molecules from consideration as a pre-filter when performing a substructure

search against a database. Only molecules that present all bits of the query are to be passed for the more resource-intensive subgraph isomorphism search [59].

Fingerprints based on a predefined library of structural patterns, like the MACCS (Molecular ACCess System or MDL keys) or the PubChem fingerprint, were originally constructed and optimized for substructure searching. In this technique, the predefined pattern matching sets an on-bit on the bitmap. The MACCS key was first defined as 166-bit and 960-bit versions [60]. Patterns encode, for example, the presence of elements or atoms from different groups, rings of different sizes, oxygen at different counts, and chemical moieties like amide, NH_2 , $\text{C}=\text{C}$, or CH_2 connected to heteroatoms. The PubChem fingerprint contains 115 hierarchical element counts, 148 ring systems, 64 bonded atom pairs, 89 examples of C, N, O, P, and Si within different environments, 44 detailed atom neighbors, and 421 simple or complex subgraph patterns; altogether 881 bits represent the molecules [61]. In the case of library-based fingerprints, important features are encoded to be able to represent and search a given chemical space. If molecules contain undefined structural patterns, this information will not be encoded.

Linear path-based hashed fingerprints (chemical hashed fingerprint, CFP) offer an alternative approach. This method exhaustively identifies all the linear paths in the molecule up to a predefined length, typically using 5–7 bond path lengths [62]. (An example up to length two is shown in Figure 15.3a. Additionally, rings are identified and represented with ring type and ring size attributes. The collected features are mapped to a predefined binary vector bit position using hash functions. The molecule in Figure 15.3 is encoded with 14 patterns up to path length 2, plus 2 additional bits for the ring. The hashing algorithm offers 2 additional parameters: the fingerprint length and the bits per feature. Decreasing the fingerprint length increases the chance that two given paths will map to the same position, a “bit collision.” Bit collision is characteristic of the hashed-type fingerprints; structural keys do not have this uncertainty. The bits per feature parameter allows the method to assign more bits to a given pattern, to balance the chance of bit collision. The increasing fingerprint length results in larger bit vectors, requires more memory, and may limit the in-memory search of extra-large libraries. The most common bit lengths are 512, 1024, and 2048. The path length, the fingerprint length, and the bits per feature parameters are the defining characteristics of the fingerprint. The number of *on* bits defines the “darkness” of the fingerprint, which can be optimized for the chemical space and the objective of the search.

15.4.1.2 Feature Fingerprints

While substructure fingerprints were designed and optimized for database searches, feature fingerprints were constructed to represent the characteristics of the molecule in terms of structure–activity properties. Accordingly, they do not preserve substructure relationships and are not suitable for pre-filtering during substructure search (compare Figure 15.3b,d panels). Their strength is to provide bit vectors for model building (machine learning) and yield better similarity values for the identification of active molecules during virtual screening. The extended connectivity fingerprint (ECFP, Figure 15.3) is the most frequently used example. ECFP [59] is a radial

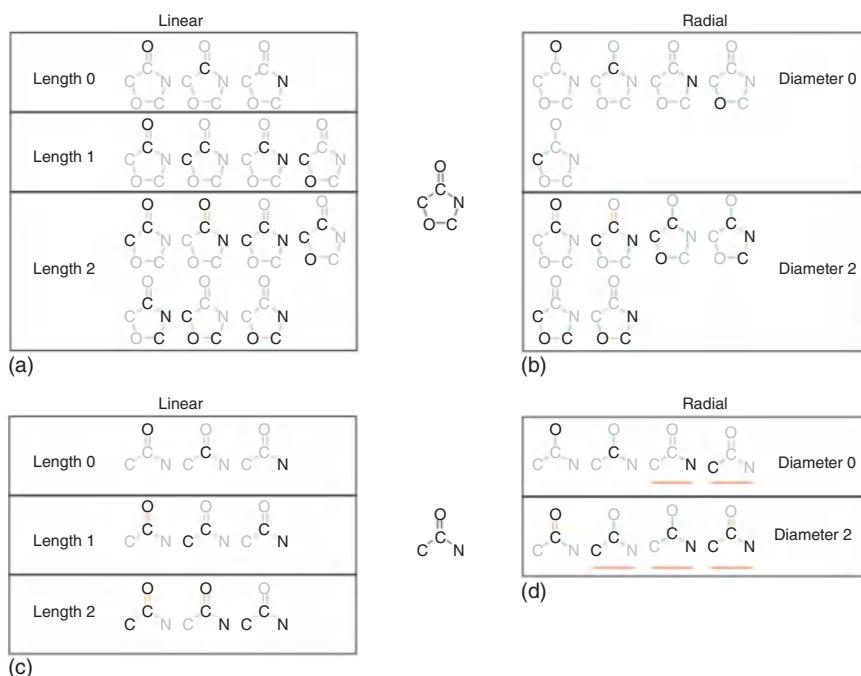


Figure 15.3 Comparison of linear and radial fingerprinting techniques. Orange bonds illustrate the current scope. Turquoise bonds represent atom environments that are taken into consideration in the case of ECFP radial fingerprint generation. Red underlining highlights ECFP patterns that violate the substructure relation between the upper and lower molecules. The figure was created with Marvin Pro [63]. Source: Ákos Tarcay.

fingerprint that encodes circular atom environments starting from each atom and expanding to a given diameter. The generated atom-centered patterns also take the environment of the selected region into account. The generated circular patterns are hashed using a modified Morgan algorithm [64] to 232-bit integer values (atom identifiers). The unique list of the generated atom identifiers is mapped onto a predefined bit string length. The modulo function is a straightforward method to map integers to a bit position. Similar to linearly hashed fingerprints, multiple features may be represented with the same bit code. Consequently, the absence of a bit is determinative, but the presence is only suggestive. ECFP generation is highly customizable: the maximal diameter, the final bit string size, and the consideration of atomic properties used for generating the atom identifiers influence the fingerprint projection and the corresponding similarity space. This flexibility is exploited in the functional class fingerprints (FCFPs), representing another level of abstraction, where the pharmacophore role of the atom is encoded. In the case of FCFP, each atom is identified by a six-bit code, where a given bit is *on* if the atom plays the associated role. The atom roles are: hydrogen-bond acceptor and donor; negatively and positively ionizable; aromatic; and halogen [59, 65]. As a result of this abstraction, a given bit position might represent a series of different radial substructural graph patterns that display functional similarity.

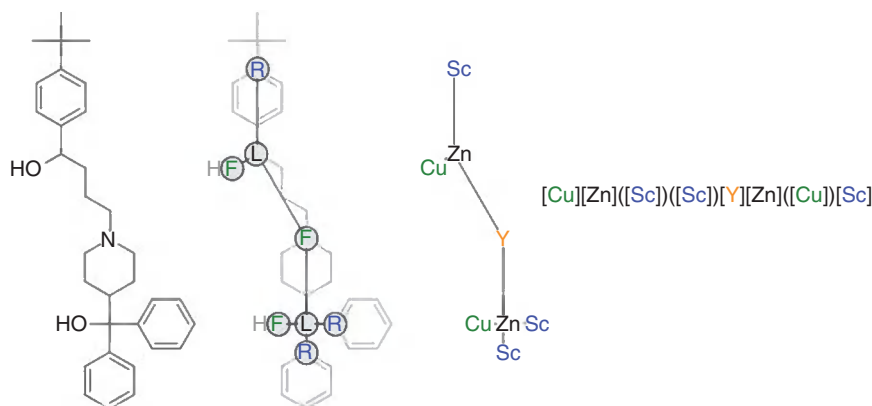


Figure 15.4 Graph reduction. In the first layer rings, linkers, and features are extracted. The extracted features can be further specified. Each species of the pharmacophore types is assigned to a rare heavy atom that is expressed as a SMILES string. The figure was created with Marvin Pro [63]. Source: Akos Tarcsay.

Another layer of abstraction was introduced by the reduced graph fingerprint [66–69]. Reduced graphs are generated from the structure in a way that the vertices are pharmacophoric points that are linked to preserve the original connectivity. The reduction is done in a hierarchical manner, where the top level corresponds to the highest level of abstraction, such as rings, linkers, and features. On subsequent levels, rings can be classified as aromatic or aliphatic, and features include donor or acceptor. On the next level, an aromatic ring can have donor or acceptor properties. Mapping the pharmacophore point at a given resolution into rare heavy atoms using a dictionary, the abstract graph can be expressed as a SMILES representation. This pseudomolecule can be represented with standard fingerprinting techniques to calculate similarity (Figure 15.4).

The radial nature of ECFP was combined with ring encoding and methods from natural language processing and data mining to construct the MiniHashFingerprint (MHFP). The MHFP differs from ECFP since attachment information on the boundaries of the radial fragments is not taken into account, only the subgraph pattern. As a result, MHFP extracts fewer unique features compared to ECFP. Processing 1.7 M compounds from ChEMBL24 resulted in 200 k and 500 k unique hash values using diameter 4 with MHFP or ECFP. Using diameter 6, the unique hash counts were 2 M and 3 M, demonstrating the large possible variety of radial regions in medicinal chemistry space [58]. According to the original authors, MHFP performed better in similarity-based virtual screening across 88 benchmark targets from ChEMBL 24. When folded to a 2048 dimension vector, it provided comparable enrichment to that of the 16 384 dimension ECFP4, indicating more efficient capturing of the structural information. MHFP enables approximate k-nearest neighbor (ANN) searches in sparse and high-dimensional binary chemical spaces without folding.

Structural keys linear and radial fingerprints extract information up to a predefined subgraph size. As a result, identical fingerprints may be generated by larger molecules that share similar structural patterns, even if the similar patterns are

present in different orders. For example, if two peptides are composed of the same residues, but in a different sequence, these fingerprinting techniques result in equivalence or high similarity. Peptide pairs AlaAspAlaLysAla and AlaAlaAspLysAla are encoded with the same chemically hashed linear fingerprint with path length 7 as well as with ECFP diameter 4, two of the most commonly used options. However, atom-pair fingerprints can distinguish between these two molecules. Atom pairs are defined as all atom pairs and the shortest topological distance that separates them to produce the (hashed) atom pair fingerprint. For a molecule with n heavy atoms, the total number of atom pairs will be $n*(n - 1)/2$. This technique does not encode the fine details of the chemical structure and provides a significantly different similarity space. The approach of atom pair fingerprinting was combined with the radial ECFP approach that resulted in the MAP4 fingerprint [70] to bridge the gap between small molecule and biopolymer structural similarity calculations. MAP4 encodes atom pairs and their bond distances similarly to the atom pair fingerprint; however, in MAP4 atom characteristics are replaced by the circular substructure around each atom of the pair, written in SMILES format using the MinHash as in the case of MHFP. MAP4 outperforms substructure fingerprints in small molecule benchmarking studies and at the same time outperforms other atom-pair fingerprints in a peptide benchmark designed to evaluate performance on large molecules [70].

In summary

- Substructure-preserving fingerprints, for example, linear hashed chemical fingerprints or structural keys, are used for rapid elimination of hit candidates in large databases before resource-intensive sub-graph isomorphism tests (Table 15.3).
- For structure-activity analysis, machine learning, and enrichment of active molecules based on the similarity principle that structurally related molecules have a higher chance of activity similarity, configurable radial fingerprints (ECFP, MHFP) or reduced graph-based representations are more suitable (Table 15.3).
- For large biopolymers, fingerprints encoding all possible topological distances are required, and atom pair or MAP4 fingerprints are more appropriate (Table 15.3).

15.4.2 Additional Representations

FTrees is a reduced graph approach that was designed to search enumerated databases using a topological approach [71]. Chemical functional groups are identified and marked as separate nodes, with the bonds between nodes categorized as linkers. Features such as donor counts, acceptor counts, and volume are calculated for each node, and these combined subtrees are merged to form the structure's Feature Tree. Similarity comparisons are then made that match subtree features. Since these comparisons preserve the local properties over the chemical structure, this is a useful method for users seeking to perform scaffold hopping. FTreesFS represents an evolution of FTrees, to extend it to much larger, non-enumerated libraries. Monomer subtrees are compared and dynamically built up into products on the fly, allowing for a more efficient expansion into a much larger chemical space, where the rules for fragment attachments are based on known synthetic possibilities to join them.

Table 15.3 Summary of fingerprinting methods and their applications.

Type	Example	Attributes	Workflow
Structural keys	MACCS, PubChem FPS	Predefined, substructure preserving, no bit collision	Database search, general
Chemical hashed fingerprint	CFP	Linear, customizable, substructure preserving, hashed, bit collision	Database search, general
Radial fingerprint	ECFP, Morgan, MFHP	Radial, customizable, hashed, bit collision	Virtual screening, machine learning, similarity search
Feature fingerprint	FCFP, Reduced graphs	Feature-based, abstract, pharmacophore-oriented, hashed, bit collision	Virtual screening, machine learning, similarity search
Atom pairs	Atom Pairs	Graph distances, substructure preserving, distinguishes biopolymers, hashed, bit collision	Virtual screening, machine learning, similarity search
Combined	MAP4	Radial and distance-based, applicable for small molecules and biopolymers, hashed, bit collision	Virtual screening, similarity search

3D shape representations are also used for a variety of use cases. Here, the comparison relies on knowledge of only a single known drug that is active against a target and not a confident crystal structure of the target, which may be difficult to obtain [72]. Assuming a non-promiscuous reference structure, selectivity can be assumed to the target from hits. 3D shape representations are also desirable because they provide the potential to overcome fingerprint-based approaches' limitations with respect to scaffold hopping [73]. The similarity expression is based on the volume overlap of the aligned structures. Representation can be extended with the biological activity profile of a molecule and chemical features such as electrostatic properties to generate a match score [74]. Overall, 3D shape search may be seen as a compromise between fingerprint searching and docking processes; its lower complexity makes it accessible to a broader audience, and its lower computational demand allows it to scale to larger datasets than docking [75, 76].

15.4.3 Commonly Used Similarity Definitions

Various functions are available to quantitatively express the comparison between two chemical structures. The function is either a distance or a similarity metric. Using the vectorial representation of the structures, simple functions like the Euclidean distance give a measure of their correspondence. Distance metrics (D) should obey four rules: (i) positive for nonidentical objects, (ii) distance from the object itself equals zero, (iii) symmetric, i.e. the distance from A to B equals the distance from B to A, and (iv) triangular inequality, the distance between A to B is smaller or equal to the distance from A to C plus the distance from C to B [77].

Similarity metrics (S) are scaled in the opposite direction, with higher values meaning higher similarity. These metrics should obey the following three rules: (i) for nonidentical objects, the similarity is lower than 1, (ii) identical objects have similarity of 1, and (iii) if the similarity function is symmetric, the similarity calculated between A and B equals similarity between B and A. Therefore, while the distance is a positive number (without maximum), similarity is always between 0 and 1, inclusive. Similarity can be trivially converted to dissimilarity: Dissimilarity = 1 – Similarity.

Structure-based similarity metrics rely on the binary fingerprint representation.

For the similarity expression, the following symbols are used:

- *a* is the number of on bits in molecule A
- *b* is number of on bits in molecule B
- *c* is the number of bits that are on in both molecules
- *d* is the number of common off bits
- *n* is the bit length (total number of bits) of the fingerprint: $n = a + b - c + d$.

The most commonly used similarity expressions are:

- Tanimoto coefficient: $S(A, B) = \frac{c}{a+b-c}$
- Soergel distance or Tanimoto dissimilarity: $D(A, B) = 1 - \frac{c}{a+b-c}$
- Euclidean distance: $D(A, B) = \sqrt{a + b - 2c}$
- Manhattan distance: $D(A, B) = a + b - 2c$
- Dice coefficient: $S(A, B) = \frac{2c}{a+b-c}$
- 15.Tversky: $S(A, B) = \frac{c}{\alpha(a-c) + \beta(b-c) + c}$
- 15.Cosine: $S(A, B) = \frac{c}{\sqrt{ab}}$

A statistical evaluation comparing similarity metrics based on their correspondence revealed that Cosine, Dice, Tanimoto, and Soergel similarities were identified as the best (equivalent) similarity metrics [78]. The Tversky metric has two parameters (α and β) that are weights for the compared A and B compound pair. In a symmetric case, when the weights are $\alpha = 1$ and $\beta = 1$, the Tversky and Tanimoto similarities are equal. In the asymmetric case ($\alpha = 1$ and $\beta = 0$), the Tversky similarity equals the common bits divided by the on bits of A, in other words (if A is the query and B is the target) this is the number of common bits divided by the number of query bits. This case calculates the substructure likeness of the query vs. the target, a similarity value of 1 means that all the query patterns (bits) are common with the target. While in the opposite asymmetric case ($\alpha = 0$ and $\beta = 1$), the metric calculates the common bits divided by the target bits, this is a superstructure likeness, since in this case a value of 1 means that all the target bits are shared with the query structure. Tanimoto similarity is used most frequently, as a standard element in numerous software applications to quickly find similar structures.

As noted earlier, the calculated similarity space for a chemical dataset is markedly influenced by the fingerprinting technique. An example case to highlight the differences is shown in Figure 15.5. Similarity spaces of 1 k randomly selected structures from ChEMBL (v31) hERG target with activity data (Target id ChEMBL240) are

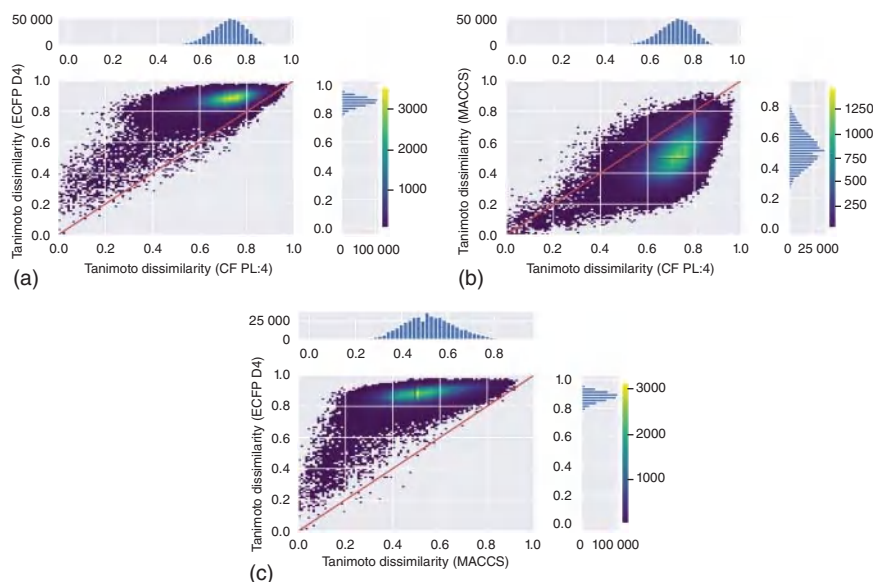


Figure 15.5 Comparison of Tanimoto dissimilarity spaces using different fingerprinting techniques. (a) ECFP diameter 4 (ECFP D4) vs. linear hashed chemical fingerprint with length 4 (CFP L:4), (b) MACCS key vs. CFP L:4, (c) ECFP D4 vs. MACCS key. 1000 randomly selected unique compounds from hERG target data (ChEMBL240, ChEMBL version 31, fingerprints were generated with JChem v22.13.). Source: Ákos Tarcsay.

shown using the Tanimoto similarity metric and ECFP with diameter 4, chemical hashed linear fingerprint with path length 4, and MACCS key. MACCS key-based similarity space identifies the structures to be more similar than CFPs, while ECFP4 identifies them to be the least similar.

15.5 Similarity Search Applications

Similarity search is available in all major cheminformatics software solutions and toolkits, including OpenBabel, CDK, OCChem TK, JChem, RDKit, Canvas, MOE, or PipelinePilot [79]. Similarity searching of 1 billion molecules requires 34 GB, 68 GB, 136 GB, and 252 GB of memory for 256, 512, 1024, and 2048-bit fingerprints [80], respectively. To be able to effectively process large data volumes, specialized applications were developed with demonstrated similarity-based screening results. Chemaxon introduced the MadFast Similarity Search, a Java application for ultra-fast parallel in-memory similarity search that offered multiple similarity metrics calculated on different fingerprints. Fast structure import processes 1 M per minute, and memory usage is about 250–350 MB per million. MadFast can calculate the 40 most similar structures in about 5 s for 1 billion molecules on an Amazon r3.8xlarge instance (244 GB of memory, 32 virtual CPUs, 2 × 320 GB of SSD-based instance storage, a 64-bit platform, and 10 Gigabit Ethernet) [29, 81]. This technology was superseded by the JChem Microservices DB module, where

720 M Enamine REAL molecules needed ~70 GB of RAM (AWS EC2 r6.4xlarge, 16 VCPU, 128 GB RAM). Using a Tanimoto similarity cutoff of 0.8, the top 100 hits were retrieved in 1 s in 90% of the cases using 52 query molecules [55]. OpenEye Scientific Software developed fingerprint search software and can store databases in memory in the cloud with fingerprints precomputed, in the Molecules as a Service (MaaS) module in Orion. A 2D similarity search of 800 M molecules from Enamine REAL takes 3 s, or less. To hold multiple large databases, the biggest Amazon Web Services (AWS) SSD memory instance is needed (768 GB memory and 96 logical processors on 48 physical cores) [82]. The Chemfp project reported 1000 nearest-neighbor searches of the 1.8 M 2048-bit Morgan fingerprints of ChEMBL 24 averaging 27 ms/query. The same search of 970 M PubChem fingerprints averages 220 ms per query [83]. Schrödinger offers a proprietary, very fast similarity comparison tool, GPUSimilarity, in which commercially available compound libraries containing approximately 1.6 billion compounds are hosted on a GPU-powered server [29].

15.6 Substructure Search

15.6.1 Atom by Atom Search

Traditionally, a substructure search is executed by performing a colored graph-matching algorithm. Considering the multiple algorithms available for matching colored graphs, the Ullmann method [84] has been the most widely used in the case of chemical substructure matching for the last 20 years. In 2004, the VF2 algorithm [85] was published, offering the advantage of higher performance in the case of typical chemical graphs.

15.6.2 Ullmann Algorithm

The Ullmann algorithm constructs a possible mapping matrix (M) to identify possible query target atom mapping pairs. The algorithm iterates on this matrix and clears impossible query target atom pair matches based on local properties (atom number, number of connections, etc.) and neighbor connectivity until the matrix is changed. After the iteration is finished (the matrix is not changing anymore) a backtracking algorithm is used for the identification of a specific query and target atom mapping. A possible query target atom match is selected, and the matrix is copied and cleared again based on local properties and neighbor connectivities. If the clearing process leads to a zero row – no possible matching for a query atom – we step back to the previous query atom and search for another possible target match. This selection and matrix clearing are iterated recursively until all atoms are matched or there is no possible match for the specific query target pair.

This algorithm works well for chemical graphs – but not with general graphs, where it may explode exponentially – in the optimal case, its computational cost is proportional to the square of the query atom count multiplied by the target atom count. The memory allocation is on the order of the mapping matrix (M) multiplied by the number of query atoms: query atom count squared times the target

atom count. With large structures like peptides, this computational resource need should be considered. Detailed examples of the matrix operations are available in ref. [86].

15.6.3 VF2 Algorithm

The VF2 algorithm uses a different approach without building a full mapping matrix; therefore, its initialization cost is much smaller compared to the Ullmann algorithm. It starts with a query atom (q_1) and matches (based on local properties) to a target atom (t_1). Takes another connected atom (q_2) and matches it to another t_2 atom, then checks if all neighbors of q_2 that were previously mapped to a target atom (t_x t_y t_z) are also connecting to t_2 . For both VF2 and Ullmann algorithms, the ordering of query atoms plays a crucial role. Processing the most selective query atoms at the beginning of the recursion is more effective. Detailed VF2 graph matching workflow example is available in ref. [86].

The VF2 algorithm was developed further in 2015 to VF2 Plus [87], and VF2 Plus was developed even further in 2018 to VF2++ [88]. Using the importance of the node order and using more efficient cutting rules, the VF2++ algorithm is able to match graphs of thousands of nodes while scaling practically linearly with time, including preprocessing. It is shown that VF2++ consistently outperforms VF2 Plus on biological graphs. In the case of random sparse graphs, VF2++ is faster than VF2 Plus, and it also has a practically linear behavior both in the case of induced subgraphs and graph isomorphism problems.

In practice, these graph-matching algorithms are not used alone. The target molecule set is prefiltered for possible matches, and these possible hits are provided for the specific graph matching algorithm. This filtering is typically executed using substructure-preserving fingerprints. For both the query and the target structures, the substructure-preserving fingerprints are calculated, and if all bits of the query fingerprint are available in the target fingerprint, then it defines a possible match.

In a typical scenario, the substructure search is executed in two phases. First, the so-called screening phase is executed, where the possible matches are identified, which is followed by the atom-by-atom matching (graph matching) phase. The screening phase is multiple times faster than the atom-by-atom search phase, as it needs only fingerprint comparisons. The target fingerprints can be calculated upfront and can be stored in memory for optimal performance. If the selectivity of the fingerprint is not high enough, additional properties can be taken into consideration to reduce the size of possible hits. However, even in the case of good selectivity, one might not want to execute all necessary atom-by-atom matchings since the expected result count – e.g. which can be digested by a human – is much smaller compared to the real matches. As a simple example, we can imagine searching for a very common structure (like a benzene ring) in a target set containing 1 M structures. In a typical molecule set at least 50% of the structures contain a moiety with a benzene ring. Would a chemist need to have all the ~500 k hits? Not always, in the case of a user-facing application, only the first 10–50 hits are presented, but these hits should be the most relevant ones. In this case, the ordering of the

possible hits by similarity to the query structure can help to avoid superfluous (and long-running) atom-by-atom searches.

In Chemaxon's solution, custom implementations are in place to improve fingerprint selectivity, which is needed indeed to be able to support a wide range of query features [89]. In the latest search solutions, the ordering of the hits is based on the query target similarity value, and both Ulmann and internally improved VF2++ graph matching algorithms are used depending on the complexity of the query structure to achieve the fastest atom-by-atom matching speed.

It is worth mentioning that substructure search can be executed without using a graph-matching algorithm [90]. Using graph databases, all possible subgraphs of the molecule are represented as distinct, unique nodes in the graph database. The edges of the graph database represent single-step molecule graph edits from one node to the other. One molecule graph edit is an addition or removal of a chemical atom from the molecule. The real molecules and the nodes representing subgraphs should be differentiated (by marking them differently). If the graph database nodes are representing the molecules and subgraphs in a canonical form, then the substructure search is simply a lookup in the graph database with the canonical representation of the query structure. After locating the specific query structure in the graph database, the algorithm traverses the graph to locate the closest real chemical structure (which is also a node in the graph) where the graph edit does not contain atom removal.

This approach has the advantage that the substructure search is finally simplified to a lookup in the graph database, resulting in a sublinear search speed with the size of the database. On the other hand, it should be mentioned that the storage requirement of this approach can be quite high since all possible subgraphs of the molecule should be persisted. Theoretically, the storage requirement saturates with the increasing number of structures, as the more structures persist in the database, the higher the chance that the new molecule's subgraphs are already available in the database.

15.7 Application Example

Cartridges are extensions to relational databases to store and search chemical data. In order to find compounds fulfilling certain chemical queries like substructure matches and their metadata, a complex joint query with several criteria has to be devised and run by joining several tables. When multiple conditions are provided, the order of filtering steps can make a significant difference in the execution time. In relational databases, the complex query steps operating on multiple fields are executed according to a query plan, which is created by the built-in query optimizer. The query optimizer predicts the cost of executing each filtering step and devises the optimal order of execution. As part of the cost estimation, the number of rows that will be returned by the given filtering step is estimated. In the case of queries that also involve chemical search, the query optimizer needs to estimate the cost of carrying out the chemical search as well. This cost estimation can be

provided by the cartridge extension, which can lead to more efficient execution of complex chemical queries.

Several commercially available and open-source cartridges have been developed to enable the retrieval of chemical structures from databases. The most commonly used back-ends are either Oracle or PostgreSQL, and there are open-source and commercially available solutions available. Major chemical cartridges are listed in Table 15.2. Among the open-source cartridges, RDKit's PostgreSQL and EPAM's Bingo cartridges are frequently used. As an example, the search performance of these cartridges, versions 0.76 and 1.9.1, respectively, is compared with the performance of Chemaxon's JChem PostgreSQL cartridge (JPC) version 21.13. The ChEMBL 29 database was used since it is representative in terms of chemical space and size for drug discovery collections. 52 compounds [91], covering a wide range of chemical properties, moieties, and query features, were selected as query compounds. In addition to substructure search, filtering on phys-chem properties was also applied. In these queries, the following five nonchemical search criteria were used: (i) Molecular weight < 300 Da, (ii) $\log P < 5$, (iii) $40 < \text{Topological Polar Surface Area} < 140$, (iv) Number of heavy atoms < 20, (v) passes Lipinski's rule of five [91]. These criteria represent commonly used properties run routinely and at large scale during the search for compounds in drug discovery. The chemical searches were performed on AWS t2.xlarge EC2 instances with 4 vCPUs and 16 GB of RAM. 6 searches were performed with each query. The result of the first, "warm-up," query was discarded. The average elapsed time, representing the time needed for the query to finish, has been calculated from the 5 additional runs.

The elapsed time during the combined chemical substructure and nonchemical search is shown for the 52 query compounds in Figure 15.6. Considerable difference in the search performance of the different cartridges even on this small target molecule set (~ 2 M) was observed. Based on these measurements, the JPC was

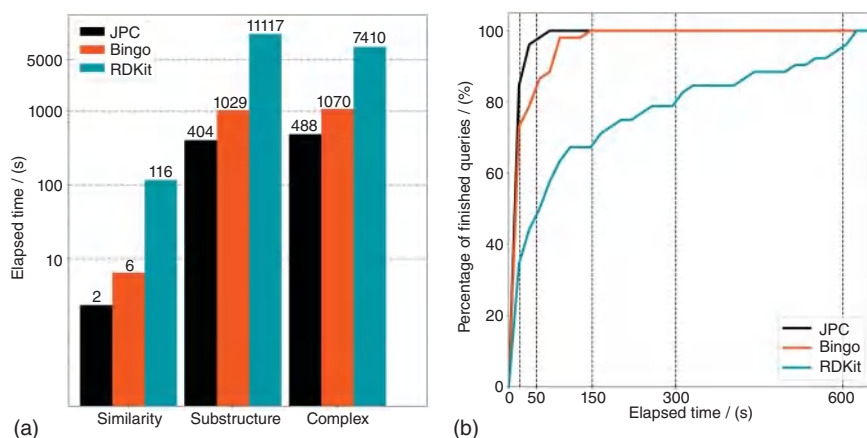


Figure 15.6 (a) Sum elapsed time of 52 query executions for similarity, substructure, and complex searches using the JChem PostgreSQL cartridge (JPC), Bingo, and RDKit cartridges, respectively. (b) Percentage of queries finished within the elapsed time on the horizontal axis for complex queries. Source: Máté Erdős.

found to be the fastest for similarity, substructure, and complex queries with 2, 404, and 488 s total elapsed time for the 52 query molecules. JPC executed the queries approximately 2.5 times faster than Bingo and in the case of the RDKit cartridge the searches took 58, 27, and 15 times longer. We observed that among the investigated cartridges JChem PostgreSQL utilized all available CPU cores, while the other two cartridges only used one core.

15.8 Summary and Outlook

The recent expansion of the multidimensional data size and exponential growth of collections impose new challenges for searching and accessing chemical data. The performance requirements of current drug discovery chemists can be served only by storing the dataset in memory, which can result in reaching capacity limits.

Navigating in complex, multidimensional data, such as experimental data for drug discovery project compounds, requires executing queries with combined conditions. Query planning in the database engine ranks the searching in the different dimensions using cost estimation to be able to slice the entire data set. The condition with the highest cost is executed in index mode in memory, and the hit list is joined with rapid sequential scans on the pre-filtered hit lists. To yield high performance, query execution planning has a cardinal role. Suboptimal query plans can result in 10-fold increases in execution time, even on small sets [92]. Chemical cartridges are creating chemical domain indices and combining fast fingerprint-based prescreening with atom-by-atom search algorithms to select the hits. There are novel attempts to navigate in datasets with high complexity using different approaches relying on different database engines such as graph databases or Elasticsearch [93–97]. These technological developments are broadening the horizon of current data traversing possibilities.

Vertical scaling of computer hardware enables assigning larger amounts of memory and CPU within a single server to enable handling large amounts of static data with low complexity, such as enumerated libraries. Examples of fingerprint storage in memory and rapid similarity comparisons on large sets were discussed in the Similarity Search Applications subchapter. A different architectural approach is database sharding, when horizontal partitioning is applied and partitions of the data are stored in multiple independent executors to yield smaller, faster, and more easily manageable parts. During sharding, the job execution is to be distributed between the shards with an equivalent load. Therefore the dataset is to be randomly sampled and evenly distributed to balance the query execution load and avoid overloading a single query node. Recent cloud technologies provide the architectural framework to manage larger clusters of executors using container technology. Amazon Elastic Container Service [98], as an example, enables the management of a large number of containers, even on a large scale (scaling out). A proof-of-concept application of using Fargate [99] technology to spin up small containers resulted in retrieval of the top 200 similarity-ordered substructure hits in less than 4 s from the ENAMINE REAL dataset containing more than 1 B

structures [55]. This study exploited the JChem Microservices technology to fetch the hit molecules. Elastic cloud infrastructure provides opportunities to overcome the limitations of a single-machine in-memory storage and scale-out to handle multibillion compounds.

Navigating in the extra-large, theoretical space of non-enumerated compounds requires a different approach. These methods open up new horizons, where, for example, reactants are prescreened and products are enumerated on the fly using possible reactions. The LEAP technique from Pfizer is an example of navigating chemical spaces defined by in-house reactants and known reactions [53]. FTrees from BioSolveIT uses the reduced graph approach on fragments and rules to combine them to enable fast searching of ultra-large potential spaces [26, 100].

Acknowledgments

We are grateful to Dóra Barna, Tim Parrott, Erneszt Kovács, Róbert Wágner, and András Strácz for their valuable comments and suggestions during the preparation of the manuscript.

References

- 1 ChEMBL database www.ebi.ac.uk/chembl [accessed 12 September 2022]
- 2 ChEMBL database schema https://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/latest/chembl_31_schema.png [accessed 12 September 2022]
- 3 Dolciemi, D., Villascaras-Fernandez, E., Kannas, C. et al. (2022). canSAR chemistry registration and standardization pipeline. *J. Cheminform.* 14 (1): 28. <https://doi.org/10.1186/s13321-022-00606-7>.
- 4 Walters, W.P. (2019). Virtual chemical libraries. *J. Med. Chem.* 62: 1116–1124. <https://doi.org/10.1021/acs.jmedchem.8b01048>.
- 5 Drugbank database <https://go.drugbank.com/drugs> [accessed 12 September 2022]
- 6 DrugCentral database <https://drugcentral.org> [accessed 12 September 2022]
- 7 ChEMBL database Drugs www.ebi.ac.uk/chembl/g/#browse/drugs [accessed 12 September 2022]
- 8 PDB ligand database <http://ligand-expo.rcsb.org/ld-download.html> [accessed 12 September 2022]
- 9 US EPA Tox https://comptox.epa.gov/dashboard/chemical-lists/TOXVAL_V5 [accessed 12 September 2022]
- 10 Crystallography Open Database (COD) <http://cod.crystallography.net/cod> [accessed 12 September 2022]
- 11 US EPA All <https://comptox.epa.gov/dashboard> [accessed 12 September 2022]
- 12 BindingDB <https://www.bindingdb.org/rwd/bind/index.jsp> [accessed 12 September 2022]
- 13 Cambridge Crystallographic Data Centre (CCDC) <https://www.ccdc.cam.ac.uk> [accessed 12 September 2022]

- 14 Enamine database <https://enaminestore.com/search> [accessed 12 September 2022]
- 15 Molport database <https://www.molport.com> [accessed 12 September 2022]
- 16 Aldrich Market Select <https://www.aldrichmarketselect.com> [accessed 12 September 2022]
- 17 Mcule database <https://mcule.com/database> [accessed 12 September 2022]
- 18 SureChEMBL <https://www.surechembl.org/search> [accessed 12 September 2022]
- 19 PubChem database <https://pubchem.ncbi.nlm.nih.gov> [accessed 12 September 2022]
- 20 ChemSpider database <http://www.chemspider.com> [accessed 12 September 2022]
- 21 Zinc database <https://zinc20.docking.org> [accessed 12 September 2022]
- 22 GDB database <https://gdb.unibe.ch/downloads> [accessed 12 September 2022]
- 23 SAVI database https://cactus.nci.nih.gov/download/savi_download [accessed 12 September 2022]
- 24 Chem-Space database <https://chem-space.com/search>
- 25 Enamine REAL compounds database <https://enamine.net/compound-collections/real-compounds> [accessed 12 September 2022]
- 26 Lessel, U., Wellenzohn, B., Lilienthal, M., and Claussen, H. (2009). Searching fragment spaces with feature trees. *J. Chem. Inf. Model.* 49 (2): 270–279. <https://doi.org/10.1021/ci800272a>.
- 27 Qiyue, H., Zhengwei, P., Scott, C.S. et al. (2012). Pfizer global virtual library (PGVL): a chemistry design tool powered by experimentally validated parallel synthesis information. *ACS Comb Sci.* 14 (11): 579–589. <https://doi.org/10.1021/co300096q>.
- 28 MASSIVE database https://www.merckgroup.com/research/science-space/presentations/2018-05-24_The_power_of_in-silico.pdf [accessed 12 September 2022]
- 29 Warr, W.A., Nicklaus, M.C., Nicolaou, C.A., and Rarey, M. (2022). Exploration of ultralarge compound collections for drug discovery. *J. Chem. Inf. Model.* 62 (9): 2021–2034. <https://doi.org/10.1021/acs.jcim.2c00224>.
- 30 Ferrero, E., Brachat, S., Jenkins, J.L. et al. (2020). Ten simple rules to power drug discovery with data science. *PLoS Comput. Biol.* 16 (8): e1008126. <https://doi.org/10.1371/journal.pcbi.1008126>.
- 31 Harrow, I., Balakrishn, R., McGinty, H.K. et al. (2022). Maximizing data value for biopharma through FAIR and quality implementation: FAIR plus Q. *Drug Discov. Today* 27 (5): 1441–1447. <https://doi.org/10.1016/j.drudis.2022.01.006>.
- 32 Wild D, *Introducing Cheminformatics, 2012–2013 LULU, (ASIN: B00G5TS7B4)* 2012
- 33 RDKit Cartridge <https://www.rdkit.org/docs/Cartridge.html> [accessed 12 September 2022]
- 34 Bingo Cartridge <https://lifescience.opensource.epam.com/bingo/index.html> [accessed 12 September 2022]
- 35 BioVia Direct Cartridge <https://www.3ds.com/products-services/biovia/products/scientific-informatics/biovia-direct> [accessed 12 September 2022]
- 36 MolCart <https://www.molsoft.com/molcart.html> [accessed 12 September 2022]

- 37 MolSQL <https://www.scilligence.com/web/molsql> [accessed 12 September 2022]
- 38 Jchem Oracle Cartridge <https://docs.chemaxon.com/display/docs/jchem-oracle-cartridge.md> [accessed 12 September 2022]
- 39 <https://docs.chemaxon.com/display/docs/jchem-choral.md> [accessed 12 September 2022]
- 40 <https://docs.chemaxon.com/display/docs/jchem-postgresql-cartridge.md> [accessed 12 September 2022]
- 41 Sachem cartridge <http://bioinfo.uochb.cas.cz/sachem> [accessed 12 September 2022]
- 42 O'Donnell, T.J. (2008). Design and Use of Relational Databases in Chemistry. CRC Press <https://doi.org/10.1201/9781420064438>.
- 43 Csizmadia, F. (2000). JChem: Java applets and modules supporting chemical database handling from web browsers. *J. Chem. Inf. Comput. Sci.* 40 (2): 323–324. <https://doi.org/10.1021/ci9902696>.
- 44 <https://docs.chemaxon.com/display/docs/administration-guide-jchem-manager.md> [accessed 12 September 2022]
- 45 Gironda-Martínez, A., Donckele, E.J., Samain, F., and Neri, D. (2021). DNA-encoded chemical libraries: a comprehensive review with successful stories and future challenges. *ACS Pharmacol. Transl. Sci.* 4: 1265–1279.
- 46 Neto, L.R.S., Moreira-Filho, J.T., Neves, B.J. et al. (2020). *Front Chem.* 8: 93. <https://doi.org/10.3389/fchem.2020.00093>.
- 47 Grygorenko, O.O., Radchenko, D.S., Dziuba, I. et al. (2020). Generating multi-billion chemical space of readily accessible screening compounds. *iScience* 23 (11): 101681. <https://doi.org/10.1016/j.isci.2020.101681>.
- 48 Wart, W. (2021). Report on an NIH workshop on ultralarge chemistry databases. *ChemRxiv* <https://doi.org/10.26434/chemrxiv.14554803.v1>.
- 49 Dossetter, A.G., Griffen, E.J., and Leach, A.G. (2013). Matched molecular pair analysis in drug discovery. *Drug Discov. Today* 18 (15–16): 724–731. <https://doi.org/10.1016/j.drudis.2013.03.003>.
- 50 Hall, R.J., Murray, C.W., and Verdonk, M.L. (2017). The fragment network: a chemistry recommendation engine built using a graph database. *J. Med. Chem.* 60 (14): 6440–6450. <https://doi.org/10.1021/acs.jmedchem.7b00809>.
- 51 <https://chemaxon.com/products/chemical-structure-representation-toolkit> [accessed 20 September 2022]
- 52 Dhaked, K.D. and Nicklaus, M. (2021). Tautomeric conflicts in forty small-molecule databases. *ChemRxiv* <https://doi.org/10.26434/chemrxiv.14779254.v1>.
- 53 Hu, Q., Peng, Z., Kostrowicki, J., and Kuki, A. (2011). LEAP into the Pfizer global virtual library (PGVL) space: creation of readily synthesizable design ideas automatically. *Methods Mol. Biol.* 685: 253–276. https://doi.org/10.1007/978-1-60761-931-4_13.
- 54 <https://docs.chemaxon.com/display/docs/searching-in-markush-targets-tables.md> [accessed 20 September 2022]
- 55 <https://chemaxon.com/presentation/substructure-search-jcm-enamine-apac> [accessed 20 September 2022]

- 56 Maggiora, G., Vogt, M., Stumpfe, D., and Bajorath, J. (2014). Molecular similarity in medicinal chemistry. *J. Med. Chem.* 57 (8): 3186–31204. <https://doi.org/10.1021/jm401411z>.
- 57 Grohe, M., Rattan, G., and Woeginger, G.J. (2018). Graph similarity and approximate isomorphism. In: *Graph Similarity and Approximate Isomorphism*, 1–16. Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.
- 58 Probst, D. and Reymond, J.L. (2018). A probabilistic molecular fingerprint for big data settings. *J. Cheminform.* 10 (1): 66. <https://doi.org/10.1186/s13321-018-0321-8>.
- 59 Rogers, D. and Hahn, M. (2010). Extended-connectivity fingerprints. *J. Chem. Inf. Model.* 50 (5): 742–754. <https://doi.org/10.1021/ci100050t>.
- 60 Durant, J.L., Leland, B.A., Henry, D.R., and Nourse, J.G. (2002). Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* 42 (6): 1273–1280. <https://doi.org/10.1021/ci010132r>.
- 61 https://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.pdf [accessed 20 September 2022]
- 62 <https://docs.chemaxon.com/display/docs/chemical-hashed-fingerprint.md> [accessed 20 September 2022]
- 63 Marvin Pro Version 22.11 <https://chemaxon.com/products/marvin-pro> [accessed 20 September 2022]
- 64 Morgan, H.L. (1965). The generation of a unique machine description for chemical structures—a technique developed at chemical abstracts service. *J. Chem. Doc.* 5 (2): 107–113. <https://doi.org/10.1021/c160017a018>.
- 65 <https://docs.chemaxon.com/display/docs/extended-connectivity-fingerprint-ecfp.md> [accessed 20 September 2022]
- 66 Birchall, K. and Gillet, V.J. (2010). Reduced graphs and their applications in chemoinformatics. In: *Chemoinformatics and Computational Chemical Biology*, 197–212. https://doi.org/10.1007/978-1-60761-839-3_8.
- 67 Gillet, V.J., Willett, P., and Bradshaw, J. (2003). Similarity searching using reduced graphs. *J. Chem. Inf. Comput. Sci.* 43 (2): 338–345. <https://doi.org/10.1021/ci025592e>.
- 68 Barker, E.J., Gardiner, E.J., Gillet, V.J. et al. (2003). Further development of reduced graphs for identifying bioactive compounds. *J. Chem. Inf. Comput. Sci.* 43 (2): 346–356. <https://doi.org/10.1021/ci0255937>.
- 69 Pogány, P., Arad, N., Genway, S., and Pickett, S.D. (2019). De novo molecule design by translating from reduced graphs to SMILES. *J. Chem. Inf. Model.* 59 (3): 1136–1146. <https://doi.org/10.1021/acs.jcim.8b00626>.
- 70 Capecchi, A., Probst, D., and Reymond, J.L. (2020). One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *J. Cheminform.* 12 (1): 43. <https://doi.org/10.1186/s13321-020-00445-4>.
- 71 Rarey, M. and Dixon, J.S. (1998). Feature trees: a new molecular similarity measure based on tree matching. *J. Comput. Aided Mol. Des.* 12 (5): 471–490. <https://doi.org/10.1023/a:1008068904628>.
- 72 Lo, Y.C., Senese, S., Damoiseaux, R., and Torres, J.Z. (2016). 3D chemical similarity networks for structure-based target prediction and scaffold hopping. *ACS Chem. Biol.* 11 (8): 2244–2253. <https://doi.org/10.1021/acscchembio.6b00253>.

- 73 Kalászi, A., Szisz, D., Imre, G., and Polgár, T. (2014). Screen3D: a novel fully flexible high-throughput shape-similarity search method. *J. Chem. Inf. Model.* 54 (4): 1036–1049. <https://doi.org/10.1021/ci400620f>.
- 74 Riniker, S., Wang, Y., Jenkins, J.L., and Landrum, G.A. (2014). Using information from historical high-throughput screens to predict active compounds. *J. Chem. Inf. Model.* 54 (7): 1880–1891. <https://doi.org/10.1021/ci500190p>.
- 75 <https://www.schrodinger.com/products/shape-screening> [accessed 20 September 2022]
- 76 <https://docs.eyesopen.com/applications/rocs/pub.html> [accessed 20 September 2022]
- 77 Willett, P., Barnard, J.M., and Downs, G.M. (1998). Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* 38 (6): 983–996.
- 78 Bajusz, D., Rácz, A., and Héberger, K. (2015). Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminform.* 7: 20. <https://doi.org/10.1186/s13321-015-0069-3>.
- 79 Cereto-Massagué, A., Ojeda, M.J., Valls, C. et al. (2015). Molecular fingerprint similarity search in virtual screening. *Methods* 71: 58–63. <https://doi.org/10.1016/j.ymeth.2014.08.005>.
- 80 https://www.nextmovesoftware.com/talks/Sayle_RecentAdvancesInChemicalSearch_ICCS_202206.pdf [accessed 20 September 2022]
- 81 https://wp.chemaxon.com/app/uploads/2018/05/MFSS_JPC_Cartridge_2018_ICCS-Poster.pdf [accessed 20 September 2022]
- 82 <https://www.eyesopen.com/news/openeye-orion-2020.2-update> [accessed 20 September 2022]
- 83 Dalke, A. (2019). The chemfp project. *J. Cheminform.* 11 (1): 76. <https://doi.org/10.1186/s13321-019-0398-8>.
- 84 Ullmann, J.R. (1976). An algorithm for subgraph isomorphism. *J. ACM (JACM)* 23: 31–42.
- 85 Cordella, L.P., Foggia, P., Sansone, C., and Vento, M. (2004). A (sub)graph isomorphism algorithm for matching large graphs. *IEEE Trans. Pattern Anal. Mach. Intell.* 26 (10): 1367–1372.
- 86 <https://docs.chemaxon.com/display/docs/graphmatching.md> [accessed 20 September 2022]
- 87 Carletti, V. and Foggia, P. (2015). VentoM, VF2 plus: an improved version of VF2 for biological graphs. In: *Graph-Based Representations in Pattern Recognition, at Beijing*. https://doi.org/10.1007/978-3-319-18224-7_17.
- 88 Jüttner, A. and Madarasi, P. (2018). VF2++—an improved subgraph isomorphism algorithm. *Discrete Appl. Math.* 242: 69–81. <https://doi.org/10.1016/j.dam.2018.02.018>.
- 89 <https://docs.chemaxon.com/display/docs/query-features-jcb.md> [accessed 20 September 2022]
- 90 *19th EuroQSAR Meeting in Vienna, Austria, August 2012*, <https://www.nextmovesoftware.com/products/SmallWorldPoster.pdf> [accessed 20 September 2022]

- 91 https://docs.chemaxon.com/display/docs/database_queries_suppinfo.md [accessed 12 October 2022]
- 92 <https://depth-first.com/articles/2021/08/11/the-rdkit-postgres-ordered-substructure-search-problem> [accessed 20 September 2022]
- 93 <https://www.slideshare.net/NextMoveSoftware/chemical-similarity-using-multiterabyte-graph-databases-68-billion-nodes-and-counting> [accessed 20 September 2022]
- 94 <https://github.com/rdkit/neo4j-rdkit> [accessed 20 September 2022]
- 95 https://chemaxon.com/presentation/neo4j_presentation [accessed 20 September 2022]
- 96 <https://chemaxon.com/presentation/cheminfo-stories-2021-virtual-ugm-jchem-elasticsearch-plugin> [accessed 20 September 2022]
- 97 Matter, H., Buning, C., Stefanescu, D.D. et al. (2020). Using graph databases to investigate trends in structure-activity relationship networks. *J. Chem. Inf. Model.* 60 (12): 6120–6134. <https://doi.org/10.1021/acs.jcim.0c00947>.
- 98 <https://aws.amazon.com/ecs> [accessed 20 September 2022]
- 99 <https://docs.aws.amazon.com/AmazonECS/latest/userguide/what-is-fargate.html> [accessed 20 September 2022]
- 100 <https://www.biosolveit.de/products/infinisee> [accessed 20 September 2022]

16

Visualization, Exploration, and Screening of Chemical Space in Drug Discovery

José J. Naveja^{1,2}, Fernanda I. Saldívar-González³, Diana L. Prado-Romero³, Angel J. Ruiz-Moreno⁴, Marco Velasco-Velázquez⁵, Ramón Alain Miranda-Quintana^{6,7}, and José L. Medina-Franco³

¹University Cancer Center, Third Department of Medicine, University Medical Center of the Johannes Gutenberg-University, Langenbeckstraße 1, 55131 Mainz, Germany

²Institute of Molecular Biology gGmbH, Ackermannweg 4, 55128 Mainz, Germany

³National Autonomous University of Mexico, DIFACQUIM Research Group, Department of Pharmacy, Avenida Universidad 3000, Mexico City, 04510, Mexico

⁴University of Groningen, University Medical Center Groningen, Hanzeplein 1, 9713 GZ Groningen, Groningen, The Netherlands

⁵Universidad Nacional Autónoma de México, Departamento de Farmacología, Facultad de Medicina, Avenida Universidad 3000, Mexico City, 04510, Mexico

⁶University of Florida, Department of Chemistry, Gainesville, FL 32611, USA

⁷University of Florida, Quantum Theory Project, Gainesville, FL 32611, USA

16.1 Introduction

With computer-aided drug design (CADD) and new artificial intelligence (AI) techniques, it has been possible to accelerate the generation of knowledge from big data in biological, chemical, and pharmaceutical medicine [1]. The methods developed in CADD, which have been optimized with machine learning (ML) algorithms, can use the vast chemical space combined with its biological information to obtain compounds with safety, efficacy, and low toxicity. CADD has led to the identification and development of many drugs used in the clinic and clinical development [2]. Figure 16.1 shows the chemical structures of drugs in clinical use and clinical development where CADD methods have contributed to their identification or development.

In the last two decades, improvements to structure- and ligand-based drug design methods developed in CADD have been described, many of them driven by AI and its subfields, ML and deep learning (DL) [3–6]. For instance, in structure-based drug design (SBDD), the prediction of three-dimensional structures with the AlphaFold2 neural network has generated the most complete and accurate picture of the human proteome [7], even highlighting its applications in cases where no similar structure is known [8]. Other notable applications of DL are predictions of chemical reactions [9], synthesis automation, and de novo design [10].

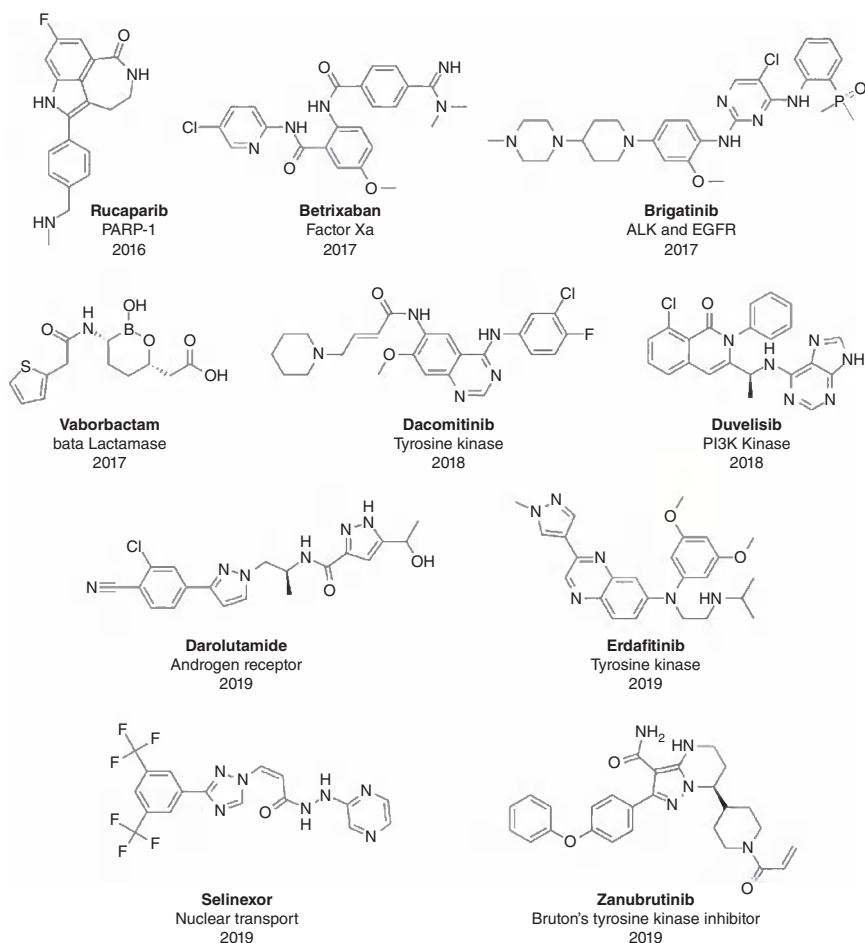


Figure 16.1 Chemical structures of exemplary drugs recently developed with the aid of computer-aided drug design. The main target and approval year are indicated.

The goal of the chapter is to discuss progress on selected concepts and applications of CADD. Because of its broad scope, this manuscript is not meant to be a comprehensive review of CADD. It discusses progress on representative concepts, resources, and applications of CADD that are part of multidisciplinary efforts to advance drug discovery. Throughout the manuscript, we emphasize public resources. The manuscript is organized into six sections. After this introduction, the next section analyzes the role of bioactivity data in CADD and discusses advances and opportunities in SBDD and ligand-based drug design (LBDD). Section 16.3 addresses the chemical space and chemical multiverse concepts to analyze the content and diversity of chemical libraries. Emphasis is placed on constellation plots. Section 16.4 explores exemplary applications of CADD to identify bioactive compounds. Therein, we introduce the concept of ViSAS: Virtual Screening of Analog Series, an implementation built upon the analog series formalism by Bajorath et al. [11, 12]. Section 16.5

describes recent advances in the development and application of extended similarity methods. Section 16.6 presents a summary of conclusions and perspectives.

16.2 Exploiting Bioactivity Data in the Artificial Intelligence Era

In the last decade, there has been an important increase in the amount of open bioactivity data in public data repositories. Databases with biological activity annotations have been extensively reviewed [13–15]. With the abundance of publicly available bioactivity data, it becomes imperative to extract, curate, and explore information of interest for drug discovery, so data-driven drug discovery models have emerged [16].

Analyzing information derived from large-scale structure–activity relationships (SAR) data can contribute to understanding the mechanism associated with the structural transformations of compounds that modify their activity [17, 18]. Similarly, available bioactivity data plays an important role in computational chemogenomics [19, 20]. Other approaches focus on identifying biologically relevant regions of chemical space to guide the design and synthesis of new compounds [21].

Recently, the importance of disclosing inactivity data in the public domain has been highlighted. In this regard, López-López et al. introduced the notion of structure-inactivity relationships (SIR), highlighting the importance of including reliable data on inactive compounds in the development of descriptive and predictive models [22]. In order to generate information and knowledge, the quantity and quality of data are vital as they improve the development and observed performance of chemoinformatics and AI models. As a scientific community, we should prioritize access to complete data, e.g. activity and inactivity data (negative results) that enable researchers to access the “big picture” of the available knowledge. Additionally, data curation and the construction of reliable databases are major issues. However, combined efforts could facilitate access to new, interesting data. Examples include natural products, metallodrugs, safety, preclinical, and toxicological databases that complement the current data available in the public domain and offer a new perspective on the known data.

AI methods enable the parallel study of very large volumes of diverse data for LBDD. However, SBDD approaches have not yet fully explored the utility of AI, although much research is in progress. One of the reasons is that experimental structural data are still sparse compared to compound activity and physicochemical data. The current protocols’ limitations only partially enable the generation of reliable 3D conformational states or binding modes. On the other hand, recent progress in AI-driven *de novo* structure prediction (see Section 16.4.3) has provided an unprecedented wealth of putatively reliable structural templates, with coverage recently approaching the entire protein universe.

There are several recent reviews discussing examples of AI applications in drug design and development [4, 23–26]. The challenges and opportunities that AI faces were recently discussed [27]. It has been emphasized that sufficient knowledge and

correct application are necessary. For this reason, it has been proposed to integrate “augmented intelligence” models into drug design, which shows a trend toward almost total automation (“*Human-assisted*”). This model of partnership between human intelligence and AI aims to improve cognitive performance, including learning, decision-making, and the generation of new experiences, by leveraging the capabilities offered by AI models and the medicinal chemist’s own expertise [28].

16.2.1 Databases Annotated with Biological Activity

Given the usefulness of chemical databases, many companies and researchers have taken it upon themselves to compile and put on web servers databases with diverse information. In an effort to classify these databases, Masoudi-Sobhanzadeh et al. distinguish five classes useful for drug repositioning based on data content, including raw data, target-based, specific data, drug design, and tool-based databases (including web servers) [15]. Other databases that are having a major current trend in drug discovery are virtual compound libraries and de novo-designed libraries [29]. Also, an effort is ongoing to build and curate a compound database with metal-containing molecules relevant in drug discovery [30].

Natural products have been sources of bioactive compounds [31]. The application of chemoinformatics approaches and AI to further advance natural product research is a current trend [32–34]. As early as 2012, there were efforts to put together natural product databases in the public domain [35]. The most recent large public compound collection is the Collection of Open Natural Products, COCONUT [36]. In Latin America, there is an ongoing effort to put together and curate compound databases composed of natural products from the region [37, 38].

16.2.2 Opportunities for AI in Ligand-Based Drug Design

In addition to in vitro and in vivo methods, in silico methods can mitigate serendipity. During rational drug design, serendipity might occur, leading to unexpected but potentially positive results such as the discovery of Lyrica (pregabalin) [39]. However, the increase of data in chemical databases with biological annotations limits the chance of serendipitous positive results and calls for enhanced methods for the identification of molecules with clinical applications [6]. Such tasks can be addressed by AI [6, 40]. Increasing numbers of AI algorithms are being developed for predicting the relationship between chemical structure and biological activity. For example, DeepChem is an open-source platform with tools for applying AI algorithms that allow the prediction not only of biological activity but also of multiple drug properties [41]. DeepChem has supported the development and benchmarking of new AI models [42–44].

Efficacy prediction using diverse inputs is one of the main objectives of applying AI in drug design. For example, Wang et al. used a model based on Support Vector Machines to discover nine new compounds and their interactions with four key targets [45]. Yu et al. used two random forest models with high sensitivity and specificity to predict possible drug-protein interactions by combining pharmacological and chemical data [46]. KinomeX can efficiently investigate the overall selectivity of

compounds for the kinase family and specific subfamilies of kinases, which can aid in developing new chemical modifiers [47]. Another example is PyRMD, an AI algorithm that can be trained to recognize the distinctive pharmacophoric features from the target bioactivity data available at the ChEMBL [48].

Identification of toxic effects in the early stages of drug design allows for removal of undesirable characteristics of bioactive compounds. At present, multiple AI-based methods are employed to assess toxicity by predicting off-target ligand binding. For example, Ligand Express uses proteome-screening data to find receptors that can interact with a specific small molecule, predicting on- and off-target interactions and suggesting the drug's potential side effects [49]. Other AI web-based tools that help predict toxicity include LimTox, pkCSM, admetSAR, and Toxtree [50]. A particularly remarkable case is DeepTox, an ML-based algorithm that predicted the toxicity of 12 707 environmental compounds and drugs during the Tox21 Data Challenge [51].

After a molecule has been virtually screened for potential bioactivity and toxicology, a chemical synthesis pathway is required for its evaluation. Despite knowledge of hundreds of thousands of transformation steps, novel molecules cannot be efficiently synthesized due to novel structural features or conflicting reactivities [52]. AI can help to identify possible and less complicated synthesis routes for compounds simultaneously or sequentially with prediction of bioactivity [53]. Computer-aided synthesis planning can also suggest millions of structures that can be synthesized and predict multiple synthesis routes for each of them [54].

New AI methods can support multiple applications, such as analog series identification, de novo drug design signatures study, SAR visualization, reactivity predictions, similarity searching, and visualization of chemical space. Two examples of such methods are Extended Similarity Indices, developed by the research group of Miranda-Quintana [55, 56], and the SAR Matrix approach and its DL extension by Bajorath et al. [57].

A strategy still to be consolidated is data expansion [58] using multiple layers of inputs. This approximation could allow the generation of the most representative similarity search to identify chemical mimetics capable of reverting disease signatures.

16.2.3 Opportunities in Structure-Based Drug Design

SBDD has reached notable maturity over the past decades, especially structure-based virtual screening [10]. In recent years, DL has been used in attempts to improve the performance of SBDD methods further. Perhaps the most well-known example is protein structure prediction. De novo structure prediction with AlphaFold [8] and RoseTTAfold [59] or other programs has yielded many protein models of near-experimental accuracy.

16.3 Chemical Space and Chemical Multiverse

Chemical space, molecular space, or “chemical universe” [60], recently extended to “chemical multiverse” [61], is a concept that is central and distinctive of

chemoinformatics [62]. Chemical space refers to all possible molecules as well as multi-dimensional conceptual spaces representing their structural and functional properties [63]. In contrast to cosmic space, chemical space is relative to the structural and functional properties used to construct or define a given chemical space. Since there is no molecular representation that captures all structural and functional properties, the concept of “chemical multiverse” has been introduced to account for the alternative chemical spaces of a compound data set that can be generated by different sets of descriptors [61]. Structural representation is the most relevant feature in basically any computational study [64], and it is an area under constant research [65]. In virtual screening, defining the chemical space to be explored is crucial, as it defines the applicability domain that will be searched. In practice, it is common to conduct virtual screening campaigns focused on regions of the medicinally relevant chemical space [66, 67]. Nonetheless, it is becoming a regular practice to explore novel regions of the chemical space, given by the emerging large and ultra-large chemical libraries [68].

The chemical space concept has practical applications in many areas of chemistry. In drug discovery, chemical space was used as a spatial navigation framework, which was helpful for understanding and generating knowledge of the pharmacokinetic properties and molecular diversity of biologically relevant compounds [69]. As the number of compounds and their information in chemical databases increased, more sophisticated molecular descriptors and visualization techniques were developed to expand their applications. For example, explorations of chemical space have considerably improved our understanding of biology and have led to the development of many tools for the exploration of SAR and structure–property relationships (SPR) [70, 71]. The availability of software libraries and the rise of AI [9] have led to the emergence of several tools that integrate ML methods as versatile tools to design, generate, and visualize the chemical space of small molecules [72, 73].

16.3.1 Recent Progress on Chemical Space

The chemical space concept has been of interest for a number of years. However, the rapid and continued increase of large and ultra-large libraries has renewed the interest of the scientific community to analyze their coverage in chemical space [29]. There are several reviews of chemical space, covering different aspects such as the enumeration of chemical compounds using *de novo* design, calculation of molecular descriptors, progress on visualization methods, web servers, and applications to study SPR [72–74]. A recent development in this area is the chemical library networks (CLNs) [75], further elaborated in Section 16.5.

16.3.2 Chemical Multiverse and Constellation Plots

As mentioned at the beginning of Section 16.3, the term chemical multiverse has been introduced as an expanded view of the chemical space [61]. This novel term is based on the chemical space concept that implies that a set of m molecules described with different descriptors would lead to distinct chemical spaces. Varnek and Baskin

point out that “unlike real physical space, a chemical space is not unique: each ensemble of graphs and descriptors defines its own chemical space” [62].

In physics, Everett’s multiverse [76] is “a hypothetical collection of potentially diverse observable universes, each of which would comprise everything that is experimentally accessible by a connected community of observers.” In analogy with the cosmic multiverse, the chemical multiverse was defined as “the group of numerical vectors that describe it differently from the same set of molecules” [61]. As reviewed recently [61], different chemical space representations can lead to alternative spaces, and the relationships between chemical compounds could change. It has been shown that the concept of chemical multiverse is applicable to different types of molecules, such as small organic molecules and peptides for drug discovery applications, food chemicals, and natural products. Eventually, the chemical multiverse can be expanded to any type of compound, including inorganic compounds.

A common limitation of most visualization methods of chemical space is that they capture a single type of molecular representation, emphasizing the dependence of the chemical space on the structure representation. To address this issue, constellation plots, generally depicted in Figure 16.2, combine, in a single graph, multiple structural representations, providing a broader perspective of the contents, diversity, and, if desired, a property of interest (e.g. biological activity, either experimental or predicted). Specifically, constellation plots combine a coordinate-based chemical space representation of analog series. Constellation plots facilitate the identification of entire zones in chemical space enriched with active compounds (“bright” SAR) or with predominantly or all inactive molecules (“dark” regions or “black holes”). In analogy with cosmic space, the name “constellations” is associated with clusters of analog series with similar chemical structures (given by similar coordinates in the two-dimensional plot). Combining multiple structural representations is founded on the general notion that multiple and well-integrated approaches perform overall better than individual methods [77–80]. Since constellation plots combine various structural representations, these plots are visual representations of chemical multiverses.

Virtually any property of interest can be depicted in a constellation plot, such as experimental activity data or results from virtual screening. This can be useful to identify, for instance, promising analog series for prioritization in experimental screening or additional computational studies.

Constellation plots have already been used to aid the visualization of chemical space for different practical applications. For example, the authors analyzed the results of a docking-based virtual screening of 2789 molecules from a commercial virtual library focused on inhibitors of DNA methyltransferase (DNMT). The docking scores were visually represented on the plot, enabling the rapid identification and grouping of analogs (“constellations”) of compounds to be prioritized for further screening [70]. Constellation plots have also been used to explore the SAR of 827 inhibitors of AKT1 obtained from a public database and the structure-multiple-activity relationships – SmART – of 286 molecules experimentally tested as inhibitors of three DNMTs and assembled from public sources [70, 81–83] in a consistent cell-selective analog series of chemical compounds. This

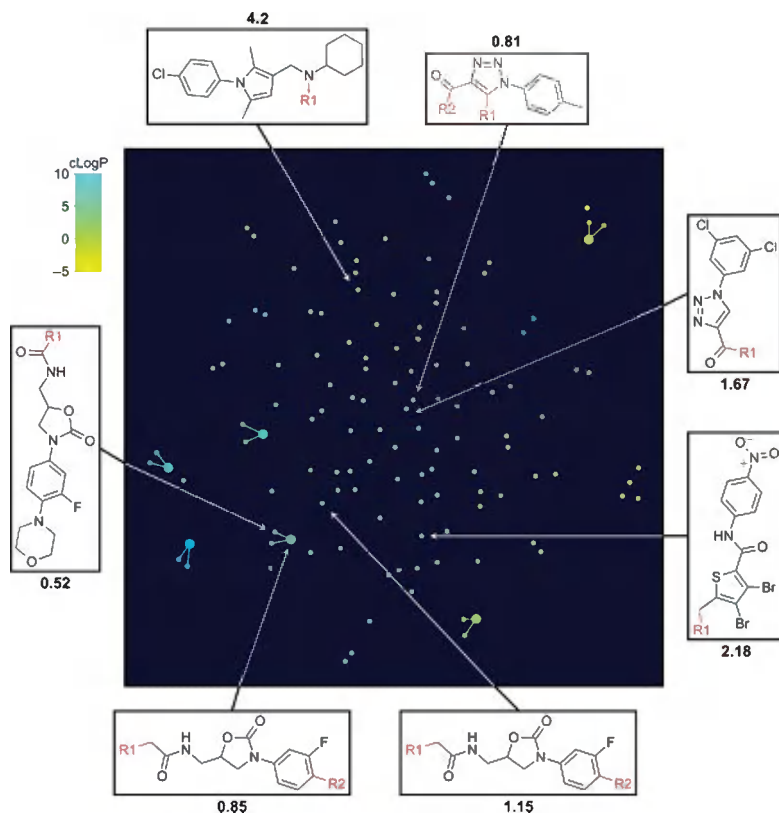


Figure 16.2 The general form of a constellation plot is illustrated in this image. Every core is represented by a dot, the size of which is proportional to the number of compounds mapped to it. Edges represent cores connected by at least one shared molecule in the dataset. The color coding can represent any feature, such as the average scores of the molecules represented by the corresponding core in virtual screening. In this example, the color indicates the average of the cLogP values of the compounds sharing the core structure.

analysis was done through a systematic analysis of high-throughput screening data of 41 821 compounds consistently assayed against the same panel of 73 human cancer cell lines used by the National Cancer Institute of the United States. In that study, the most relevant analog series were identified as measured by a therein-developed combined selectivity and consensus score. Also, all 3750 cores of the entire data set were used as queries or reference structures to virtually screen the entire ZINC 15 database, identifying 82 409 purchasable analogs for 1980 of the 3750 cores [82].

One more recent application of the constellation plots was to contribute to a comprehensive SAR analysis of 851 compounds tested as tubulin inhibitors and their bioactivity data in different cancer cell lines. A total of 147 analog series were identified and analyzed in a constellation plot. Visual analysis of the plot identified “bright” and “dark” regions in chemical space, i.e. analog series with overall high and low activity, respectively, as inhibitors of tubulin [83]. The code to generate constellation plots is freely available at <https://github.com/navejaromero/analog-series>.

16.4 Hit Identification, Optimization, and Development of Bioactive Compounds

One of the most frequent approaches to identify active compounds from large compound libraries is through the computational filtering of possibly large or extremely large screening compound databases, followed by experimental validation.

16.4.1 Virtual Screening

Virtual screening is an approach devised to identify promising molecules [84, 85]. Hits selected based on other features, such as physicochemical properties and smooth SAR, become leads ready to undergo optimization cycles [86]. With the rise of large and ultra-large chemical databases, virtual screening has evolved as a natural way to exploit their contents and diversity [87]. Table 16.1 lists a few examples of successful cases of virtual screening [2].

16.4.2 VISAS: General Approach that Expands Bioactive Molecules

Sometimes a focused virtual screening approach might be desirable. One way of delimiting the chemical space implies carefully selecting the collection for the screen. For instance, the database could be confined to drug-like molecules, natural products, synthesizable compounds, or focused or targeted libraries. Thus, the hits will comply with relevant selection criteria, depending on the project. A second factor to consider is the similarity threshold to define hit compounds. A more flexible hit definition puts novelty and scaffold-hopping in the spotlight, whereas stringent criteria would be suitable for identifying lead compounds, SAR analysis, and hit expansion. We consider that the most extreme hit definition requires that the query and hit molecules are chemical analogs, i.e. molecules with a

Table 16.1 Examples of recent successful virtual screening campaigns.

Hit compound(s) activity	Virtual screening approach	Reference
11 β -HSD1 inhibitors.	Growth-based screening of the ZINC database (1.8 million compounds).	[88]
Two SARS-CoV-2 M ^{pro} inhibitors with IC ₅₀ values in the micromolar range.	Docking-based virtual screening of an in-house focused library.	[89]
Thirty-two inhibitors of Notum with IC ₅₀ values lower than 500 nM.	Docking-based virtual screening of 1.5 million compounds in a synthetic and commercial library (ChemDiv).	[90]
Four histone deacetylase inhibitors with nanomolar activity vs. HDAC 1, 3, and 6.	Pharmacophore model and docking of an in-house database of 22 700 molecules.	[91]
Six compounds with activity against <i>Mycobacterium tuberculosis</i> peptide deformylase.	Docking-based virtual screening of a commercial compound library with 7120 small molecules.	[92]

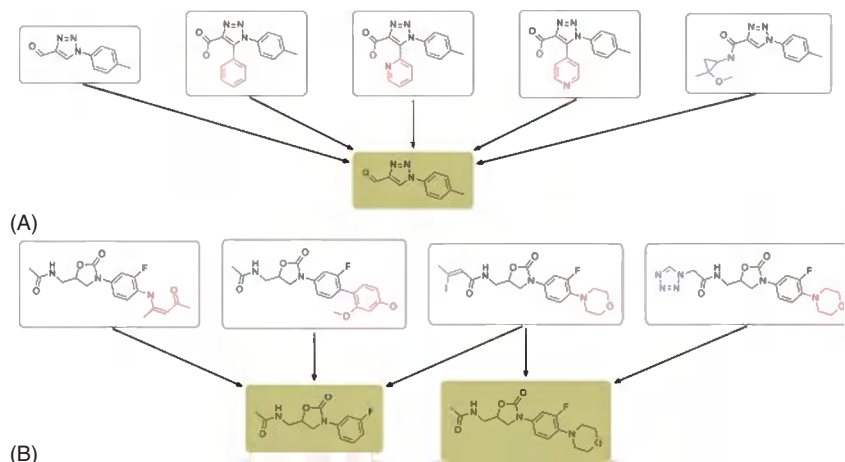


Figure 16.3 The general concept of analog series. All molecules in series A share a common core, which, for some applications, could be used to summarize it. Series B is somewhat more complex and requires at least two minimally overlapping cores for a comprehensive representation. Note that our definition of analog series allows every molecule to map to multiple cores. For clarity, not all putative cores are shown in this figure. See reference [93] for more details on the fragmentation-and-indexing algorithm employed. Source: Adapted from Naveja et al. [93].

close synthetical relationship. Since queries and hits might as well have arisen from an organic synthesis project, it might be understood as a “pseudo-optimization” algorithm enabling the rapid extraction of purchasable or readily available analogs for experimental SAR exploration. We term this approach ViSAS (Virtual Screening based on Analog Series) since the practical implementation builds upon the analog series formalism by Bajorath et al. [11, 12] Figure 16.3 depicts two exemplary analog series according to the definition presented in [93]. Briefly, the process of finding putative cores for a molecule begins with fragmenting the molecule (for instance, using RECAP retrosynthetic rules [94]) and subsequently filtering for relevant fully connected fragments that include most of the original structure (we require that at least two-thirds of the heavy atoms from the molecule must be included in the fragment’s structure). Fragments obtained through this procedure are termed putative cores. Although this method allows every molecule to map to more than a single core, large analog series can be summarized in a few cores that comprehensively map all molecules in the series (Figure 16.3). Nevertheless, keeping a record of all putative cores permits the later inclusion of new molecules, which is the principle on which we base the virtual screening approach presented here.

Algorithms and applications related to the automatic identification of analog series in large data sets have been reviewed [12]. For over a decade, the analog series algorithms derived from matched molecular pair analysis have demonstrated a compelling balance between chemical interpretability and scalability [95]. Recent developments have emphasized the ability of analog series for SAR and activity

cliffs rationalization [11, 96, 97]. However, other industrial applications, such as the evaluation of progress in lead optimization [98], highlight the potential for analog series analyses to assist drug discovery teams dealing with organic synthesis and biological evaluation [12].

The formulation of virtual screening from the analog series emerges from the definition of chemical analogs: two molecules are considered analogs if they share a common core structure. Therefore, a typical fragment-and-index approach lists all possible matching cores for molecules in a dataset. Any new molecule that could be reduced to a fragment matching the fragment list would be an analog of the molecule(s) in the dataset indexed to this fragment. It remains only to define a fragmentation procedure and the requirements of a fragment to be considered a valid core. Many different such approaches have been reviewed elsewhere [12, 95]. For instance, exhaustive methods may consider every possible substructure to be a valid core. Nonetheless, such strategies might lead to practical limitations. For instance, even relatively small libraries of somewhat complex molecules might lead to a combinatorial explosion during exhaustive substructure enumeration. Furthermore, synthetic interpretability is not prioritized in this approach, thus leading to a harder rationalization of the results. Therefore, matched molecular pairs obtained through retrosynthetic fragmentation [99] gradually developed into several applications relying on analog series computational identification [11], such as analog series-based scaffolds [100], compact chemical space representations of analog series in constellation plots [81], and the novel SAR rationalization approaches [93, 97].

Another application of analog series yet to be fully harnessed is virtual screening in ultra-large libraries. While most virtual screening methods focus on identifying single molecules with a desired predicted property, working with analog series up front has the potential of readily identifying a whole family of compounds to be prioritized for additional computational analysis or tested experimentally for a richer and in-depth SAR analysis. In essence, ViSAS is a substructure search algorithm (see Figure 16.4). However, the valid substructures to search are delimited before a direct comparison between queries and compounds in the database to search occurs. This allows the fragmentation of the databases to be computed in advance, thus reducing the substructure search to a text-matching problem. Moreover, the inherent hierarchical structure of analog series can be represented as scaffold networks and R-group tables, allowing prompt local SAR analyses early on.

We fragmented ZINC15 to prepare it for virtual screening. Although fragmentation is time-consuming, fragment-and-index approaches require fragmenting each molecule only once. This implies that updates would be faster, as only new molecules have to be processed and added to the dictionary. Any new molecule that is processed undergoes a standard washing procedure consisting of salt removal, extraction of the largest fragment, charge neutralization, and removal of stereochemistry information. Afterward, the washed molecule is searched in the list of processed SMILES, to avoid processing a compound twice. This list maps every unique washed SMILES to the identifiers – IDs – of the compounds mapping to it after the washing procedure. Any new SMILES are fragmented as described in [93, 101]. The fragmentation procedure is easy to run in parallel, as every molecule can be processed

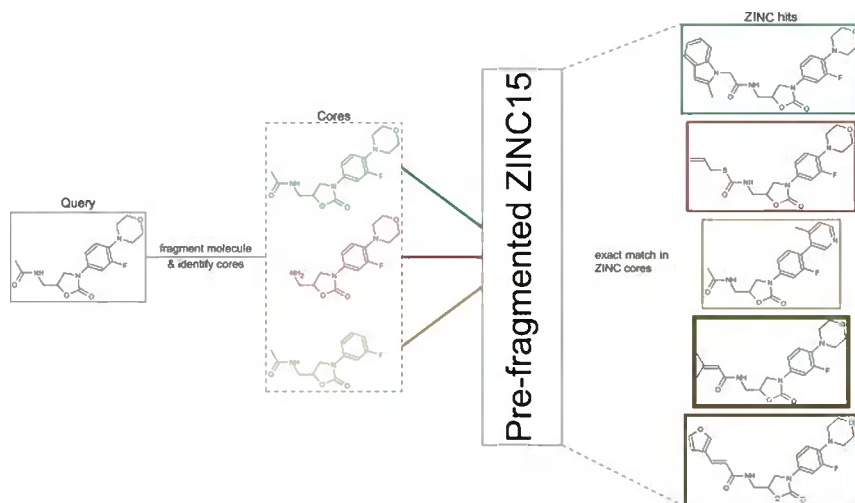


Figure 16.4 Virtual screening of analog series (ViSAS) concept. In this example, one query molecule is fragmented through RECAP rules, and only fragments retaining at least two-thirds of the heavy atoms in the query are considered cores. The cores are then used for searching for exact matches in the precomputed cores of the ZINC database. This allows searching for chemical analogs in ultra-large libraries (in this case, >740 million unique molecules). For each core, an R-group table with the matching compounds can be computed.

independently. We provide bash scripts for downloading and processing ZINC at <https://github.com/navejaromero/analog-series>. Also, the post-fragmentation ZINC library can be downloaded from Zenodo (10.5281/zenodo.6562818). To further show the application of the analog series in virtual screening, in the next section, we discuss a case study using public data to address a global health issue.

16.4.2.1 ViSAS on an Antituberculosis Chemical Dataset

The process of hit expansion itself using the processed ZINC database has been described by Madariaga-Mazón et al. [102]. In this case study, for the purpose of illustrating ViSAS, we used 118 published antituberculosis compounds as queries [103]. The fragmenting procedure identified 261 cores, which were then used for the text search in the preprocessed ZINC database. 3091 computational hits were identified. Sixty-seven cores matched at least one molecule in ZINC; however, only seven minimally overlapping cores resulted in more than two hits (Figure 16.5). Note that this method only finds analogs by matching cores; it is not designed to directly add more cores to the chemical space, but only to enrich those that are already represented in the queries. Nonetheless, the results might be fragmented again and used for another round of virtual screening; this would increase the coverage of the chemical space at the cost of adding more diverse analogs. In this example, the total size of the database increased ~26-fold. However, a significant number of hits could be found only for a few cores, whose SAR could be characterized. For instance, the most represented core had over a thousand hits and two substitution sites. A selection of these hits might be readily acquired and tested experimentally (Figure 16.6).

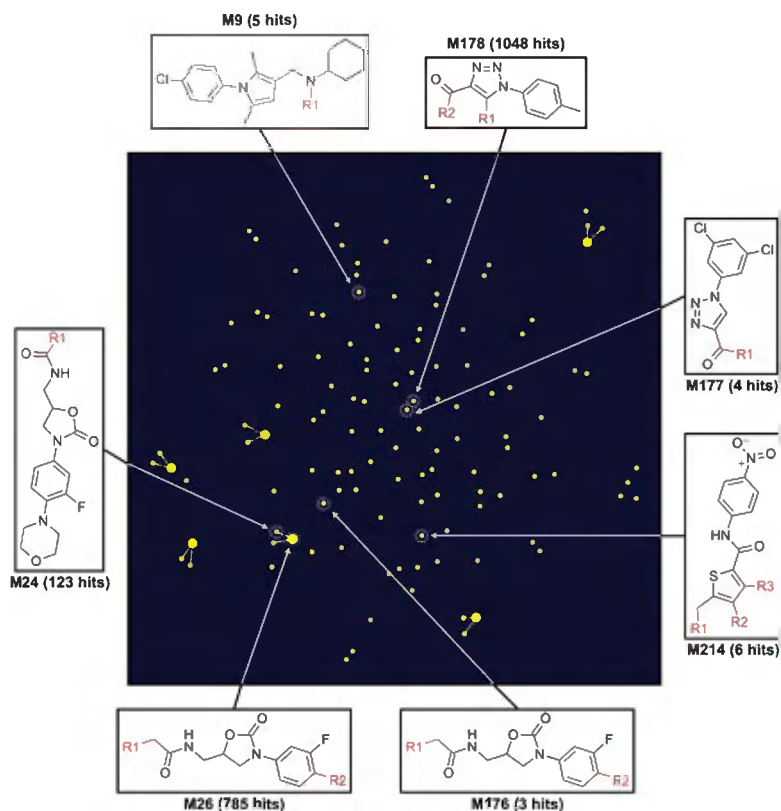
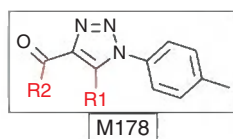


Figure 16.5 Constellation plot depicting the core chemical space of a collection of 118 molecules with antituberculosis activity from the core's viewpoint. Every dot represents a valid retrosynthetic core. Larger points represent cores to which two molecules are mapped. Six complex analog series were found, forming constellations in the original data set. ZINC15 was searched for analogs of any of the cores, successfully finding more than a single molecule for seven of them (structures shown and dots highlighted with a clear halo). For simplicity, only 124 cores summarizing the whole core space are plotted; these were selected for minimal overlapping as described in [93].

16.4.3 De Novo Design Libraries

The main goal of de novo design is the proposal of novel chemical entities. This computational approach considers a series of constraints to construct new molecular structures. These constraints could include desired biological effects, drug-likeness, pharmacokinetic properties, toxicity, or chemical feasibility. The desired biological effect is considered the primary constraint for the reason that all programs contemplate this objective [105]. In addition, de novo design software has to address three tasks: the assembly of the new molecules, molecule scoring, and optimization of the molecules [106]. More than forty algorithms for de novo design have been published since the early 1990s. Taking into account the available information, structure-based or ligand-based approaches can be selected. The three-dimensional coordinates of the receptor are fundamental for the first and active binders for the



ID	R1	R2	Price (USD)
ZINC000065288225		[R]OH	\$8.00
ZINC000011730479		[R]OH	\$6.30
ZINC000011730472		[R]OH	\$8.30
ZINC000013355576	[R]H		\$1.79
ZINC000065036590	[R]H		\$87.00
ZINC000040871814	[R]H	[R]OC	\$7.80

Figure 16.6 R-group table showing a selection of the 1048 analogs matching **M178**, the most populated core from the antituberculosis collection in [103], matching the processed ZINC database. Prices as of May 2022, according to the ZINC Express website [104]. Source: Adapted from [103, 104].

second [105]. A recurrent ligand-based strategy is the definition of a pharmacophore model from an ensemble of known actives. PhDD [107] is a pharmacophore-based de novo design method, it incorporates the assessment of synthetic accessibility, and the bioactivity of the proposed molecules is estimated with a fit value to the pharmacophore model. Another example of ligand-based de novo design is the reaction-based software DOGS [108]. This program recommends a synthetic route for each compound.

Most recent software considers fragments above atoms as building blocks. Fragments' databases come from different sources, including external inputs such as catalogs or fragment-like compounds. Databases can also be constructed from the virtual fragmentation of complete drug molecules to increase the probability of obtaining a drug-like molecule with synthetic accessibility [109]. This last strategy can be exploited not only with complete drugs, but also with known active compounds against our target of interest. To follow the strategy of using bioactivity data, the selected program has to admit this type of focused fragments. Examples of software with this characteristic are LUDI [110] and LigBuilder [111].

It is possible that the software incorporates the fragmentation step from complete molecules entered by the user, like alvaBuilder (Alvascience, alvaBuilder, version 1.0.6, 2021, www.alvascience.com). In case the virtual fragmentation has to be made before entering the information into the program, it is necessary to construct a database of active modulators. The threshold value to separate actives and inactives can be established in 10 μ M [112]. Another strategy is to analyze the bioactivity data

of the selected compounds for the target of interest, for example, the median can be calculated to set a different limit to the particular database. Once the active set is ready, molecular fragmentation can be done with algorithms like RECAP [94], and the resulting fragments with suitable properties are selected as input.

To deal with the tasks of molecular generation and the increasing amount of available bioactivity data, AI has been applied to automated de novo design. Taking into account the scoring of molecules, ML approaches like target prediction, which classifies compounds into active and inactive, or quantitative structure–activity relationships (QSAR) could be applied [113]. Inverse QSAR or inverse quantitative structure–property relationships – QSPR – are also applied to de novo design. These methodologies seek to correlate desired properties, including biological activity, to molecular structural features [114].

Research work from 2018 proposed an approach based on a generative model that made use of a recurrent neural network for de novo drug design. The model was trained with a large molecular set from the ChEMBL database. With this training, the model learned the grammar of SMILES, the chosen molecular representation for the molecules. To generate focused libraries, the model was fine-tuned with active modulators of a specific target. This was another strategy that took advantage of bioactivity data to generate novel molecules [113].

16.4.3.1 Case Study: DNMT-Focused Libraries

This example is centered on the discovery of new hits against DNMT. We hypothesized that automated de novo design could lead to molecules that expand the epigenetic-relevant chemical space, due to the expected novelty of the compounds. We selected AlvaBuilder version 1.0.6 [115], to construct the molecules. This software selects fragments as construction blocks and incorporates a genetic algorithm to optimize the search for suitable compounds. To create the fragments and linkers, the user has to select a database of entire molecules. The user enters data to create the training set, the molecules are then fragmented into ring systems, linkers, and side chains. Biologically active modulators could constitute the training set. For the selection of the training set, we constructed a database of DNMT1 inhibitors with an IC_{50} of 10 μ M or less. Bioactivity data was obtained from ChEMBL, version 29 (2021). Since the bioactivity data reported for DNMT1 inhibitors could differ between experimental techniques [116], we only maintained 422 molecules with IC_{50} values (48% of those with annotated bioactivity). After data curation, we had 259 unique compounds with the desired inhibitory concentration.

The scoring function is also customized by the user; the score is a conglomeration of a set of rules. The first rule finds molecules that have a target value for the selected descriptor, from the 91 available. The second rule calculates the similarity to a reference, and the third evaluates if the molecules contain or do not contain a molecular pattern. The result of the scoring aggregates with either an arithmetic or geometric mean and exhibits values from zero (worst) to one (best). To establish the scoring function for the case study, we selected seven descriptors: molecular weight, donor atoms for H-bonds, acceptor atoms for H-bonds, consensus LogP model, LogS

aqueous solubility, synthetic accessibility score (SAscore), and topological polar surface area. Synthetic accessibility is one of the major concerns in *de novo* design. Therefore, we included this quantitative estimation in addition to other physicochemical properties.

We calculated the descriptors of the active molecules with alvaDesc (Alvascience, alvaDesc [software for molecular descriptors calculation] version 2.0.10, 2021, www.alvascience.com). This program has the same algorithms for the computation of descriptors as alvaBuilder. With this information, we set up the donor atoms for H-bonds to be ≥ 2 and the SAscore to ≤ 5.979 . For the rest of the descriptors, the range was designated to the mean $\pm$ the standard deviation of the calculated numerical values for the active inhibitors. The final score was aggregated with the arithmetic mean of the selected rules. We defined a population size of 65 and a maximum number of iterations of 100 for the genetic algorithm. With the same training set and scoring function, we obtained 10 sets of new molecules.

With the ten different sets, we computed similarity matrices with the Platform for Unified Molecular Analysis – PUMA – server [117]. The results confirmed that the predicted physicochemical properties are highly similar, with Tanimoto coefficients between 0.969 and 0.983. The similarity results were expected due to the definition of the scoring function. Since we confirmed that molecular properties were alike, we also wanted to compute the structural similarity between the compounds. We calculated two different fingerprints: MACCS keys (166-bits) and Morgan radius 2 with RDKit node for KNIME. Preliminary results showed that the similarity between inter and intraset is lower than the one calculated with molecular properties. Cumulative distribution functions computed with PUMA showed median similarity values from 0.471 to 0.590 with MACCS keys and 0.114–0.149 with Extended Connectivity Fingerprints radius 4, both results presenting intersets similarity. Overall, the results showed that new molecules exhibit highly similar properties. These molecular properties were established as secondary constraints by the scoring function. Nevertheless, the sets exhibit less structural similarity, according to the selected fingerprints. The calculated structural diversity is expected for a *de novo* design. In this case, it could also be influenced by the initial diversity of the training set. This is encouraging due to the probability that the desired bioactivity could also be transferred to the novel molecules.

16.5 Extended Similarity Methods

Binary similarity indices [118] and metrics are core elements of the machinery used to explore chemical space, classify molecules, design new drugs, and screen molecular libraries in search of promising compounds [119–125]. However, as well-studied and ubiquitous as they are, these indices have a fundamental drawback, given by the fact that they can only compare two molecules at a time. This means that if we want to estimate the similarity of N compounds, we will need $O(N^2)$ operations, which greatly limits the scaling of these algorithms and restricts them to narrow sections of chemical space. Motivated by these issues, a new family of similarity indices [55, 56]

(extended or n -ary similarity indices) was recently proposed that can compare multiple molecules at the same time. In this section, we briefly review the characteristics of these indices and some exemplary applications.

16.5.1 The Extended Similarity Framework

The defining characteristic of the extended similarity indices [55, 56] is that they are capable of comparing any number of molecules at the same time. Remarkably, the procedure leading to this generalization is extremely simple, which contributes to the ease of implementation of these indices. The starting point is having all the molecules that are going to be analyzed in a suitable representation. For now, all the cheminformatic-related applications of the n -ary indices have mostly relied on binary fingerprints (e.g. any type of fingerprint, including MACCS keys, RDKit fingerprints, and circular or Morgan fingerprints), but there has been extensive work on generalizing the domain of definition of these indices, so one could also use arbitrary sequence representations [126], latent-space descriptor-based approaches [127], or even coordinate-based 3D representations [128]. The key is that all the molecules will be encoded by equal-length “vectors.” Then, we just need to form a cumulative vector, σ , with components σ_k equal to the sum of each of the components of the molecules to be analyzed (if the molecular representations are aligned in a matrix-like array, then σ will correspond to the sums of the columns of this matrix). This is the most time-consuming step in the calculation of the n -ary indices; however, it is easy to see that it will scale as $O(N)$, so it is dramatically more efficient than using any binary comparison (e.g. recent benchmarks have been able to compare tens of billions of molecules using extended indices in a regular laptop). The next step is to define a coincidence threshold, γ , that is going to be used to determine which of the components of σ are going to be identified as corresponding to similarity or dissimilarity descriptors. This classification is extremely easy to perform since we just need to notice that $|2^*\sigma_k - n| > \gamma$ indicates that σ_k will be a similarity, while $|2^*\sigma_k - n| \leq \gamma$ indicates dissimilarity (with n being the number of molecules to be compared). We can gain even more detail into the nature of the similarity descriptors: if $2^*\sigma_k - n > \gamma$, the similarity will be given by the dominant contribution of “on” bits in the fingerprints (they mostly share the presence of the same feature), but if $n - 2^*\sigma_k > \gamma$, then the similarity will be given by the preponderance of “off” bits in the fingerprints (the given feature is mostly absent from the molecules). Of course, this means that unless we select $\gamma = n - 1$ (which is an extreme choice), we are going to consider similar descriptors that do not necessarily correspond to a perfect coincidence of the features. To properly account for this, we have to introduce weight functions in the formalism, which will penalize these partial coincidences. Then, with all these ingredients, we just need to substitute the (weighted) 1- and 0-similarity, together with the dissimilarity descriptors, in the same expressions used to define the binary similarity indices, and we have their corresponding extended version.

These new indices give more freedom at the time of performing any similarity-based analysis because now we can study correlations between an arbitrary number

of molecules [56]. Reassuringly, it has already been shown that they are internally and externally consistent with respect to the newly introduced hyper-parameter γ [118, 129, 130]. The former implies that they will rank multiple datasets in the same way, largely independently of the value of γ . The latter reflects the fact that the ranking obtained from extended indices and the ranking obtained from standard binary indices will also be the same over most γ values.

16.5.2 Global Description: Chemical Diversity, Chemical Library Networks, and Clustering

The ability of the extended similarity indices to require $O(N)$ operations to compare N objects makes them immediately attractive to explore large sections of chemical space, potentially dealing with numbers of molecules out of reach for current approaches. The first direct application related to this is quantifying the chemical diversity of large molecular libraries [75]. The key insight here is that while raw similarity values can be hard to interpret (except in the trivial cases when they are 0 or 1), we can easily determine the relative degree of diversity with respect to a given reference dataset. This makes it easier to readily interpret what one means when saying that libraries are more or less diverse. Several detailed benchmarks showed that the combination of RDKit fingerprints and the extended Tanimoto index provides the most robust measure of chemical diversity.

The hyper-efficiency of the n -ary indices motivated their use in representations of chemical space spanning millions of molecules. The inspiration here came from Maggiora and Bajorath's chemical space networks (CSNs) [131]. The CSNs start from a set of molecules and use these molecules as nodes/vertices in a graph. Then, one decides whether to connect these nodes with edges depending on the (binary) similarity between the involved molecules. As powerful and intuitive as this process can be, it has the same problem as other binary-based approaches, since it demands $O(N^2)$ operations. The extended similarity indices, on the other hand, can be used to define CLNs [75], which borrow inspiration from the CSNs, but now the nodes of the graph correspond to complete libraries, and the edges are associated to the extended similarity resulting from comparing any two such datasets. Preliminary studies have applied this methodology to representing chemical spaces with more than 18 million molecules, which is several orders of magnitude more than what has been represented using CSNs. This provides an unprecedented opportunity to map extremely large sections of chemical space, study how the connectivity patterns depend on factors like molecular representation, and potentially see how the relations between large libraries evolve in a dynamic way when elements are added or removed from them. This versatility strongly points out to the potential of CLNs to be used in polypharmacology and drug repurposing [132, 133].

The structure of the extended similarity indices naturally leads to a new way of performing hierarchical clustering. [56, 134] Notice that if in any given moment (e.g. k^{th} iteration) we have clusters $c_1(k)$, $c_2(k)$, ..., $c_N(k)$, we can proceed to the $(k + 1)^{\text{th}}$ iteration by combining the two clusters that maximize their joint extended similarity. In other words, we have a new linkage criterion. While the scaling of this algorithm is

the same as those relying on standard linkage criteria like single, average, complete, etc., the n -ary clustering has two key advantages. On one hand, this new clustering algorithm has proven to be more robust than current methods, as quantified by the V-measure [135]. Moreover, with the extended clustering, we can provide a very convenient estimate of the number of clusters in the data, without any extra computational cost. Recent studies have shown that this new method is capable of readily classifying various JAK inhibitors derived from different scaffolds [56].

16.5.3 Local Description: Diversity Selection and Medoid Calculation

A less obvious advantage of the n -ary indices is their ability to shed light on the local structure of large chemical spaces by singling out a molecule or sampling a few species. The first of such applications that was explored in detail was the use of extended indices in diversity selection [56] (e.g. selecting a maximally diverse subset from a given library). There are several ways to do this using binary comparisons (like the MaxMin [136] and MaxSum [137] algorithms), but they scale as $O(N^2)$. On the other hand, with the extended similarity indices one can just directly maximize the diversity of the selected set by minimizing the extended similarity of the molecules that are going to be picked (e.g. the Max_nDis [56] or ECS_MeDiv [128] algorithms). This simple procedure scales linearly, so it is in the perfect position to handle very large datasets. Moreover, several benchmarks have shown that the Max_nDis algorithm can result in sets that are more than three times more diverse than those selected by MaxMin or MaxSum.

Perhaps even more surprisingly, the extended indices allow us to find the most representative elements of a set with great ease [134]. This is a key task in several fields, known as the medoid problem. However, the usual solutions either scale as $O(N^2)$, or use several approximations and stochastic tools to get down to $O(N\log N)$ [138]. The main insight to approach this problem using extended indices is to introduce the concept of complementary similarity. That is, given an element in the set, the complementary similarity is the extended similarity of all the elements in the set, except for the chosen one. It is then clear that the medoid will correspond to the element in the set that has the lowest possible value of complementary similarity. This simple recipe provides excellent results when applied to the analysis of biological ensembles [134] and has also been used to explore epigenetic-focused libraries [139]. An enticing possibility of this algorithm is classifying all the molecules in a library depending on how “central” or “outlier” they are, and information that we can use to select either “stars” or “satellites” in order to represent chemical space regions with more detail.

16.6 Conclusion and Outlook

CADD has made contributions to identify and advance drug candidates that are in clinical use. Over the past few decades, and more remarkably in recent years, AI (arguably better called augmented intelligence), altogether with even richer

databases, is advancing drug research at a tremendous speed. AI has made clear and dramatic advances in SBDD and LBDD. However, the scientific community should employ AI correctly, keeping in mind that “fashion vanishes with time but quality (reasoning) is timeless.”

As part of the large data sources currently available to train AI models, large and ultra-large compound databases are emerging, many of them being designed with the aid of *de novo* design. The chemical space is expanding, boosting the proposal and evolution of novel visual and graphical representations. The concept of chemical multiverse has recently emerged to capture alternative chemical spaces of a compound data set given by different molecular representations. In this context, the constellation plots, which are based on the concept of analog series, are visual representations of chemical multiverses that facilitate SP(A)R analysis and the assessment of virtual screening results. We point out that the community should look into underrepresented regions of chemical spaces, such as metal-containing compounds, food chemicals, and other molecules from natural sources.

Systematic computational searches in chemical and biological spaces are a common practice in drug discovery, with several documented successful cases. To this end, ViSAS is a general approach that expands bioactive molecules' chemical space by finding analogs in libraries of purchasable compounds. The hits can be arranged in an R-group table, similar to a molecule optimization campaign; however, this is probably a first in virtual screening. Although ViSAS is formalized in this chapter, its principles and applications have been published and discussed. We anticipate that local SAR analysis with ML can help select the most relevant analog series to test experimentally. Similarly, ViSAS can be extended to the analysis of *de novo* libraries.

In order to speed and facilitate the navigation of the large and ultra-large compound libraries, the extended or *n*-ary similarity indices were recently proposed. These novel indices have found a broad range of applications, such as quantifying the chemical diversity of (large) molecular libraries, the graphical representation of chemical space through CLNs, clustering, diversity selection, and identifying the most representative compound of a compound data set.

A practical perspective on the continued improvement of CADD and AI, besides the methodological challenges, is the convenience of formal training and education at the undergraduate and graduate levels. Several practitioners learn and practice CADD on the fly as the research needs to emerge. Formal training would prepare better researchers and practitioners of CADD and AI, and it will be advantageous for the continued improvement of communication between multidisciplinary research teams so that experts in CADD can communicate more effectively with medicinal chemists and other drug discovery team members.

Acknowledgments

F.I.-S.-G and D.L.P.-R are thankful to CONACyT for the granted scholarship numbers 848061 and 888207, respectively. JJN is grateful to the Alexander von Humboldt Foundation for a postdoctoral scholarship and to CONACYT for the National

Researchers Program. Authors thank grant support from the General Direction of Academic Staff Affairs (DGAPA), UNAM, Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (UNAM-DGAPA-PAPIIT), grants IN201321 and IV200121. R.A.M.-Q. acknowledges support from the University of Florida in the form of a startup grant and a UFII SEED award.

Abbreviations

AI	artificial intelligence
CADD	computer-aided drug design
CLN	chemical library networks
CSN	chemical space networks
DL	deep learning
DNMT	DNA methyltransferase
LBDD	SARligand-based drug design
ML	machine learning
QSAR	quantitative structure–activity relationships
RECAP	retrosynthetic combinatorial analysis procedure
SAR	structure–activity relationships
SBDD	structure-based drug design
ViSAS	virtual screening of analog series

References

- 1 Lee, J.W., Maria-Solano, M.A., Vu, T.N.L. et al. (2022). Big data and artificial intelligence (AI) methodologies for computer-aided drug design (CADD). *Biochem. Soc. Trans.* 50 (1): 241–252.
- 2 Sabe, V.T., Ntombela, T., Jhamba, L.A. et al. (2021). Current trends in computer aided drug design and a highlight of drugs discovered via computational techniques: a review. *Eur. J. Med. Chem.* 224: 113705.
- 3 Zhao, L., Ciallella, H.L., Aleksunes, L.M., and Zhu, H. (2020). Advancing computer-aided drug discovery (CADD) by big data and data-driven machine learning modeling. *Drug Discov. Today* 25 (9): 1624–1638.
- 4 Jiménez-Luna, J., Grisoni, F., Weskamp, N., and Schneider, G. (2021). Artificial intelligence in drug discovery: recent advances and future perspectives. *Expert Opin. Drug Discovery* 16 (9): 949–959.
- 5 Schneider, P., Walters, W.P., Plowright, A.T. et al. (2020). Rethinking drug design in the artificial intelligence era. *Nat. Rev. Drug Discov.* 19 (5): 353–364.
- 6 Mak, K.K. and Pichika, M.R. (2019). Artificial intelligence in drug development: present status and future prospects. *Drug Discov. Today* 24 (3): 773–780.
- 7 Tunyasuvunakool, K., Adler, J., Wu, Z. et al. (2021). Highly accurate protein structure prediction for the human proteome. *Nature* 596 (7873): 590–596.

- 8 Jumper, J., Evans, R., Pritzel, A. et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596 (7873): 583–589.
- 9 Miljković, F., Rodríguez-Pérez, R., and Bajorath, J. (2021). Impact of artificial intelligence on compound discovery, design, and synthesis. *ACS Omega* 6 (49): 33293–33299.
- 10 Bajorath, J. (2022). Deep machine learning for computer-aided drug design. *Front. Drug Discov.* 2. Available from: <https://www.frontiersin.org/articles/10.3389/fddsv.2022.829043/full>.
- 11 Stumpfe, D., Dimova, D., and Bajorath, J. (2016). Computational method for the systematic identification of analog series and key compounds representing series and their biological activity profiles. *J. Med. Chem.* 59 (16): 7667–7676.
- 12 Naveja, J.J. and Vogt, M. (2021). Automatic identification of analogue series from large compound data sets: methods and applications. *Molecules* 26 (17): <https://doi.org/10.3390/molecules26175291>.
- 13 González-Medina, M., Jesús Naveja, J., Sánchez-Cruz, N., and Medina-Franco, J.L. (2017). Open chemoinformatic resources to explore the structure, properties and chemical space of molecules. *RSC Adv.* 7 (85): 54153–54163.
- 14 Mendez, D., Gaulton, A., Bento, A.P. et al. (2019). ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* 47 (D1): D930–D940.
- 15 Masoudi-Sobhanzadeh, Y., Omid, Y., Amanlou, M., and Masoudi-Nejad, A. (2020). Drug databases and their contributions to drug repurposing. *Genomics* 112 (2): 1087–1095.
- 16 Kunimoto, R., Bajorath, J., and Aoki, K. (2022). From traditional to data-driven medicinal chemistry: a case study. *Drug Discov. Today* 27 (8): 2065–2070.
- 17 Hopkins, A.L. (2008). Network pharmacology: the next paradigm in drug discovery. *Nat. Chem. Biol.* 4 (11): 682–690.
- 18 Nogales, C., Mamdouh, Z.M., List, M. et al. (2022). Network pharmacology: curing causal mechanisms instead of treating symptoms. *Trends Pharmacol. Sci.* 43 (2): 136–150.
- 19 Jacoby, E. (2011). Computational chemogenomics. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 1 (1): 57–67.
- 20 Brown, J.B. *Computational Chemogenomics*. New York: Springer 12 p.
- 21 Saldivar-González, F.I., Lenci, E., Trabocchi, A., and Medina-Franco, J.L. (2019). Exploring the chemical space and the bioactivity profile of lactams: a chemoinformatic study. *RSC Adv.* 9 (46): 27105–27116.
- 22 López-López, E., Fernández-de Gortari, E., and Medina-Franco, J.L. (2022). Yes SIR! On the structure–inactivity relationships in drug discovery. *Drug Discov. Today* 27 (8): 2353–2362.
- 23 Bender, A. and Cortés-Ciriano, I. (2021). Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 1: ways to make an impact, and why we are not there yet. *Drug Discov. Today* 26 (2): 511–524.
- 24 Bender, A. and Cortes-Ciriano, I. (2021). Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 2: a discussion of chemical and biological data. *Drug Discov. Today* 26 (4): 1040–1052.

- 25 Bajorath, J. (2022). Artificial intelligence in interdisciplinary life science and drug discovery research. *Future Sci. OA.* 8 (4): FSO792.
- 26 Bajorath, J. (2021). State-of-the-art of artificial intelligence in medicinal chemistry. *Future Sci. OA.* 7 (6): FSO702.
- 27 Bajorath, J., Chávez-Hernández, A.L., Duran-Frigola, M. et al. (2022). Chemoinformatics and artificial intelligence colloquium: progress and challenges to develop bioactive compounds. *ChemRxiv* Available from: <https://chemrxiv.org/engage/chemrxiv/article-details/62f1a15d42ddf532a9b420af>.
- 28 Definition of Augmented Intelligence (2022). Gartner information technology glossary. *Gartner* Available from: <https://www.gartner.com/en/information-technology/glossary/augmented-intelligence>.
- 29 Warr, W.A., Nicklaus, M.C., Nicolaou, C.A., and Rarey, M. (2022). Exploration of ultralarge compound collections for drug discovery. *J. Chem. Inf. Model.* 62 (9): 2021–2034.
- 30 Medina-Franco, J.L., López-López, E., Andrade, E. et al. (2022). Bridging informatics and medicinal inorganic chemistry: toward a database of metallodrugs and metallodrug candidates. *Drug Discov. Today* 27 (5): 1420–1430.
- 31 Newman, D.J. and Cragg, G.M. (2020). Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *J. Nat. Prod.* 83 (3): 770–803.
- 32 Medina-Franco, J.L. and Saldívar-González, F.I. (2020). Cheminformatics to characterize pharmacologically active natural products. *Biomolecules* 10 (11): 1566.
- 33 Kirchmair, J. (2020). Molecular informatics in natural products research. *Mol. Inform.* 39 (11): e2000206.
- 34 Saldívar-González, F.I., Aldas-Bulos, V.D., Medina-Franco, J.L., and Plisson, F. (2022). Natural product drug discovery in the artificial intelligence era. *Chem. Sci.* 13 (6): 1526–1546.
- 35 Yongye, A.B., Waddell, J., and Medina-Franco, J.L. (2012). Molecular scaffold analysis of natural products databases in the public domain. *Chem. Biol. Drug Des.* 80 (5): 717–724.
- 36 Sorokina, M., Merseburger, P., Rajan, K. et al. (2021). COCONUT online: collection of open natural products database. *J. Cheminform.* 13 (1): 2.
- 37 Medina-Franco, J.L. (2020). Towards a unified Latin American natural products database: LANaPD. *Future Sci. OA* 6 (8): FSO468.
- 38 Gómez-García, A., Jiménez, D.A., Zamora, W.J. et al. (2023). Navigating the chemical space and chemical multiverse of a unified latin american natural product database: LANaPDB. *Pharmaceuticals* 16 (10): 1388.
- 39 Barenie, R., Darrow, J., Avorn, J., and Kesselheim, A.S. (2021). Discovery and development of pregabalin (Lyrica): the role of public funding. *Neurology* 97 (17): e1653–e1660.
- 40 Paul, D., Sanap, G., Shenoy, S. et al. (2021). Artificial intelligence in drug discovery and development. *Drug Discov. Today* 26 (1): 80–93.
- 41 Wu, Z., Ramsundar, B., Feinberg, E.N. et al. (2018). MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* 9 (2): 513–530.

- 42 Yang, K., Swanson, K., Jin, W. et al. (2019). Analyzing learned molecular representations for property prediction. *J. Chem. Inf. Model.* 59 (8): 3370–3388.
- 43 Minnich, A.J., McLoughlin, K., Tse, M. et al. (2020). AMPL: a data-driven modeling pipeline for drug discovery. *J. Chem. Inf. Model.* 60 (4): 1955–1968.
- 44 Altae-Tran, H., Ramsundar, B., Pappu, A.S., and Pande, V. (2017). Low data drug discovery with one-shot learning. *ACS Cent. Sci.* 3 (4): 283–293.
- 45 Wang, F., Liu, D., Wang, H. et al. (2011). Computational screening for active compounds targeting protein sequences: methodology and experimental validation. *J. Chem. Inf. Model.* 51 (11): 2821–2828.
- 46 Yu, H., Chen, J., Xu, X. et al. (2012). A systematic prediction of multiple drug-target interactions from chemical, genomic, and pharmacological data. *PLoS ONE* 7 (5): e37608.
- 47 Li, Z., Li, X., Liu, X. et al. (2019). KinomeX: a web application for predicting kinome-wide polypharmacology effect of small molecules. *Bioinformatics* 35 (24): 5354–5356.
- 48 Amendola, G. and Cosconati, S. (2021). PyRMD: a new fully automated AI-powered ligand-based virtual screening tool. *J. Chem. Inf. Model.* 61 (8): 3835–3845.
- 49 Cyclica launches ligand express *Cyclica*. 2022. Available from: <https://cyclicarx.com/press-releases/cyclica-launches-ligand-express-a-disruptive-cloud-based-platform-to-revolutionize-drug-discovery>.
- 50 Yang, X., Wang, Y., Byrne, R. et al. (2019). Concepts of artificial intelligence for computer-assisted drug discovery. *Chem. Rev.* 119 (18): 10520–10594.
- 51 Mayr, A., Klambauer, G., Unterthiner, T., and Hochreiter, S. (2016). DeepTox: toxicity prediction using deep learning. *Front. Environ. Sci.* 3. Available from: <https://www.frontiersin.org/articles/10.3389/fenvs.2015.00080>.
- 52 Collins, K.D. and Glorius, F. (2013). A robustness screen for the rapid assessment of chemical reactions. *Nat. Chem.* 5 (7): 597–601.
- 53 Hessler, G. and Baringhaus, K.H. (2018). Artificial intelligence in drug design. *Molecules* 23 (10): 2520.
- 54 Corey, E.J. and Wipke, W.T. (1969). Computer-assisted design of complex organic syntheses. *Science* 166 (3902): 178–192.
- 55 Miranda-Quintana, R.A., Bajusz, D., Rácz, A., and Héberger, K. (2021). Extended similarity indices: the benefits of comparing more than two objects simultaneously. Part 1: theory and characteristics. *J. Cheminform.* 13 (1): 32.
- 56 Miranda-Quintana, R.A., Rácz, A., Bajusz, D., and Héberger, K. (2021). Extended similarity indices: the benefits of comparing more than two objects simultaneously. Part 2: speed, consistency, diversity selection. *J. Cheminform.* 13 (1): 33.
- 57 Yoshimori, A. and Bajorath, J. (2021). Iterative DeepSARM modeling for compound optimization. *Artifi. Intel. Life Sci.* 1: 100015.
- 58 Gupta, R., Srivastava, D., Sahu, M. et al. (2021). Artificial intelligence to deep learning: machine intelligence approach for drug discovery. *Mol. Divers.* 25 (3): 1315–1360.

- 59 Baek, M., DiMaio, F., Anishchenko, I. et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 373 (6557): 871–876.
- 60 Ruddigkeit, L., Blum, L.C., and Reymond, J.L. (2013). Visualization and virtual screening of the chemical universe database GDB-17. *J. Chem. Inf. Model.* 53 (1): 56–65.
- 61 Medina-Franco, J.L., Chávez-Hernández, A.L., López-López, E., and Saldívar-González, F.I. (2022). Chemical multiverse: an expanded view of chemical space. *Mol. Inform.* 41: e2200116.
- 62 Varnek, A. and Baskin, I.I. (2011). Chemoinformatics as a theoretical chemistry discipline. *Mol. Inform.* 30 (1): 20–32.
- 63 Maggiora, G.M. (2014). Introduction to molecular similarity and chemical space. In: *Foodinformatics: Applications of Chemical Information to Food Chemistry* (ed. K. Martinez-Mayorga and J.L. Medina-Franco), 1–81. Cham: Springer International Publishing.
- 64 Chuang, K.V., Gunsalus, L.M., and Keiser, M.J. (2020). Learning molecular representations for medicinal chemistry. *J. Med. Chem.* 63 (16): 8705–8722.
- 65 Wigh, D.S., Goodman, J.M., and Lapkin, A.A. (2022). A review of molecular representation in the age of machine learning. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 12: e1603.
- 66 Polinsky, A. (2008). Chapter 12 – Lead-likeness and drug-likeness. In: *The Practice of Medicinal Chemistry*, Thirde (ed. C.G. Wermuth), 244–254. New York: Academic Press.
- 67 Lipinski, C.A. (2004). Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discov. Today Technol.* 1 (4): 337–341.
- 68 Warr, W. (2021). Report on an NIH workshop on ultralarge chemistry databases. *ChemRxiv*. Available from: <https://chemrxiv.org/engage/api-gateway/chemrxiv/assets/orp/resource/item/60c75883bdbb89984ea3ada5/original/report-on-an-nih-workshop-on-ultralarge-chemistry-databases.pdf>.
- 69 Lipinski, C. and Hopkins, A. (2004). Navigating chemical space for biology and medicine. *Nature* 432 (7019): 855–861.
- 70 Medina-Franco, J.L., Naveja, J.J., and López-López, E. (2019). Reaching for the bright StARs in chemical space. *Drug Discov. Today* 24 (11): 2162–2169.
- 71 Medina-Franco, J.L., Sánchez-Cruz, N., López-López, E., and Díaz-Eufracio, B.I. (2021). Progress on open chemoinformatic tools for expanding and exploring the chemical space. *J. Comput. Aided Mol. Des.* 36: 341–354.
- 72 Osolodkin, D.I., Radchenko, E.V., Orlov, A.A. et al. (2015). Progress in visual representations of chemical space. *Expert Opin. Drug Discovery* 10 (9): 959–973.
- 73 Saldívar-González, F.I. and Medina-Franco, J.L. (2022). Approaches for enhancing the analysis of chemical space for drug discovery. *Expert Opin. Drug Discovery* 17 (7): 789–798.
- 74 Wawer, M., Lounkine, E., Wassermann, A.M., and Bajorath, J. (2010). Data structures and computational tools for the extraction of SAR information from large compound sets. *Drug Discov. Today* 15 (15–16): 630–639.

- 75 Dunn, T.B., Seabra, G.M., Kim, T.D. et al. (2022). Diversity and chemical library networks of large data sets. *J. Chem. Inf. Model.* 62 (9): 2186–2201.
- 76 Everett, H. (1957). *Hugh Everett theory of the universal wavefunction*. Thesis. Princeton University.
- 77 Ren, X., Shi, Y.S., Zhang, Y. et al. (2018). Novel consensus docking strategy to improve ligand pose prediction. *J. Chem. Inf. Model.* 58 (8): 1662–1668.
- 78 Willett, P. (2013). Combination of similarity rankings using data fusion. *J. Chem. Inf. Model.* 53 (1): 1–10.
- 79 Medina-Franco, J.L., Maggiora, G.M., Giulianotti, M.A. et al. (2007). A similarity-based data-fusion approach to the visual characterization and comparison of compound databases. *Chem. Biol. Drug Des.* 70 (5): 393–412.
- 80 Medina-Franco, J.L., Martínez-Mayorga, K., Bender, A. et al. (2009). Characterization of activity landscapes using 2D and 3D similarity methods: consensus activity cliffs. *J. Chem. Inf. Model.* 49 (2): 477–491.
- 81 Naveja, J.J. and Medina-Franco, J.L. (2019). Finding constellations in chemical space through core analysis. *Front. Chem.* 7: 510.
- 82 Naveja, J.J. and Medina-Franco, J.L. (2020). Consistent cell-selective analog series as constellation luminaries in chemical space. *Mol. Inform.* 39 (12): e2000061.
- 83 López-López, E., Cerda-García-Rojas, C.M., and Medina-Franco, J.L. (2021). Tubulin inhibitors: a chemoinformatic analysis using cell-based data. *Molecules* 26 (9): 2483.
- 84 Muegge, I. and Oloff, S. (2006). Advances in virtual screening. *Drug Discov. Today Technol.* 3 (4): 405–411.
- 85 Schneider, G. (2010). Virtual screening: an endless staircase? *Nat. Rev. Drug Discov.* 9 (4): 273–276.
- 86 Zhao, H. (2007). Scaffold selection and scaffold hopping in lead generation: a medicinal chemistry perspective. *Drug Discov. Today* 12 (3–4): 149–155.
- 87 Sadybekov, A.A., Sadybekov, A.V., Liu, Y. et al. (2022). Synthon-based ligand discovery in virtual libraries of over 11 billion compounds. *Nature* 601 (7893): 452–459.
- 88 Liu, Z., Singh, S.B., Zheng, Y. et al. (2019). Discovery of potent inhibitors of 11 β -Hydroxysteroid dehydrogenase type 1 using a novel growth-based protocol of in silico screening and optimization in CONTOUR. *J. Chem. Inf. Model.* 59 (8): 3422–3436.
- 89 Amendola, G., Ettari, R., Previti, S. et al. (2021). Lead discovery of SARS-CoV-2 main protease inhibitors through covalent docking-based virtual screening. *J. Chem. Inf. Model.* 61 (4): 2062–2073.
- 90 Steadman, D., Atkinson, B.N., Zhao, Y. et al. (2022). Virtual screening directly identifies new fragment-sized inhibitors of carboxylesterase notum with Nanomolar activity. *J. Med. Chem.* 65 (1): 562–578.
- 91 Peng, Z., Zhao, Q., Tian, X. et al. (2022). Discovery of potent and isoform-selective histone deacetylase inhibitors using structure-based virtual screening and biological evaluation. *Mol. Inform.* e2100295.

- 92 Li, X., Jiang, Q., and Yang, X. (2022). Discovery of inhibitors for mycobacterium tuberculosis peptide deformylase based on virtual screening in silico. *Mol. Inform.* 41 (3): e2100002.
- 93 Naveja, J.J., Pilón-Jiménez, B.A., Bajorath, J., and Medina-Franco, J.L. (2019). A general approach for retrosynthetic molecular core analysis. *J. Cheminform.* 11 (1): 61.
- 94 Lewell, X.Q., Judd, D.B., Watson, S.P., and Hann, M.M. (1998). RECAP--retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* 38 (3): 511–522.
- 95 Wassermann, A.M., Dimova, D., Iyer, P., and Bajorath, J. (2012). Advances in computational medicinal chemistry: matched molecular pair analysis. *Drug Dev. Res.* 73 (8): 518–527.
- 96 Kunitomo, R., Dimova, D., and Bajorath, J. (2017). Application of a new scaffold concept for computational target deconvolution of chemical Cancer cell line screens. *ACS Omega* 2 (4): 1463–1468.
- 97 Hu, H. and Bajorath, J. (2020). Increasing the public activity cliff knowledge base with new categories of activity cliffs. *Future Sci. OA* 6 (5): FSO472.
- 98 Vogt, M., Yonchev, D., and Bajorath, J. (2018). Computational method to evaluate progress in lead optimization. *J. Med. Chem.* 61 (23): 10895–10900.
- 99 de la Vega de León, A. and Bajorath, J. (2014). Matched molecular pairs derived by retrosynthetic fragmentation. *Medchemcomm.* 5 (1): 64–67.
- 100 Dimova, D., Stumpfe, D., Hu, Y., and Bajorath, J. (2016). Analog series-based scaffolds: computational design and exploration of a new type of molecular scaffolds for medicinal chemistry. *Future Sci. OA.* 2 (4): FSO149.
- 101 Naveja, J.J., Vogt, M., Stumpfe, D. et al. (2019). Systematic extraction of analogue series from large compound collections using a new computational compound-core relationship method. *ACS Omega* 4 (1): 1027–1032.
- 102 Madariaga-Mazón, A., Naveja, J.J., Medina-Franco, J.L. et al. (2021). DiaNat-DB: a molecular database of antidiabetic compounds from medicinal plants. *RSC Adv.* 11 (9): 5172–5178.
- 103 Makarov, V., Salina, E., Reynolds, R.C. et al. (2020). Molecule property analyses of active compounds for mycobacterium tuberculosis. *J. Med. Chem.* 63 (17): 8917–8955.
- 104 Bobrowski, T.M., Korn, D.R., Muratov, E.N., and Tropsha, A. (2021). ZINC express: a virtual assistant for purchasing compounds annotated in the ZINC database. *J. Chem. Inf. Model.* 61 (3): 1033–1036.
- 105 Hartenfeller, M. and Schneider, G. (2011). Enabling future drug discovery by de novo design. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 1 (5): 742–759.
- 106 Schneider, G. and Clark, D.E. (2019). Automated de novo drug design: are we nearly there yet? *Angew. Chem. Int. Ed. Engl.* 58 (32): 10792–10803.
- 107 Huang, Q., Li, L.L., and Yang, S.Y. (2010). PhDD: a new pharmacophore-based de novo design method of drug-like molecules combined with assessment of synthetic accessibility. *J. Mol. Graph. Model.* 28 (8): 775–787.

- 108 Hartenfeller, M., Zettl, H., Walter, M. et al. (2012). DOGS: reaction-driven de novo design of bioactive compounds. *PLoS Comput. Biol.* 8 (2): e1002380.
- 109 Fischer, T., Gazzola, S., and Riedl, R. (2019). Approaching target selectivity by de novo drug design. *Expert Opin. Drug Discovery* 14 (8): 791–803.
- 110 Böhm, H.J. (1992). The computer program LUDI: a new method for the de novo design of enzyme inhibitors. *J. Comput. Aided Mol. Des.* 6 (1): 61–78.
- 111 Yuan, Y., Pei, J., and Lai, L. (2020). LigBuilder V3: a multi-target de novo drug design approach. *Front. Chem.* 8: 142.
- 112 Ertl, P. (2022). Magic rings: navigation in the ring chemical space guided by the bioactive rings. *J. Chem. Inf. Model.* 62 (9): 2164–2170.
- 113 Segler, M.H.S., Kogej, T., Tyrchan, C., and Waller, M.P. (2018). Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent. Sci.* 4 (1): 120–131.
- 114 Gantzer, P., Creton, B., and Nieto-Draghi, C. (2020). Inverse-QSPR for de novo design: a review. *Mol. Inform.* 39 (4): e1900087.
- 115 Mauri, A., and Bertola, M. (2023). AlvaBuilder: a software for de novo molecular design. *J. Chem. Inf. Model.* <https://doi.org/10.1021/acs.jcim.3c00610>.
- 116 Guianvarc’h, D. and Arimondo, P.B. (2014). Challenges in developing novel DNA methyltransferases inhibitors for cancer therapy. *Future Med. Chem.* 6 (11): 1237–1240.
- 117 González-Medina, M. and Medina-Franco, J.L. (2017). Platform for unified molecular analysis: PUMA. *J. Chem. Inf. Model.* 57 (8): 1735–1740.
- 118 Miranda-Quintana, R.A., Cruz-Rodes, R., Codorniu-Hernandez, E., and Batista-Leyva, A.J. (2010). Formal theory of the comparative relations: its application to the study of quantum similarity and dissimilarity measures and indices. *J. Math. Chem.* 47 (4): 1344–1365.
- 119 Johnson, M.A. and Maggiora, G.M. (1990). *Concepts and Applications of Molecular Similarity*. Wiley.
- 120 Bender, A. and Glen, R.C. (2004). Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem.* 2 (22): 3204–3218.
- 121 Schuffenhauer, A. and Brown, N. (2006). Chemical diversity and biological activity. *Drug Discov. Today Technol.* 3 (4): 387–395.
- 122 Eckert, H. and Bajorath, J. (2007). Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discov. Today* 12 (5–6): 225–233.
- 123 Koutsoukas, A., Paricharak, S., Galloway, W.R.J.D. et al. (2014). How diverse are diversity assessment methods? A comparative analysis and benchmarking of molecular descriptor space. *J. Chem. Inf. Model.* 54 (1): 230–242.
- 124 Bajorath, J. (2017). Representation and identification of activity cliffs. *Expert Opin. Drug Discovery* 12 (9): 879–883.
- 125 Martínez-Mayorga, K., Madariaga-Mazon, A., Medina-Franco, J.L., and Maggiora, G. (2020). The impact of chemoinformatics on drug discovery in the pharmaceutical industry. *Expert Opin. Drug Discovery* 15 (3): 293–306.

- 126 Bajusz, D., Miranda-Quintana, R.A., Rácz, A., and Héberger, K. (2021). Extended many-item similarity indices for sets of nucleotide and protein sequences. *Comput. Struct. Biotechnol. J.* 19: 3628–3639.
- 127 Rácz, A., Dunn, T.B., Bajusz, D. et al. (2022). Extended continuous similarity indices: theory and application for QSAR descriptor selection. *J. Comput. Aided Mol. Des.* 36 (3): 157–173.
- 128 Rácz, A., Mihalovits, L.M., Bajusz, D. et al. (2022). Molecular dynamics simulations and diversity selection by extended continuous similarity indices. *J. Chem. Inf. Model.* 62 (14): 3415–3425.
- 129 Miranda-Quintana, R.A., Kim, T.D., Heidar-Zadeh, F., and Ayers, P.W. (2019). On the impossibility of unambiguously selecting the best model for fitting data. *J. Math. Chem.* 57 (7): 1755–1769.
- 130 Miranda-Quintana, R.A., Bajusz, D., Rácz, A., and Héberger, K. (2021). Differential consistency analysis: which similarity measures can be applied in drug discovery? *Mol. Inform.* 40 (7): e2060017.
- 131 Maggiora, G.M. and Bajorath, J. (2014). Chemical space networks: a powerful new paradigm for the description of chemical space. *J. Comput. Aided Mol. Des.* 28 (8): 795–802.
- 132 Miljković, F. and Bajorath, J. (2020). Data structures for computational compound promiscuity analysis and exemplary applications to inhibitors of the human kinome. *J. Comput. Aided Mol. Des.* 34 (1): 1–10.
- 133 Gordon, D.E., Jang, G.M., Bouhaddou, M. et al. (2020). A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* 583 (7816): 459–468.
- 134 Chang, L., Perez, A., and Miranda-Quintana, R.A. (2021). Improving the analysis of biological ensembles through extended similarity measures. *Phys. Chem. Chem. Phys.* 24 (1): 444–451.
- 135 Rosenberg, A. and Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 410–420. Prague, Czech Republic: Association for Computational Linguistics.
- 136 Ashton, M., Barnard, J., Casset, F. et al. (2002). Identification of diverse database subsets using property-based and fragment-based molecular descriptions. *Quant struct-act relatsh.* 21 (6): 598–604.
- 137 Snarey, M., Terrett, N.K., Willett, P., and Wilton, D.J. (1997). Comparison of algorithms for dissimilarity-based compound selection. *J. Mol. Graph. Model.* 15 (6): 372–385.
- 138 Eppstein, D. (2004). Wang, fast approximation of centrality. *J. Graph. Algorithms Appl.* 8 (1): 39–45.
- 139 Flores-Padilla, E.A., Juárez-Mercado, K.E., Naveja, J.J. et al. (2021). Chemoinformatic characterization of synthetic screening libraries focused on epigenetic targets. *Mol. Inform.* e2100285.

17

SAR Knowledge Bases for Driving Drug Discovery

Nishanth Kandepedu, Anil Kumar Manchala, and Norman Azoulay

GOSTAR, Excelra Knowledge Solutions Pvt Ltd, Hyderabad, India

17.1 Introduction

Pharmaceutical researchers are working in an information-rich environment. Data on small molecules' activity and affinity toward biological targets is published in abundance, and medicinal chemistry data is becoming increasingly interconnected with data from the fields of bioinformatics and systems biology. A steady stream of publications containing useful information about novel compounds and their biological activities has resulted from decades of expansion in the pharmaceutical industry and academic drug discovery efforts [1]. Ten years ago, nearly 30 000 new compounds were getting published annually in some of the leading medicinal chemistry journals, with an annual growth rate of 4.4% [2, 3]. Current technological developments are accelerating the synthesis and testing of compounds, along with the emergence and expansion of the allied fields of chemical biology and chemogenomics. This has resulted in an exponential increase in the amount of data being published. Data in scientific publications and patents is captured in a way that makes it inaccessible to computational search and retrieval. As a result, the traditional publication paradigm has the potential to significantly impede the discoverability and utility of medicinal chemistry data.

The explosion of data necessitates the development of electronic means for capturing, querying, and extracting relevant data for further analysis and to gain valuable insights. The emphasis of this chapter is on databases that contain structure–activity relationship (SAR) data that are drawn from published sources, while we will also highlight some other potential sources of relevance. Furthermore, it is essential to acknowledge that the number of databases holding SAR information is expanding at a rapid pace, making it inevitable that even with a narrow focus, only a limited viewpoint of the databases and their coverage is presented.

This chapter is structured to first summarize the origin, evolution, and overview of relevant databases (at times referred to as knowledge base) that currently focus on SAR data of small molecules at a scale. It will be followed by a review of the various applications of these SAR databases in drug discovery and a note on their future direction.

17.2 The Origins of SAR Knowledge Bases

The origins of modern-day databases used in drug discovery can be traced back to the 1880s, when Beilstein published the first edition of “Handbuch der organischen Chemie” in two volumes in 1881 and 1883, with a total of 2200 pages and 15 000 compounds listed in it [4]. Even though there have been numerous revisions since then, the 1990s saw a critical turning point with the advent of computational advancement in chemistry, leading to the modernization of chemistry databases. In 1998, Frank Brown introduced the word chemoinformatics to describe the then-emerging field of computer applications in chemistry [5]. Chemoinformatics has taken over as the de facto standard for computer and informatics applications in chemistry since Johann Gasteiger’s Handbook of Chemoinformatics was published in 2003 [6]. Initial efforts in chemoinformatics were primarily focused on developing the corresponding database search engines and converting printed collections of chemical data, such as mass spectra and chemical literature, into electronic formats [7]. Eugene Markush revolutionized scientific intellectual property by publishing a patent in 1924, which also marked the beginning of the Markush structure, allowing an applicant to claim a multitude of chemical structures using a single generic structure [8]. The American Chemical Society’s Chemical Abstract Service (CAS) established a research and development department in 1955, setting the stage for creating computer-based chemical knowledge bases [9]. CAS launched the Markush structure database, MARPAT, in 1988 and SciFinder in 1995 as comprehensive chemical search engines [10]. The Derwent World Patents Index, formerly known as Farmdoc, comprises patent applications and grants from 44 different patent-granting bodies throughout the world [11]. The following are some key historical events that have led us to modern-day medicinal chemistry SAR databases (Figure 17.1).

Brown and Fraser proposed the concept of SAR in 1865 as a link between a molecule’s chemical structure and its biological activity. Drug discovery researchers use SAR data (at times referred to as bioactivity data) to identify the chemical

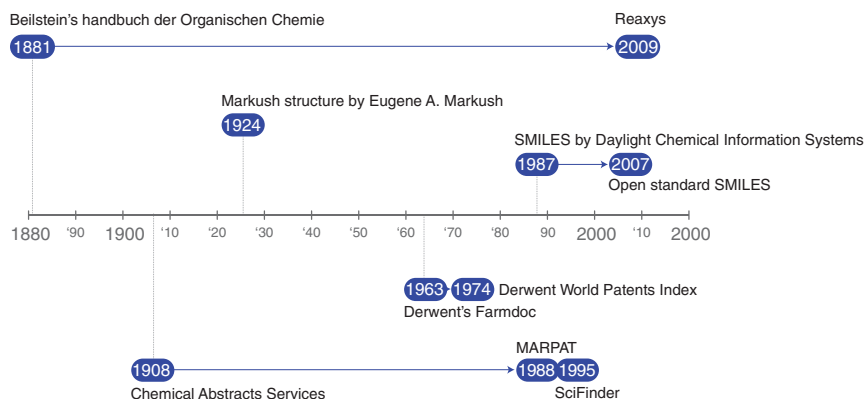


Figure 17.1 Evolution of chemical databases.

group responsible for biological effects in an organism [12]. Drug discovery in the twenty-first century relies heavily on databases with information on small molecule interactions with biological targets, drug metabolism, and pharmacokinetics. The recent increase in the number of SAR knowledge bases available to support drug discovery and development is evidence of this. There are many freely available chemical compound databases on the web, while others charge a fee. Some of these databases aim to index a specific subset of data, while others are comprehensive [13].

The binding database (BindingDB), launched at the University of Maryland in the late-1990s, is widely regarded as the first publicly available database and primarily collected quantitative protein-small molecule binding affinity data [14]. The National Institute of Health (NIH) launched the PubChem database in 2004, initially as a centralized repository for the NIH's Molecular Libraries Program (MLP). However, it now includes data contributed by several non-MLP organizations. For instance, PubChem now has a sizable amount of information on chemical compounds' bioactivity curated from several thousand scientific articles by data contributors like the binding database [15]. The ChEMBL database, launched in 2009, is another publicly available repository of bioactive molecules with drug-like properties. The database, known initially as StARlite, was created by Inpharmatica, a biotech company that Galapagos later acquired [16]. The Wellcome Trust grant helped the European Molecular Biology Laboratory (EMBL) collect the data in 2008, which led to the establishment of the ChEMBL chemogenomics group at the EMBL- European Bioinformatics Institute [17, 18].

Global Online Structure–Activity Relationship (GOSTAR) is a proprietary database introduced by GVK Biosciences in 2008 that provides comprehensive, homogenized, and detailed metadata around assay conditions that are typically missing from public databases [19]. Elsevier's proprietary database, Reaxys Medicinal Chemistry (RMC), was released in 2013 and can be used independently or in conjunction with their Reaxys chemistry database. RMC obtains the data from a vast collection of journal articles published by Elsevier and other sources (Figure 17.2) [20].

Alongside these knowledge bases, there are many other public and proprietary databases on the market. An in-depth walkthrough of some of these databases will be discussed in the next section of this review.

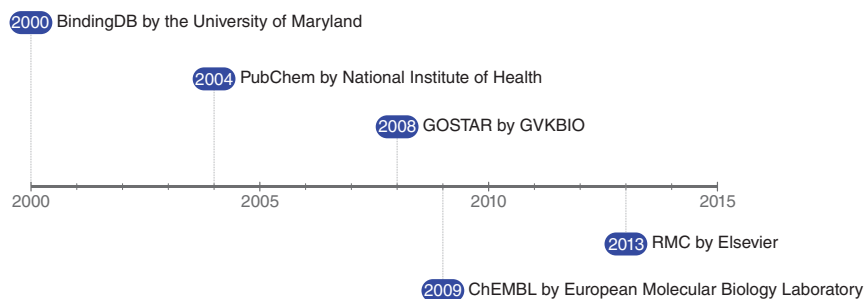


Figure 17.2 SAR knowledge bases in drug discovery.

17.3 SAR Knowledge Base Landscape

17.3.1 Open-Source SAR Databases

17.3.1.1 PubChem

The National Institute of Health (NIH) of the United States established PubChem (<http://www.pubchem.ncbi.nlm.nih.gov>) in 2004 as part of the NIH molecular libraries initiative, as one of the multiple repositories administered by the National Center of Biotechnology Information (NCBI) [21]. As of August 2022, PubChem has more than 282 million chemical substances, 112 million unique chemical structures, and 295 million bioactivity data stemming from over 1.4 million bioassays. Substance, compound, and bioassay are three interconnected databases that PubChem uses to manage this enormous quantity of data. More than 869 contributors, including pharmaceutical companies, academic laboratories, government organizations, chemical suppliers, publishers, and several other sources, supply data to PubChem [22]. The NIH Molecular Libraries Screening Centre Network (MLSCN), a consortium of 10 facilities that can do high throughput screening on large compound sets, is the most significant source of biological data for PubChem. The MLSCN's primary objective is to find small-molecule chemical probes for fundamental research; as a result, these screening libraries at times have compounds that medicinal chemists would deem nondrug-like [23]. In the last five years, the number of substances has increased from 236 million to 282 million, while the number of unique compounds has increased from 94 million to 112 million, representing a 19% growth [24]. Thieme Chemistry supplied around 700 000 compounds to PubChem in 2019, 42% of which were new to the database, and 89% lacked literature referencing, increasing the number of chemical structures with references from 1 to 1.6 million [25]. Such efforts significantly increase the findability, accessibility, interoperability, and reusability (FAIR) of chemical information pertaining to compounds. PubChem also provides compound-centric data on commercial availability, vendor information, safety, hazard, use, and manufacturing when available. Depending on the user's needs, PubChem data can be accessed in various ways, including using their web interface for small subsets of data and their web services and File Transfer Protocol (FTP) site for bulk downloads [26]. The Power User Gateway – Representational State Transfer (PUG-REST), which can be encoded into a one-line Uniform Resource Locator (URL), is the simplest programmatic method to request one or more PubChem records [27]. Additionally, it enables chemistry-based searches, which are inaccessible through other programmatic interfaces made available by PubChem [28]. PUG-View is another REST-style web service offering access to PubChem reports [29]. PUG-Simple Object Access Protocol (SOAP), PubChem-Resource Description Framework (RDF), PUG, and Entrez utilities are the other programmatic routes to access various PubChem fields [30]. It is, however, challenging for PubChem to meet the data demands of every user, given that it is a public resource accessed by millions of users with diverse needs. Additionally, handling heterogeneous data from multiple data sources from several scientific disciplines and preserving data quality are two other key areas where PubChem will need to continue to work to meet its users' demands.

17.3.1.2 ChEMBL

The ChEMBL database (www.ebi.ac.uk/chembl), originally developed by Inpharmatica, started as a collection of commercial products such as StARlite, CandiStore, and DrugStore [16]. The present-day ChEMBL database is maintained by the EBI-EMBL, based at the Wellcome Trust Genome Campus in the United Kingdom. As of August 2022, ChEMBL has more than 2.1 million unique chemical structures and 19 million bioactivity data stemming from over 1.5 million bioassays [31]. ChEMBL contains experimental readouts such as data from protein–ligand affinities, whole-cell-based assays, drug metabolism, pharmacokinetics, and toxicity that are manually curated and annotated from medicinal chemistry research publications. Approximately 40% of ChEMBL data is imported from PubChem, which includes PubChem bioassays and information on a compound's progression through the clinical stages [2]. ChEMBL data has been annotated with a variety of additional third-party identifiers, including target proteins from the Protein Data Bank (PDB), protein sequences from the UniProt Knowledge Base (UniProtKB), and chemical substances from PubChem, DrugBank, and ChemSpider, among others [2, 32, 33]. ChEMBL also serves as an open data-sharing hub for the field of neglected tropical disease research. The datasets, which contain thousands of compounds, are the outcome of chemical screening campaigns conducted by GlaxoSmithKline, Novartis Institute of Biomedical Research, Drugs for Neglected Diseases Initiative, and St. Jude Children's Research Hospital [34]. ChEMBL data can be accessed via a web interface, RDF platform, FTP site, and RESTful application programming interface (API) [35]. Web browsers, command-line programs that retrieve content from the web, or applications that use RESTful APIs, like Konstanz Information Miner (KNIME) and Slack, can all readily access the ChEMBL API [36].

The data curated from pharmacological patents in the ChEMBL database accounts for just over 2000 documents as of August 2022. Patents from the drug discovery and development stages are often considered a rich source of knowledge on novel chemotypes. It would take an average of four years for new chemicals to be published in the scientific literature and then annotated into a publicly accessible database [37, 38]. Even though SureChEMBL was introduced by ChEMBL in 2016 and offered free access to 17 million chemical structures from 14 million patents dating back to 1970, the database is filled with information on starting materials and intermediates with minimal pharmacological significance [39]. A considerable effort will be needed to narrow the gap between content extracted from scientific literature and patents, which is where ChEMBL will need to continue working to meet its users' demands.

17.3.1.3 DrugBank

The DrugBank (www.drugbank.com) was launched in 2006 in David Wishart's lab at the University of Alberta, in Canada, with information about 841 FDA-approved small molecule drugs, 113 biotech drugs, and 2133 drug targets [40]. As of August 2022, DrugBank had more than 2700 FDA-approved small molecule drugs, 6692 experimental drugs, and 271 withdrawn drugs spanning around 4900 unique targets, including enzymes, transporters, and carriers [41]. DrugBank has had several iterations since its inception, adding many data fields such as drug metabolism

and pharmacokinetics, pathways, food–drug interactions, drug–drug interactions, pricing, and other chemistry-centric information manually curated and annotated from several data sources. Like PubChem and ChEMBL, third-party identifiers are used wherever applicable for seamless integration with other databases. The DrugBank online is open to the public for specific noncommercial use cases, while using and redistributing of information from DrugBank data for any other purpose necessitates a license [42]. DrugBank data, like other databases, can be accessed via a web interface and a RESTful web API [43]. Although DrugBank offers richly annotated data, the small number of compounds covered restricts its broader applicability. DrugBank could gain comprehensive drug discovery and development applications by expanding its scope beyond drugs and drug candidates.

17.3.1.4 BindingDB

The binding database, often known as BindingDB (www.bindingdb.org), was launched in the late-1990s at the University of Maryland, in the United States, as an open-access database focused on protein-small molecule binding affinity data [2, 44]. As of August 2022, BindingDB is maintained by the Skaggs School of Pharmacy and Pharmaceutical Sciences and has more than 1 million small molecules with over 2.5 million binding data on 8800 protein targets; of those, 1.1 million data for 527 000 compounds and 4300 targets were curated by BindingDB [45]. In addition to integrating information from open-access databases like ChEMBL and PubChem, the BindingDB also regularly curates data from roughly a dozen scientific publications, particularly in the domains of chemical biology and biochemistry. BindingDB curators collect quantitative affinity data from the documents along with experimental assay conditions, such as pH, buffer, and temperature. BindingDB, in addition to a web interface, provides data access via a RESTful API, where a protein of interest or a SMILES string can be used to request data [46]. BindingDB has also been integrated with KNIME for data analysis and reporting [47].

BindingDB uses automatic and manual curation techniques to excerpt only the readouts that include a well-defined protein target and a quantitative measure of affinity or relative affinity, often an inhibitory concentration, inhibition constant, or dissociation constant value [48]. For example, whole-cell-based assay data from ChEMBL and single-concentration HTS data from PubChem are not imported into BindingDB since they are not deemed to have confirmative binding affinities toward a single protein target. For some users, this may be considered a limitation because the data is mostly drawn from chemical biology-focused papers rather than information-rich medicinal chemistry literature. With the resurgence of phenotypic drug discovery approaches in identifying a successful drug, as well as the use of absorption, distribution, metabolism, and excretion data in machine learning (ML) models, it is critical for BindingDB to expand its coverage of activity endpoints in future offerings.

17.3.1.5 Other Known Open-Source Databases

The Psychoactive Drug Screening Program launched Ki (<http://pdsp.med.unc.edu>), an open-access database hosted by the University of North Carolina, in the United

States, with 97 606 binding affinities for thousands of compounds on more than 700 receptors, ion channels, neurotransmitter transporters, and enzymes [49]. The Human Metabolome Database (HMDB) (www.hmdb.ca) is a publicly available repository of small molecule metabolites from humans with over 220 945 metabolite entries, including both hydro- and lipophilic metabolites. The HMDB also has over 8000 protein sequences linked to these metabolite entries [50]. The HMDB suite of databases also includes four other databases: DrugBank, Therapeutic Target Database, Small Molecule Pathway Database, and Food Database. ZINC database (<http://zinc.docking.org>), based at the University of California, San Francisco, in the United States, is an open-access database of over 230 million commercially available compounds [51]. Compounds are divided into categories such as target-focused, natural products, metabolites, lead-like, and fragment-like, and their availability is annotated. Compounds are available in standard molecular docking formats to enable virtual structure-based screening with precomputed three-dimensional conformations. The Protein Kinase Inhibitor Database (PKIDB) is a curated repository of over 320 clinical compounds that are inhibitors of various protein kinases [52]. The Kinase–Ligand Interaction Fingerprints and Structure (KLIFS) database is another kinase knowledge base (KKB) with more than 3300 small molecule inhibitors spanning over 307 unique kinases, capturing information on kinase–ligand interactions and binding affinities [53]. Proteolysis-targeting chimera database (PROTAC-DB) is the first and publicly available SAR knowledge base with over 3270 PROTACs, which are heterobifunctional small molecules capable of degrading protein targets of interest [54].

17.3.2 Commercial SAR Databases

The information available on proprietary databases is sparse compared to open-access bioactivity databases and has not been extensively investigated or documented in the scientific literature. This is understandable for several reasons, including the need to preserve the confidentiality of the data fields that these databases collect during the extraction process and the dynamic nature of these databases to update the information based on customers' needs. This also makes it challenging to present a guidance point to readers on commercial databases. However, we gathered information from openly accessible resources on a few proprietary SAR knowledge bases.

17.3.2.1 GOSTAR

The GOSTAR database (www.gostardb.com), launched by GVK Biosciences (currently known as Aragen Life Sciences) in 2008, is regarded as one of the few comprehensive databases to include manually curated bioactivity data from both scientific publications and patents [55]. Excelra Knowledge Solutions, formerly known as GVK Informatics, has maintained the GOSTAR database since 2014. As of August 2022, GOSTAR has more than 9 million small molecules and over 30 million bioactivity data, including 10 million binding data on 82 000 protein targets stemming from several biological sources (Figure 17.3) [56].

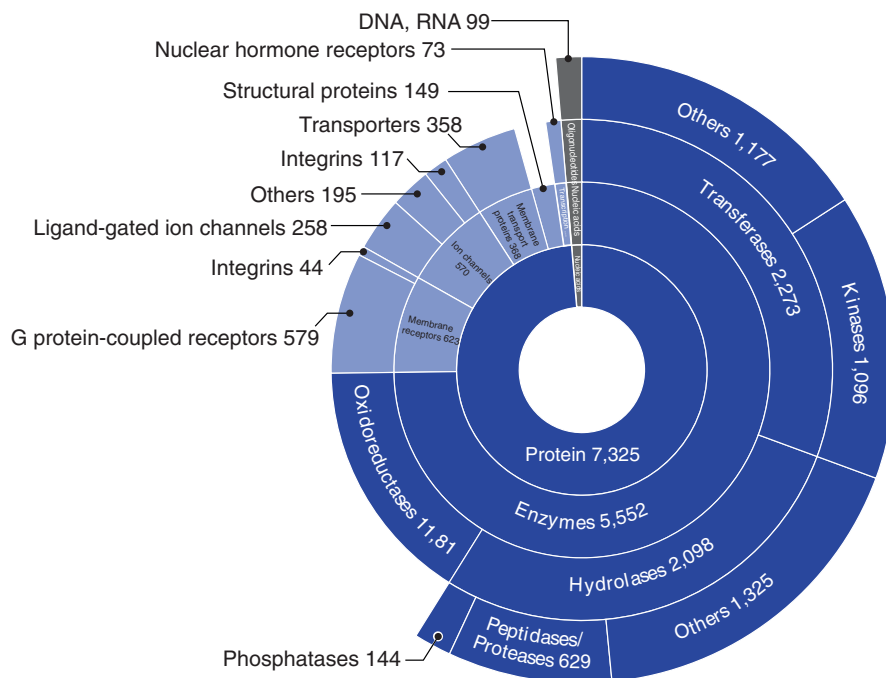


Figure 17.3 GOSTAR target coverage with the total number of targets in each family.

GOSTAR is the only SAR knowledge base in this review that uses an ISO-certified end-to-end manual curation approach. GOSTAR data is manually extracted from 204 000 journal articles and 87 000 patents without integrating data from other databases, preserving dataset quality and homogeneity. GOSTAR contains quantitative experimental data on protein–ligand affinities, whole-cell-based assays, drug metabolism, pharmacokinetics, and toxicity from patents and established medicinal chemistry journals, from their first editions through 2022 [57]. A snapshot of assay protocols, including experimental conditions such as pH, buffer, temperature, radioligands, and substrate information, is captured within GOSTAR. GOSTAR Intelligence platform is an intuitive search engine that can run simple searches such as target, compound, and bibliography-based searches as well as complex searches where an end user can combine two or more search parameters to obtain information on a specific query. It also includes a foray into analyzers such as matched molecular pair (MMP) analysis, drug–target interaction heatmaps, and property analyzers to provide valuable knowledge-based insights for medicinal chemists [58]. In addition, GOSTAR data can also be accessed via API and as a downloadable dataset [59]. The data within GOSTAR is annotated with several third-party identifiers, including target proteins from the PDB, protein sequences from the UniProtKB, and activity endpoints from the BioAssay Ontology (BAO), among others [60].

17.3.2.2 Reaxys Medicinal Chemistry

RMC (www.reaxys.com) is another proprietary SAR knowledge base launched in 2013 by Elsevier. As of August 2022, RMC has 34 million compounds and

41 million bioactivity data extracted from 540 000 documents. RMC is also a comprehensive database covering data on target-compound affinities, functional assays, absorption, distribution, metabolism, excretion, and toxicity, curated from a collection of more than 5000 drug discovery and medicinal chemistry journals [61]. In addition to curating content from primary data sources, Reaxys also integrates content from third-party databases. To maintain the quality and homogeneity of the data, RMC standardizes the chemistry and bioactivity data when integrating it from other databases and eliminates conflicting data before merging [62]. The RMC's query-building mechanisms allow a medicinal chemist to carry out knowledge-based drug design using their online platform's 19 dedicated query forms for primary lead optimization searches [63]. The RMC API enables users to access and integrate data with other applications such as KNIME and Pipeline Pilot [64]. The data within RMC is annotated with several third-party identifiers, like other databases.

17.3.2.3 Other Known Commercial Databases

In 2019, Clarivate Analytics introduced Cortellis Drug Discovery Intelligence (CDDI) (www.cortellis.com) to reinvent its Integrity platform as a competitive intelligence tool [65]. As of August 2022, CDDI has more than 600 000 small-molecule drugs and biologics and 3 million bioactivities extracted from 2.9 million publication references [66]. The KKB was launched in 2005 as a result of a merger between Eidogen and Sertanty as a solution to leverage knowledge-driven drug discovery. KKB, as the name implies, is a kinase-specific SAR database, making it the only non-comprehensive database on the list of commercial databases within this chapter. As of August 2022, KKB has more than 400 000 small-molecule drugs and 2.2 million bioactivities on 560 unique kinase targets extracted from over 9000 references [67].

The reader is referred elsewhere for comprehensive reviews of other databases of relevance [68–72].

17.4 Comparison and Complementarity of SAR Databases

There are three essential criteria users will consider while selecting SAR knowledge bases: quantity, quality, and overlap of content across multiple public and commercial databases. Public databases have remained relatively modest, focusing on specific subsets of small molecules, targets, activities, document types, etc., despite the rise in public domain data in recent years. On the other hand, commercial SAR databases are more comprehensive and large-scale, with a fee-based business model to cover the costs of curation and maintenance.

The number of compounds, bioactivities, and documents extracted is typically a simple metric used for quantifying and comparing SAR knowledge bases. Nevertheless, there are essential limits when comparing these statistics due to the slightly varying techniques by which they are connected in these databases. For instance, the PubChem statistics for bioactivities factor substances but not compounds, inflating

Table 17.1 Latest statistics on various public and proprietary SAR knowledge bases.

Database	PubChem	RMC	GOSTAR	ChEMBL	CDDI	Kinase KB
Compounds	112M	34M	9M	2.2M	0.64M	0.41M
Bioactivities	295M	41M	30M	19M	3M	2.2M
Scientific literature	34M	0.40M	0.20M	0.08M	2.9M	0.01M
Patents	42M	0.14M	0.09M	0.01M		

the number, whereas ChEMBL, the main contributor to PubChem bioassays, minimizes its assay counts to compounds [73]. The vendor dilution effect generated by the addition of commercially accessible compounds from chemical suppliers that do not have associated bioactivities makes comparing the number of compounds in databases challenging at times [74].

The compounds or bioactivities per document cannot be used as an absolute criterion to evaluate any database in general. However, since most SAR databases depend on the same set of medicinal chemistry publications and patents for their content, comparing them can provide insight into the disparity in the content extraction approaches. PubChem has 112 million compounds in total, with an average of 2 compounds and 12 bioactivities per 3 documents (Table 17.1). When PubChem figures are compared to those obtained for other databases, it is pronounced that many of the compounds and bioactivities reported in PubChem could have been deposited by data vendors and not published in scientific literature or patents. Since this chemical and biological data is unique to PubChem, it would give significant value to many drug discovery applications, particularly in academic research. However, given that some compounds and bioactivities were not peer-reviewed and published, some use cases, such as training ML models to predict various parameters of interest, may be challenging as there is no available data provenance or assay definition for reproducibility.

Comparing content statistics among commercial databases is also challenging since the figures are compiled using proprietary identifiers for structures, bioactivities, documents, and somewhat different procedures. The GOSTAR numbers and details around their extraction process have been discussed several times in scientific literature and on Excelra's website [56, 57, 75]. GOSTAR's developers claim that all the content is manually curated from data sources, whereas RMC claims to use a combination of automation and manual excerption [60, 62]. Even though the number of compounds in RMC has increased from 6.8 million in 2019 to 34 million in 2021, the amount of bioactivity data has only increased from 35 million to 41 million [61, 73]. This suggests an increase of a staggering 27 million compounds and only 6 million bioactivities, with an average of 1 bioactivity data for every 5 compounds added between 2019 and 2021 into RMC. It is unclear which procedural variations may account for this. The KKB has been one of the gold standards for databases for information on compounds acting on kinases. The KKB has 2.2 million bioactivity data from about 9000 documents. An average of 226

bioactivity data from a document mined shows the depth of information collected from a relatively small corpus of data sources (Figure 17.4) [67].

The content updates to the database would be another critical aspect while comparing SAR knowledge bases. Commercial databases will have an advantage in this case since they continuously add content by maintaining a large team of curators, as they serve a vast client base with a persistent need for new and diverse data. This is supported by the finding that GOSTAR added four times as many compounds in the last five years compared to ChEMBL (Figure 17.5) [76]. However, it is commendable that, despite limited resources, public databases have added large sets of bioactivity data in recent years (Figure 17.6).

There have been several attempts to demonstrate how complementary public and commercial databases are and the benefits of merging the two. Christopher Southan

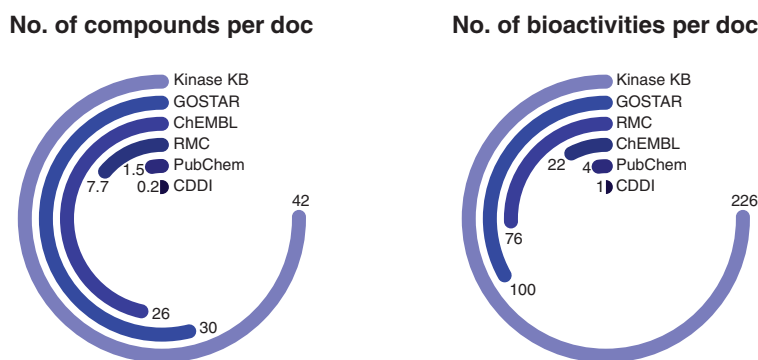


Figure 17.4 Average number of compounds and bioactivities per document in various SAR knowledge bases.

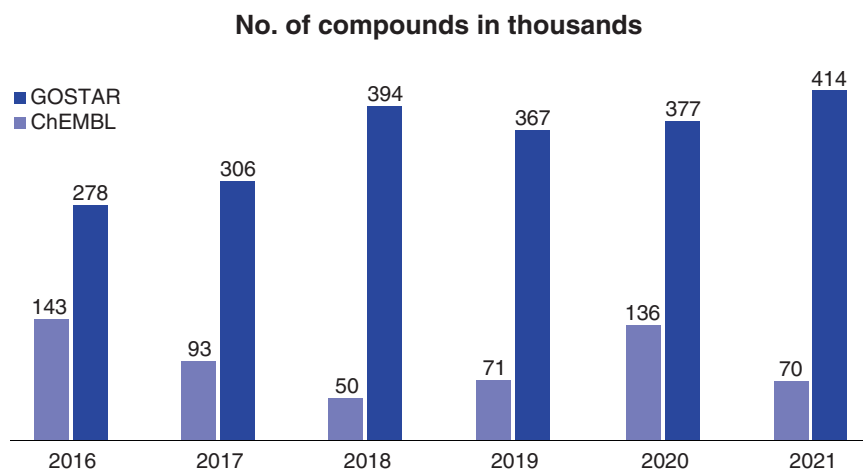


Figure 17.5 Compound growth over the last five years between ChEMBL and GOSTAR.

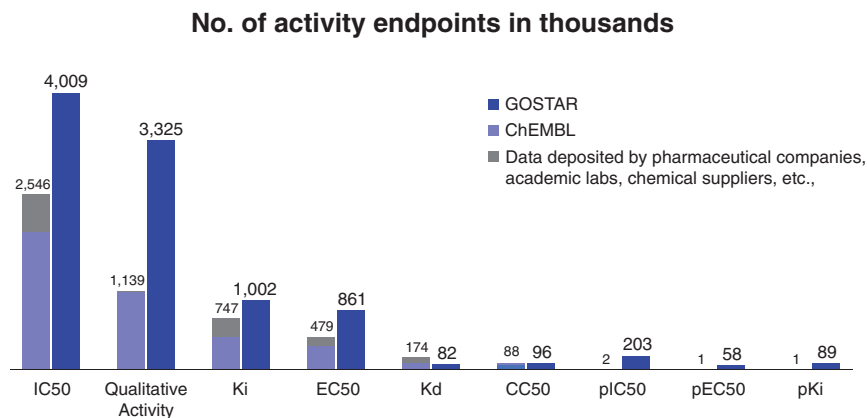


Figure 17.6 Comparison of bioactivity endpoints between ChEMBL and GOSTAR.

and colleagues did notable work in this field, investigating the overlap across several medicinal chemistry databases and demonstrating the uniqueness in some instances. In 2007, Southan et al. published a study on a three-way comparison of the PubChem, GOSTAR (previously known as GVKBIO database), and World of Molecular Bioactivity (WOMBAT) databases. The structural overlap among these three databases was around 86 000, with PubChem having 6.8 million distinct structures and GOSTAR having more than 1 million [77]. This is to be expected, given that PubChem contains compounds contributed by depositors and GOSTAR has substantial expertise in patent curation. Their recent study in 2020 demonstrates that documents, chemical structures, and protein targets overlap among three open-source databases: ChEMBL, BindingDB, and Guide to Pharmacology (GtoPdb). When ChEMBL and BindingDB are compared, there is an overlap of around 25 000 documents, 600 000 structures, and 3000 protein targets out of 73 000 documents, 2 million chemical structures, and 9000 targets [73]. Laura Isigkeit and colleagues used compounds and bioactivities from ChEMBL, PubChem, BindingDB, GtoPdb, and Probes & Drugs to build a consensus SAR dataset of 1.1 million compounds and 11 million bioactivities on 5600 targets. The dataset analysis revealed that it includes about 455 000 out of 1.1 million compounds in more than one database, with just 600 compounds in all databases. Around 987 000 bioactivity data had an exact match with bioactivity annotation out of 1.3 million bioactivity data obtained from many databases [78]. This study demonstrates the value of complementing several SAR databases for data-driven drug discovery applications since it broadens the coverage of compounds and targets as well as scaffold diversity. Other studies in the literature have similarly shown complementarity by combining in-house data from large pharmaceutical companies with information from public and commercial sources [68]. It is worth noting that the main objective of the databases in this review is to extract and disseminate data on the bioactivities of small molecules that are primarily published in the scientific literature focusing on drug discovery. As a result, the compounds in these databases are thus expected to have drug-like molecular properties.

Given that data quality is of such importance to users, it is essential to understand what errors may be found in the data, and how they may be caused. Scientific, transcription, and handling errors are the three types of data errors. While scientific errors arise due to experimental or technical faults, transcription errors occur during the publication write-up process. Data handling errors happen during data extraction. They are the consequence of a curator's scientific misinterpretation, inaccurate data entry, or occasionally a mismatch between the structures and bioactivity tables. Several studies have been published in the literature to assess data quality and quantify data errors in various SAR databases. William and colleagues conducted such a study on a chemistry database and were surprised to find 70% errors in absolute structural integrity when they expected just 10% [79]. Oprea et al. discovered that errors from one database are migrated to another during data integration [80]. A comprehensive examination of SAR databases would be beneficial for identifying and correcting data errors. However, this would be difficult given the volume of data in some of these databases. While the database owners cannot address scientific and transcription errors, data handling can be improved. The target information can occasionally be at the level of a protein subfamily, and there is no assurance that the small molecule will bind to the same protein listed in the activity information. To solve the issue, ChEMBL has created a target confidence score ranging from zero for data entries that have not yet been curated, to nine for a single protein target that has been given a high confidence level. A confidence value of one is assigned to assays with nonmolecular targets, whereas a confidence level of at least four is given to tests with protein targets [81]. Extracting unstructured data from documents, particularly patents, is often considered the most challenging step in bioactivity data curation and a root cause of data errors in knowledge bases. Manual curation would be more accurate and effective if compounds, bioassays, and their endpoints were published in a standardized way across the industry. The Minimum Information About a Bioactive Entity (MIABE) initiative is one such effort. It provides a framework for authors to create and publish structured and annotated data to support the data-driven discovery of bioactive molecules [82].

An extremely efficient but expensive method to reduce the curation error rate would be to have two or more curators extract data independently from a particular article and then cross-check the findings. GOSTAR's developers have made a product USP of this approach, subjecting all data to a three-tiered quality control process. Only data that matches the source document is included in the GOSTAR database [56, 57, 75]. KKB follows a similar procedure, with manual interventions at several levels to curate and quality-check the information. Each data entry is stamped as QC passed before being distributed to the end user [83].

17.5 Applications of SAR Knowledge Base in Modern Drug Discovery

There were few documented applications for knowledge bases in drug discovery until recently, and even those that could be identified – such as using the SAR

databases as an easily accessible source of compound and bioactivity information – would have been regarded as a little obvious. However, the time and effort saved while searching for and retrieving information should not be understated. For example, patent landscape analysis, made possible by databases like DWPI and MARPAT, is crucial for pharmaceutical companies to monitor patent applications, prevent duplication of work, and limit time spent on research. Finding a lead compound from a group of hits identified during primary screening often takes less than a year in today's fast-paced drug discovery paradigm. As a result, chemistry and associated SAR databases are an essential knowledge source for medicinal chemists seeking prior art in the chemical and biological space. Chemistry databases such as SciFinder and Reaxys can provide information on chemical abstracts for related compounds, whereas SAR knowledge bases such as GOSTAR, ChEMBL, RMC, and others can provide information on the associated bioactivities of these compounds, saving medicinal chemists' time. Besides prior art, SAR databases provide knowledge-based design for rational drug discovery [84]. Regulatory bodies approved 59 new drugs in the US, EU, and Japan in 2021, 50 of which have well-established mechanism-of-action targets for their approved indications, meaning most of them are documented and well-known to the scientific community [85]. As a result, SAR knowledge bases play an important role in modern-day data-driven drug discovery.

Knowledge-based compound design is straightforward and can be done manually for small subsets of compounds based on documented bioactivity data. However, with millions of compounds and associated bioactivity data in the literature, it is challenging and time-consuming to visualize trends and derive conclusions. One method in the medicinal chemist's toolbox for accomplishing this is the MMP analysis, which analyses similar chemical structures pairwise over a large dataset [86]. It is easier to understand changes in any physical or biological properties between MMPs and to develop and test a hypothesis since they have minor structural differences. Several studies using the MMP concept to analyze potency, drug metabolism, and pharmacokinetic parameters have been published, with the only difference being the MMP algorithm used [69, 87–90]. GOSTAR intelligence platform implements an intuitive analyzer view to do MMP analysis for compounds and targets of interest [91]. Wawer and Bajorath introduced the concept of matched molecular series (MMS) analysis, which broadens the definition of MMP analysis to include a series of analogs with identical scaffolds and one positional difference [92]. Any two compounds within an MMS are, by definition, MMPs of one another. The MMS concept has often been applied in the literature for SAR transfer between series, compound design, and bioactivity predictions [93–102]. Another valuable tool for knowledge-based design is the drug–target interaction heatmap, which provides insight into the off-target activity of structurally comparable compounds and, in some cases, can be used for drug repurposing. Heatmaps of structurally identical compounds and their reported activity against an array of targets can be visualized by end users using GOSTAR and RMC [61, 103]. Integrating such technologies into SAR knowledge bases can be of great help to researchers

since they can run a query, obtain data, and gain valuable insights for designing the next set of drugs.

The most widely reported application of SAR knowledge bases is in ML modeling. With the rise in accessibility of cutting-edge graphics processing units (GPUs) and cloud computing, which can speed up the processing of complex computations, as well as the success of ML models like deep learning, AI has expanded from essentially theoretical to practical applications [104–106]. A recent analysis of 21 000 compounds from phase I clinical trial to drug approval shows that the overall success rate stands at 6% [107]. The pharmaceutical sector has been compelled to adapt to nontraditional drug discovery technologies like ML to decrease total attrition and increase cost-effectiveness. ML has advanced end-to-end drug discovery and development applications. There have been reports on identifying new targets associated with diseases, disease pathophysiology, optimization of small-molecule lead compounds, drug efficacy, adverse drug reactions, and biomarker development [108–112]. Readers are directed to other extensive reviews on applications of ML in drug discovery and development [113–116].

The ML practice is widely believed to be split 80% on data processing and cleaning and 20% on building the algorithm, highlighting the need for annotated, large volumes, and high-quality data to make the most out of the model [114]. This section highlights numerous ML algorithms that have been used for drug discovery applications in the literature. While deep learning techniques appear promising, no single model or descriptor set has gained widespread acceptance. Target prediction with machine-learning algorithms can help accelerate the search for a new pharmacological target, limiting the number of required experiments. Deep learning outperformed all other tested methods, such as Random Forest, Support Vector Machine, K-Nearest Neighbors, Similarity Ensemble Approach, and Naive Bayes for target predictions, according to a study using various ML methods on a dataset of 45 000 compounds contained in more than 1000 assays extracted from ChEMBL [117]. The dataset encompassed a wide range of target families, and several sorts of fingerprints were used, which prevented it from being skewed by specific chemical structures or a particular structure representation of the compounds. This investigation revealed that the prediction model's performance improves with the size of the training set, proving the importance of developing large datasets for ML approaches. Another recent study used 290 000 structurally diverse compounds collected from GOSTAR, ChEMBL, PubChem, and hERGCentral to build an hERG classification model to predict potential cardiotoxicity. With an accuracy of 0.984 for the test set, the SVM classification model significantly outperformed the performance of the available commercial hERG prediction software [118]. The study showed that models created using diverse chemical space data from multiple SAR knowledge bases enable the creation of a more accurate classification model with a broader applicability domain. A further study used random forest classifier models to predict cytochrome P450 enzyme inhibitors demonstrated significant levels of robustness, as proven by good predictivity even for structurally different compounds from the training data [119]. This study used a combined dataset of 18 815 compounds from

the PubChem, ChEMBL, and ADME databases to train the model, and obtained an area under the receiver operating characteristic curve value for test compounds ranging from 0.89 to 0.92, depending on the CYP isozyme, demonstrating the value of combining data from different sources in achieving a better balance between two activity classes. The applications of ML in different stages of drug development using SAR knowledge bases are increasing constantly, and we can only provide a few examples from the literature here.

17.6 Future Direction

As drug discovery continues to harness the potential of data-driven technologies to find a drug, the presence of SAR knowledge bases in public and private realms will continue to grow. Even though the availability of numerous open-access databases provides diversity in data, search capabilities, and user interface choices, there is a pressing need for cooperation to reduce duplication of effort and enhance user value. PubChem, ChEMBL, and BindingDB are increasingly working on sharing data and curation efforts. Given the limited resources available to these initiatives, collaboration may be the only way to improve overall data quality. While proprietary knowledge bases are not immune to data errors, significant efforts are made to reduce the mistakes by engaging independent curators to review the extracted data, which is a more reliable but time-consuming technique. However, there is a need for more transparent and common industry standard practices to be followed by data providers to further improve data quality. While an increase in the publication of bioactivity data on novel compounds is more exciting than ever from a scientific standpoint, it also means that data extraction and management require enormous effort. With the advancements in artificial intelligence technologies such as natural language processing and automation, the quality and quantity of content curation will continue to improve while the extraction times will shorten. The saying the whole is greater than the sum of its parts holds true in the case of SAR knowledge bases, as with many of the applications discussed in this review. In conclusion, collaborative efforts and cutting-edge technology will likely reinforce the significance of SAR knowledge bases in drug discovery.

Acknowledgment

We acknowledge the following employees of Excelra Knowledge Solutions for helping us with details that made it much easier to analyze certain aspects of this book chapter: Renganathan Nageswaran and Ben Frew. We want to thank Vani Gudimella Tirumala, Krishna Reddy Madadi, and Navakanth Suram for providing the details on content updates and statistics for the GOSTAR database. The authors thank Hinrich W. Göhlmann of Janssen Research & Development and Eric Michael Gifford for their valuable feedback and discussions in structuring the chapter.

List of Abbreviations

ADME	absorption, distribution, metabolism, and excretion
AI	artificial intelligence
API	application programming interface
BAO	BioAssay ontology
CAS	Chemical Abstract Service
CDDI	Cortellis Drug Discovery Intelligence
CYP	cytochrome P450
DWPI	Derwent World Patents Index
EBI	European Bioinformatics Institute
EMBL	European Molecular Biology Laboratory
EU	European Union
FAIR	findability, accessibility, interoperability, and reusability
FDA	Food and Drug Administration
FTP	file transfer protocol
GOSTAR	global online structure–activity relationship
GPUs	graphics processing units
GtoPdb	Guide to Pharmacology
hERG	human ether-a-go-go-related gene
HMDB	Human Metabolome Database
HTS	high-throughput screening
ISO	International Organization for Standardization
K _i	inhibitory constant
KKB	Kinase Knowledge Base
KLIFS	kinase–ligand interaction fingerprints and structure
KNIME	Konstanz Information Miner
MIABE	minimum information about a bioactive entity
ML	machine learning
MLP	Molecular Libraries Program
MLSCN	Molecular Libraries Screening Centre Network
MMP	matched molecular pair
NICB	National Center of Biotechnology Information
NIH	National Institute of Health
PDB	Protein Data Bank
pH	potential of hydrogen
PKIDB	Protein Kinase Inhibitor Database
PROTAC-DB	proteolysis-targeting chimera Database
PUG	Power User Gateway
QC	quality control
RDF	Resource Description Framework
REST	representational state transfer
RMC	Reaxys Medicinal Chemistry
SAR	structure–activity relationship
SOAP	Simple Object Access Protocol

SVM	support vector machine
UniProtKB	UniProt Knowledge Base
URL	Uniform Resource Locator
US	United States of America
USP	unique selling point
WOMBAT	World of molecular bioactivity
ZINC	ZINC is not commercial

Disclaimer

The employees of Excelra Knowledge Solutions, the firm that owns the GOSTAR database, used their scientific expertise to write this book chapter. We are neither a publisher nor advocating the use of a specific product.

References

- 1 Portoghese, P.S. (2011). My farewell to the journal of medicinal chemistry. *J. Med. Chem.* 54: 8235.
- 2 Nicola, G., Liu, T., and Gilson, M.K. (2012). Public domain databases for medicinal chemistry. *J. Med. Chem.* 55 (16): 6987–7002.
- 3 Llanos, E.J., Leal, W., Luu, D.H. et al. (2019). Exploration of the chemical space and its three historical regimes. *Proc. Natl. Acad. Sci. U.S.A.* 116 (26): 12660–12665.
- 4 Luckenbach, R. (1981). The Beilstein handbook of organic chemistry: the first hundred years. *J. Chem. Inf. Comput. Sci.* 21 (2): 82–83.
- 5 Brown, F. (1998). Chemoinformatics: what is it and how does it impact Drug Discovery. *Annu. Rep. Med. Chem.* 33: 375–384.
- 6 Gasteiger, J. (ed.) (2003). *Handbook of Chemoinformatics - from Data to Knowledge*. Germany: Wiley-VCH. Weinheim.
- 7 Zemany, P.D. (1950). Punched card catalog of mass spectra useful in qualitative analysis. *Anal. Chem.* 22: 920–922.
- 8 Eugene, A. M. Pyrazolone dye and process of making the same. US1506316A, August 26, 1924.
- 9 Fisanick, W., Amaral, N.J., Metanomski, W.V. et al. (1998). Chemical abstract service information system. In: *The Encyclopedia of Computational Chemistry* (ed. P.V.R. Schleyer, N.L. Allinger, T. Clark, et al.), 277–315. Chichester: J. Wiley & Sons.
- 10 Chen, W.L. (2006). Chemoinformatics: past, present, and future. *J. Chem. Inf. Model.* 46 (6): 2230–2255.
- 11 Ozcan, S. and Islam, N. (2017). Patent information retrieval: approaching a method and analysing nanotechnology patent collaborations. *Scientometrics.* 111: 941–970.

- 12 Brown, A.C. and Fraser, T.R. (1865). The connection of chemical constitution and physiological action. *Trans. R. Soc. Edinb.* 25: 1968–1969.
- 13 Ekins, S., Clark, A.M., Swamidass, S.J. et al. (2014). Bigger data, collaborative tools and the future of predictive drug discovery. *J. Comput. Aided. Mol. Des.* 28: 997–1008.
- 14 Chen, X., Liu, M., and Gilson, M.K. (2001). Binding DB: a web-accessible molecular recognition database. *Combi. Chem. High-Throughput Screen* 4: 719–725.
- 15 Kim, S., Thiessen, P.A., Bolton, E.E. et al. (2016). PubChem substance and compound databases. *Nucleic Acids Res.* 44 (D1): D1202–D1213.
- 16 Warr, W.A.C.E.M.B.L. (2009). An interview with John Overington, team leader, chemogenomics at the European bioinformatics institute outstation of the European molecular biology laboratory (EMBL-EBI). *J. Comput. Aided. Mol. Des.* 23 (4): 195–198.
- 17 Gaulton, A., Bellis, L.J., Bento, A.P. et al. (2012). ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 40 (D1): D1100–D1107.
- 18 Open access drug discovery database launches with half a million compounds. <http://wellcome.ac.uk>. 18 January 2010. Retrieved 27 July 2022.
- 19 Southan, C., Boppana, K., Jagarlapudi, S.A., and Muresan, S. (2011). Analysis of in vitro bioactivity data extracted from drug discovery literature and patents: ranking 1654 human protein targets by assayed compounds and molecular scaffolds. *J. Cheminform.* 3 (14): 1–11.
- 20 Elsevier launches Reaxys Medicinal Chemistry as part of its suite of life science solutions. <http://stm-publishing.com>. 5 February 2013. Retrieved 27 July 2022.
- 21 Wang, Y., Xiao, J., Suzek, O.T. et al. (2009). PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* 37: W623–W633.
- 22 PubChem Data Counts. <http://pubchemdocs.ncbi.nlm.nih.gov>. Retrieved 29 August 2022.
- 23 Huryn, D. M.; Cosford, N. D. P. The Molecular Libraries Screening Center Network (MLSCN): identifying chemical probes of biological systems. Macor J. E. (Ed.); *Annual Reports in Medicinal Chemistry*. 2007, 42, pp. 401–416.
- 24 Southan, C. (2018). Caveat Usor: assessing differences between major chemistry databases. *ChemMedChem*. 13 (6): 470–481.
- 25 More than a million chemical-article links from Thieme Chemistry added into PubChem. <http://pubchemdocs.ncbi.nlm.nih.gov>. 15 January 2019. Retrieved 29 August 2022.
- 26 Kim, S., Thiessen, P.A., Bolton, E.E., and Bryant, S.H. (2015). PUG-SOAP and PUG-REST: web services for programmatic access to chemical information in PubChem. *Nucleic Acids Res.* 43: W605–W611.
- 27 Kim, S., Thiessen, P.A., Cheng, T. et al. (2018). An update on PUG-REST: RESTful interface for programmatic access to PubChem. *Nucleic Acids Res.* 46: W563–W570.
- 28 Kim, S. (2016). Getting the most out of PubChem for virtual screening. *Expert Opin. Drug Discov.* 11 (9): 843–855.

- 29 Kim, S., Thiessen, P.A., Cheng, T. et al. (2019). PUG-view: programmatic access to chemical annotations integrated in PubChem. *J. Cheminform.* 11 (56): 1–11.
- 30 Downloading PubChem Data. <http://pubchemdocs.ncbi.nlm.nih.gov>. Retrieved 7 November 2022.
- 31 ChEMBL 30 released. <http://chembl.blogspot.com>. (10 March 2022). Retrieved 29 August 2022.
- 32 Berman, H.M., Westbrook, J., Feng, Z. et al. (2000). The Protein Data Bank. *Nucleic Acids Res.* 28: 235–242.
- 33 Bairoch, A., Apweiler, R., Wu, C.H. et al. (2005). The universal protein resource (UniProt). *Nucleic Acids Res.* 33: D154–D159.
- 34 Gaulton, A., Hersey, A., Nowotka, M. et al. (2017). The ChEMBL database in 2017. *Nucleic Acids Res.* 45 (D1): D945–D954.
- 35 Davies, M., Nowotka, M., Papadatos, G. et al. (2015). ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic Acids Res.* 43: W612–W620.
- 36 Nowotka, M.M., Gaulton, A., Mendez, D. et al. (2017). Using ChEMBL web services for building applications and data processing workflows relevant to drug discovery. *Expert Opin. Drug Discov.* 12 (8): 757–767.
- 37 Senger, S. (2017). Assessment of the significance of patent-derived information for the early identification of compound-target interaction hypotheses. *J. Cheminform.* 9 (26): 1–8.
- 38 Mendez, D., Gaulton, A., Bento, A.P. et al. (2019). ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* 47: D930–D940.
- 39 Falaguera, M.J. and Mestres, J. (2021). Identification of the Core chemical structure in SureChEMBL patents. *J. Chem. Inf. Model.* 61 (5): 2241–2247.
- 40 Wishart, D.S., Knox, C., Guo, A.C. et al. (2008). DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* 36: D901–D906.
- 41 Statistics. <http://go.drugbank.com>. Retrieved 29 August 2022.
- 42 Wishart, D.S., Feunang, Y.D., Guo, A.C. et al. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* 46: D1074–D1082.
- 43 API Support. <http://dev.drugbank.com>. Retrieved 7 November 2022.
- 44 Chen, X., Lin, Y., and Gilson, M.K. (2002). The binding database: overview and user's guide. *Biopolymers.* 61 (2): 127–141.
- 45 About Us. www.bindingdb.org. Retrieved 29 August 2022.
- 46 BindingDB Web Services. www.bindingdb.org. 7 November 2022.
- 47 Berthold, M. R.; Cebon, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. KNIME: the Konstanz Information Miner. In: Data Analysis, Machine Learning and Applications – Proceedings of the 31st Annual Conference of the Gesellschaft für Klassifikation e.V. Studies in Classification, Data Analysis, and Knowledge Organization, Berlin, Germany, 7–9 March 2007, 319–326.
- 48 Gilson, M.K., Liu, T., Baitaluk, M. et al. (2016). BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.* 44 (D1): D1045–D1053.
- 49 PDSP Ki Database. <http://pdsp.unc.edu>. Retrieved 29 August 2022.

- 50 welcome to HMDB version 5.0. Hmdb.Ca. Retrieved 29 August 2022.
- 51 About ZINC 15 Resources. <http://zinc15.docking.org>. Retrieved 29 August 2022.
- 52 Carles, F., Bourg, S., Meyer, C., and Bonnet, P. (2018). PKIDB: a curated, annotated and updated database of protein kinase inhibitors in clinical trials. *Molecules* 23 (4): 1–18.
- 53 van Linden, O.P., Kooistra, A.J., Leurs, R. et al. (2014). KLIFS: a knowledge-based structural database to navigate kinase-ligand interaction space. *J. Med. Chem.* 57 (2): 249–277.
- 54 About PROTAC-DB. <http://cadd.zju.edu.cn/protacdb>. Retrieved 7 November 2022.
- 55 Southan, C. (2015). Expanding opportunities for mining bioactive chemistry from patents. *Drug Discov. Today Technol.* 14: 3–9.
- 56 Pharma. www.gostardb.com. Retrieved 8 September 2022.
- 57 About GOSTAR. www.gostardb.com. Retrieved 8 September 2022.
- 58 Academia. www.gostardb.com. Retrieved 8 September 2022.
- 59 Excelra launches a re-envisioned version of GOSTAR. www.gostardb.com. Retrieved 7 November 2022.
- 60 GOSTAR Best-in-class SAR knowledgebase with analysis-ready datasets [Brochure]. Retrieved September 8, 2022, from GOSTAR website: <https://www.gostardb.com/wp-content/uploads/2021/08/GOSTAR-Database-Services.pdf>
- 61 Reaxys Medicinal Chemistry. www.elsevier.com. Retrieved 8 September 2022.
- 62 Production Innovation to Generate the Best Information [White paper]. Retrieved September 8, 2022, from Ural Federal University website: <https://elar.urfu.ru/bitstream/10995/31052/3/Reaxys%20Medicinal%20Chemistry%20-%20White%20Paper%20-%20Producing%20Innovation%20-%20Decemb....pdf>
- 63 Empowering hit identification and lead optimization for success in early drug discovery [Fact Sheet: Reaxys Medicinal Chemistry]. Retrieved September 8, 2022, from the French National Centre for Scientific Research: https://bib.cnrs.fr/wp-content/uploads/2018/04/R_D-Solutions_RMC_Fact-Sheet_DIGITAL.pdf
- 64 Integrating Reaxys with other chemistry research systems. www.elsevier.com. Retrieved 7 November 2022.
- 65 Clarivate analytics launches Cortellis digital health intelligence, a first-of-its-kind solution covering the global digital health ecosystem. <http://ir.clarivate.com>. 13 August 2019. Retrieved 8 September 2022.
- 66 Cortellis Drug Discovery Intelligence. www.clarivate.com. Retrieved 8 September 2022.
- 67 The Kinase Knowledgebase. <http://www.eidogen-sertanty.com>. Retrieved 8 September 2022.
- 68 Senger, S.; Leach, A. R. SAR knowledge bases in drug discovery. Wheeler, R. A.; Spellmeyer, D. C. (Eds.); *Annu. Rep. Comput. Chem.* 2008, 4, pp. 203–216.
- 69 Muresan, S., Petrov, P., Southan, C. et al. (2011). Making every SAR point count: the development of chemistry connect for the large-scale integration of structure and bioactivity data. *Drug Discov. Today*. 16 (23–24): 1019–1030.

- 70 Sharma, R., Schürer, S.C., and Muskal, S.M. (2016). High quality, small molecule-activity datasets for kinase research. *FI000Res* 5 (Chem Inf Sci-1366: 1–13.
- 71 González-Medina, M., Naveja, J.J., Sánchez-Cruz, N., and Medina-Franco, J.L. (2017). Open chemoinformatic resources to explore the structure, properties and chemical space of molecules. *RSC Adv.* 7 (85): 54153–54163.
- 72 Wang, R., Fang, X., Lu, Y., and Wang, S. (2004). The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J. Med. Chem.* 47 (12): 2977–2980.
- 73 Southan, C. (2020). Opening up connectivity between documents, structures and bioactivity. *Beilstein J. Org. Chem.* 16: 596–606.
- 74 Southan, C., Vrkonyi, P., and Muresan, S. (2009). Quantitative assessment of the expanding complementarity between public and commercial databases of bioactive compounds. *J. Cheminformatics.* 1 (1): 1–17.
- 75 Resources. www.gostardb.com. Retrieved 8 September 2022.
- 76 Release notes. <http://chembl.blogspot.com>. Retrieved 29 August 2022.
- 77 Southan, C. and Muresan, S. (2007). Complementarity between public and commercial databases: new opportunities in medicinal chemistry informatics. *Curr. Top. Med. Chem.* 7: 1502–1508.
- 78 Isigkeit, L., Chaikuad, A., and Merk, D. (2022). A consensus compound/ bioactivity dataset for data-driven drug design and chemogenomics. *Molecules* 27 (8): 1–13.
- 79 Williams, A.J. and Ekins, S. (2011). A quality alert and call for improved curation of public chemistry databases. *Drug Discov. Today.* 16: 747–750.
- 80 Opera, T.I., Olah, M., Ostropovici, L. et al. (2003). On the propagation of errors in the QSAR literature. In: *EuroQSAR 2002 Designing Drugs and Crop Protectants: Processes, Problems and Solutions* (ed. M. Ford, D. Livingstone, J. Dearden, and H. Van de Waterbeemd), 314–315. New York: Blackwell Publishing.
- 81 Data Checks. <http://chembl.blogspot.com>. 12 October 2020. Retrieved 8 September 2022.
- 82 Orchard, S., Al-Lazikani, B., Bryant, S. et al. (2011). Minimum information about a bioactive entity (MIABE). *Nat. Rev. Drug Discov.* 10 (9): 661–669.
- 83 Content Prioritization And Content Entry and Quality Control Process. Retrieved September 8, 2022, from the Eidogen website: <http://www.eidogen.com/pdfs/ContentPrioritizationEntryQCProcessAndTargetClassification.pdf>
- 84 Dragovich, P.S., Haap, W., Mulvihill, M.M. et al. (2022). Small-molecule Lead-finding trends across the Roche and Genentech research organizations. *J. Med. Chem.* 65 (4): 3606–3615.
- 85 Avram, S., Halip, L., Curpan, R., and Oprea, T.I. (2022). Novel drug targets in 2021. *Nat. Rev. Drug Discov.* 21 (5): 328.
- 86 Tyrchan, C. and Evertsson, E. (2017). Matched molecular pair analysis in short: algorithms, applications and limitations. *Comput. Struct. Biotechnol. J.* 15: 86–90.

- 87 Lipinski, C.A., Lombardo, F., Dominy, B.W., and Feeney, P.J. (1997). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* 23: 3–25.
- 88 Keefer, C.E., Chang, G., and Kauffman, G.W. (2011). Extraction of tacit knowledge from large ADME data sets via pairwise analysis. *Bioorg. Med. Chem.* 19: 3739–3749.
- 89 Gleeson, P., Bravi, G., Modi, S., and Lowe, D. (2009). ADMET rules of thumb II: a comparison of the effects of common substituents on a range of ADMET parameters *bioorg. Med. Chem.* 17: 5906–5919.
- 90 Leach, A.G., Jones, H.D., Cosgrove, D.A. et al. (2006). Matched molecular pairs as a guide in the optimization of pharmaceutical properties; a study of aqueous solubility, plasma protein binding and oral exposure. *J. Med. Chem.* 49: 6672–6682.
- 91 Matched Molecular Pair Analysis. www.gostardb.com. Retrieved 8 September 2022.
- 92 Wawer, M. and Bajorath, J. (2011). Local structural changes, global data views: graphical substructure–activity relationship trailing. *J. Med. Chem.* 54: 2944–2951.
- 93 Wassermann, A.M. and Bajorath, J. (2011). A data mining method to facilitate SAR transfer. *J. Chem. Inf. Model.* 51: 1857–1866.
- 94 Gupta-Ostermann, D., Wawer, M., Wassermann, A.M., and Bajorath, J. (2012). Graph mining for SAR transfer series. *J. Chem. Inf. Model.* 52: 935–942.
- 95 Zhang, B., Wassermann, A.M., Vogt, M., and Bajorath, J. (2012). Systematic assessment of compound series with SAR transfer potential. *J. Chem. Inf. Model.* 52: 3138–3143.
- 96 Zhang, B., Hu, Y., and Bajorath, J. (2013). SAR transfer across different targets. *J. Chem. Inf. Model.* 53: 1589–1594.
- 97 Hunt, P., Segall, M., O’Boyle, N., and Sayle, R. (2017). Practical applications of matched series analysis: SAR transfer, binding mode suggestion and data point validation. *Future Med. Chem.* 9: 153–168.
- 98 Yoshimori, A., Horita, Y., Tanoue, T., and Bajorath, J. (2019). Method for systematic analogue search using the mega SAR matrix database. *J. Chem. Inf. Model.* 59: 3727–3734.
- 99 Mills, J.E.J., Brown, A.D., Ryckmans, T. et al. (2012). SAR mining and its application to the design of TRPA1 antagonists. *MedChemComm.* 3: 174–178.
- 100 O’Boyle, N.M., Boström, J., Sayle, R.A., and Gill, A. (2014). Using matched molecular series as a predictive tool to optimize biological activity. *J. Med. Chem.* 57: 2704–2713.
- 101 Keefer, C.E. and Chang, G. (2017). The use of matched molecular series networks for Cross target structure activity relationship translation and potency prediction. *MedChemComm.* 8: 2067–2078.
- 102 Ehmki, E.S.R. and Kramer, C. (2017). Matched molecular series: measuring SAR similarity. *J. Chem. Inf. Model.* 57: 1187–1196.
- 103 The Drug-Target Interaction Heatmap. www.gostardb.com. Retrieved 8 September 2022.

- 104 LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*. 521: 436–444.
- 105 Chen, H., Engkvist, O., Wang, Y. et al. (2018). The rise of deep learning in drug discovery. *Drug Discov. Today*. 23: 1241–1250.
- 106 Hinton, G. (2018). Deep learning – a technology with the potential to transform health care. *J. Am. Med. Assoc.* 320: 1101–1102.
- 107 Wong, C.H., Siah, K.W., and Lo, A.W. (2018). Estimation of clinical trial success rates and related parameters. *Biostatistics*. 20 (2): 273–286.
- 108 Jeon, J., Nim, S., Teyra, J. et al. (2014). A systematic approach to identify novel cancer drug targets using machine learning, inhibitor design and high-throughput screening. *Genome Med.* 6 (57): 1–18.
- 109 Ferrero, E., Dunham, I., and Sanseau, P. (2017). In silico prediction of novel therapeutic targets using gene-disease association data. *J. Transl. Med.* 15 (182): 1–16.
- 110 Riniker, S., Wang, Y., Jenkins, J., and Landrum, G. (2014). Using information from historical high-throughput screens to predict active compounds. *J. Chem. Inf. Model.* 54: 1880–1891.
- 111 Godinez, W.J., Hossain, I., Lazic, S.E. et al. (2017). A multi-scale convolutional neural network for phenotyping high-content cellular images. *Bioinformatics*. 33: 2010–2019.
- 112 Tosstorff, A., Rudolph, M.G., Cole, J.C. et al. (2022). A high quality, industrial data set for binding affinity prediction: performance comparison in different early drug discovery scenarios. *J. Comput. Aided Mol. Des.* 36 (10): 753–765.
- 113 Panteleev, J., Gao, H., and Jia, L. (2018). Recent applications of machine learning in medicinal chemistry. *Bioorg. Med. Chem. Lett.* 28 (17): 2807–2815.
- 114 Vamathevan, J., Clark, D., Czodrowski, P. et al. (2019). Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* 18 (6): 463–477.
- 115 Dara, S., Dhamercherla, S., Jadav, S.S. et al. (2022). Machine learning in Drug Discovery: A review. *Artif. Intell. Rev.* 55 (3): 1947–1999.
- 116 Vijayan, R.S.K., Kihlberg, J., Cross, J.B., and Poongavanam, V. (2022). Enhancing preclinical drug discovery with artificial intelligence. *Drug Discov. Today*. 27 (4): 967–984.
- 117 Mayr, A., Klambauer, G., Unterthiner, T. et al. (2018). Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chem. Sci.* 9 (24): 5441–5451.
- 118 Sato, T., Yuki, H., Ogura, K., and Honma, T. (2018). Construction of an integrated database for hERG blocking small molecules. *PLoS One* 13 (7): 1–18.
- 119 Plonka, W., Stork, C., Šicho, M., and Kirchmair, J. (2021). CYPlebrity: machine learning models for the prediction of inhibitors of cytochrome P450 enzymes. *Bioorg. Med. Chem.* 46: 1–11.

18

Cambridge Structural Database (CSD) – Drug Discovery Through Data Mining & Knowledge-Based Tools

Francesca Stanzione¹, Rupesh Chikhale¹, and Laura Friggeri^{1,2}

¹Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, United Kingdom

²Cambridge Crystallography Data Center, Inc, One Boston Place, Suite 2600, Boston, MA 02108, US

18.1 Introduction

During drug discovery campaigns structural data related to target proteins, and drug molecules, or a combination of two, is extensively used for driving drug development and optimization. Although the value of protein–ligand structural information is very well recognized, knowledge derived from small molecule structures complements the knowledge from protein–ligand crystal structures and can significantly boost drug discovery projects [1, 2]. Access to millions of small molecule crystallographic structures, including preferred conformations and interactions, provides invaluable insights not only on the energetics of molecular conformation (bond lengths, angles, and torsion preferences) but also on molecular recognition (small molecules interacting with other small molecules in a crystal structure or with a protein). Such small molecule crystal structures and related databases are additionally used in three-dimensional (3D) searches to identify new potential drug candidates to elaborate pharmacophore models, three-dimensional quantitative structure–activity relationships (3D-QSAR), docking, and to guide de novo design [2]. Among small molecule crystallographic databases, the Cambridge Structural Database (CSD) is the most well-known worldwide repository of small molecule crystal structures. While it has been successfully used by scientists looking for patterns in crystallization and physicochemical properties of crystalline materials, it has also been proven successful in early stage drug discovery to guide and rationalize drug design and drug development [2–4].

In this chapter, we provide an overview of the CSD, and the computational software associated with the database, with a focus on how they have been successfully used in life science and drug discovery campaigns.

18.2 The Cambridge Structural Database (CSD) and CSD-Based Tools

The CSD is the world's leading and most comprehensive repository for small-molecule organic and metal–organic crystal structures. It contains every small molecule crystal structure ever published and now totals over one million structures. The CSD includes structures from X-ray and neutron diffraction analyses from a single-crystal study or a powder, study forming a unique database of high-resolution 3D structures [5]. In addition to the structures deposited to support scientific articles, the CSD includes many others published directly as CSD Communications [6].

Each entry in the CSD is manually curated by expert structural chemistry editors prior to deposition in the database. Each structure is enriched with chemical representations alongside bibliographic, chemical, and physical property information (e.g. common names, bioactivity, natural source, and cross-references to other enantiomers, racemates, or polymorphs) (Figure 18.1).

The CSD comes with several data subsets to easily navigate within millions of structures present in the database. The data subsets are classified into pesticides, approved drug molecules listed by DrugBank, and metal–organic frameworks (MOFs). MOFs are a class of porous materials, and currently, because of the global demand for clean energy storage, the CSD MOF subset can have a high impact on advancing MOF research in emergent technologies, such as catalysis, gas storage, and drug delivery [7–9]. The CSD Drug Subset contains 14 320 compounds (as of the date of preparation of this chapter) and consists of many solvates, co-crystals, or hydrated forms of drugs reported in DrugBank [10]. A single-component CSD Drug Subset is also available and includes 2282 CSD entries (as of the date of preparation of this chapter) where a drug molecule is the only component in the crystal structure [8]. A recent addition to the CSD Drug Subset is the CSD COVID-19 subset, with 303 structures of interest in the fight against COVID-19 (as of the date of preparation of this chapter). Statistical analysis and comparison of organic molecules in the CSD, internal crystal structure databases of two major pharmaceutical companies (Pfizer and AstraZeneca), and the CSD Drug Subset revealed a significant deviation from the chemical space occupied by current molecules under investigation by the pharmaceutical industry. New drugs and under-investigation compounds tend to be bigger, more complex, and more lipophilic with a high degree of deviation from the wisdom of Lipinski's rule. Interestingly, when the Pfizer and AstraZeneca databases are compared to the organic molecules in the CSD, such deviation is less pronounced.

Since its foundation, the CSD has not only grown in number but also in complexity [2, 8, 11, 12]. This has necessitated the development of search tools such as Conquest [13] and WebCSD [14, 15] to enable access to structural information contained in the CSD. Analysis tools have also been developed to enable chemical knowledge to be extracted from raw crystallographic data [16–18].

The wealth of data in the CSD was used to generate two knowledge-based libraries: (i) a library of intramolecular geometries – Mogul [19] and (ii) a library

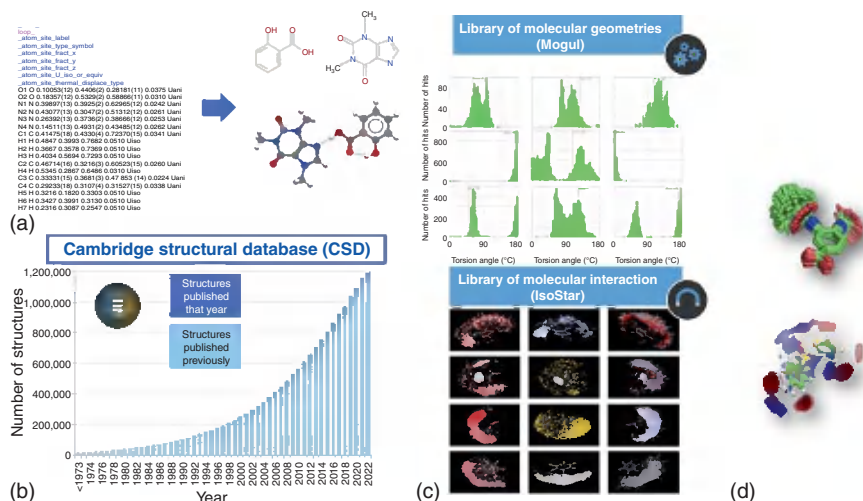


Figure 18.1 (a) Process of validating and enriching a structure before entering the CSD. Experimentally determined coordinates of atoms in a crystal are submitted in the form of a crystallographic information file (CIF), an example of CIF file is displayed in text. CIF contains information about the crystal structure as well as any details of the diffraction experiment and any data processing undertaken. Before entering the CSD, the CIF file is validated and curated, the 2D diagram and the 3D structure are generated, and the entry is enriched by crystallographic editors. (b) Yearly growth of the CSD. (c) Library of molecular geometries (Mogul) and library of molecular interactions derived by the CSD (IsoStar). Mogul is a precomputed library of geometric properties such as bond lengths, valence angles, torsion angles, and ring conformations and provides results as histograms. IsoStar is a precomputed library of intermolecular interactions derived from the CSD and the PDB. The interaction distributions are displayed as scatterplots or contour surfaces and provide a picture of preferred interactions between functional groups. (d) Two examples of intelligent software developed using Mogul and IsoStar. Top: different small molecule conformations generated using CSD Conformer Generator. Bottom: interaction map generated using full interaction maps (FIM) tool.

of intermolecular interactions – IsoStar [20], providing click-of-a-button access to millions of individual details of geometrical and chemical information derived from the CSD and (in the case of IsoStar) protein–ligand complexes derived from the Protein Data Bank (PDB) [21].

While Mogul describes molecular geometric preferences in small molecules, IsoStar gives a picture of how functional groups interact, providing an intuitive visual representation of the frequency and directionality of intermolecular interactions (Figure 18.1).

By extracting data from these two knowledge-based libraries, additional software has been developed not only to support scientists working with structural data but also to extract new insights from proprietary, experimental, and predicted data [13, 22–31] (Figure 18.1). A detailed description of the tools is beyond the scope of the chapter, but we will highlight some of the tool usage in the later sections of the chapter.

18.3 How CSD and CSD Knowledge-Based Tools Can Aid Drug Discovery

Drug discovery campaigns are long, difficult, and expensive where, early identification of hit compound(s) is a key to the success of the campaign. From target validation to lead optimization, crystallographic data and knowledge-based methods such as those that use the CSD can be useful to hasten and optimize the overall drug design and development.

For example, the availability of a 3D structure gives insight into the potential conformations of a molecule and, furthermore, reveals the interactions that a molecule makes with itself and with neighboring molecules. Comparison of these conformations and interactions with those made with a protein can be revealing (i.e. adjusting conformation to optimize the protein's interactions and understanding geometrical strains) [1, 4, 32].

Here we are going to highlight the contribution of crystallographic data, and knowledge-based tools derived from the CSD, in the different stages of the drug discovery pipeline with a focus on a few significant published examples.

18.3.1 Target Identification and Target Validation

Target identification and target validation are the first critical steps in discovering a new drug, if a target cannot be validated, then it will not proceed in the drug development process. Moreover, insufficient validation of drug targets can lead to costly clinical failures [33, 34] and low drug approval rates [34]. As a result, there is a common consensus that robust target validation is crucial and deserves greater emphasis to facilitate the development of new drugs. Recommendations for target validation and assessment have been widely discussed in several reviews [34, 35].

In general, the target identification process involves the application of different techniques to identify “druggable” proteins (i.e. proteins that can be targeted by a drug) and to demonstrate that drug effects on the target can provide a therapeutic benefit with an acceptable safety window [36]. If a drug for the desired target has already been identified, then the target is druggable, if no drugs have been identified, then the druggability of the target should be assessed and validated.

Common computational methods used to assess target druggability rely on evolutionary relationships (i.e. proteins from the same family are known to be targeted by drugs) and 3D structural information, including properties and/or other descriptors [36–40]. Structure-based methods rely on the availability of experimental 3D structures or high-quality homology models and consist of three main components (i) identifying cavities or pockets in the target structure; (ii) calculating physicochemical and geometric properties of the pockets; and (iii) assessing how these properties fit a training set of known druggable targets [37, 38, 41–44].

A variety of algorithmic approaches exist to help to identify cavities, for example, LIGSITE [45] uses grid-based ray tracing to identify pockets; GHECOM [46] uses recursive sphere-based detection, whereas newer algorithms, such as CavVis, use

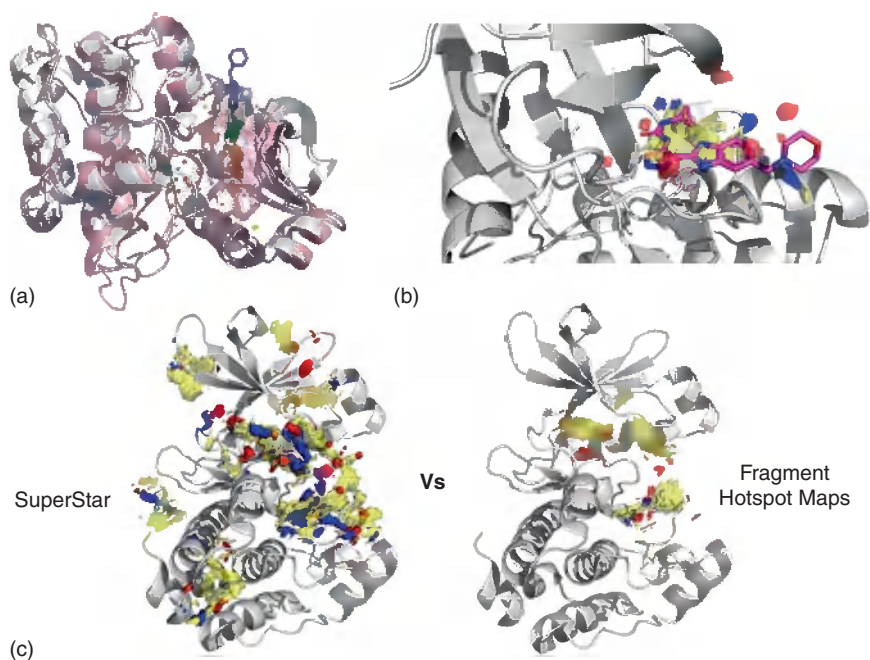


Figure 18.2 (a) Result of a cavity comparison. Alignment and superimposition of Aurora Kinase (gray) in complex with an inhibitor (blue) (PDB code 2W1G) and the Human PDK1 Kinase Domain (pink) in complex with ATP (green) (PDB code 4A07). The cavity of the Aurora Kinase was determined using LIGSITE. The surface points (pale yellow dots) are generated to define the shape of the binding cavity, and the pseudocenters are assigned based on the physicochemical properties of the amino acids in the cavities: hydrogen bond acceptors (red), hydrogen bond donors (blue), aromatics (yellow), pi (gray), and hydrogen bond donor/acceptor (green). (b) Binding site interaction maps of Aurora Kinase (gray) in complex with an inhibitor (magenta) (PDB code 2W1G). The map is generated using SuperStar and displayed as contoured maps, hydrogen bond acceptor (red), hydrogen bond donor (blue), and hydrophobic region (yellow). High contours indicate regions where the probe group has a high propensity to occur. (c) SuperStar contoured maps vs. Fragment Hotspot maps were generated for the full Aurora Kinase protein (gray). Hydrogen bond acceptor (red), hydrogen bond donor (blue), and hydrophobic region (yellow). High contours indicate regions where the probe group has a high propensity to occur. The figure was generated using The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.

methods inspired by computer graphics, [47]. All these different approaches provide insight into the size and shape of protein cavities that could be targeted by a drug. Identified cavities can be then compared with cavities of related and unrelated proteins to provide an early assessment of cross-reactivity between targets, bioselectivity, or to elucidate the function of orphan proteins [48–52] (Figure 18.2a).

Furthermore, determining the physicochemical properties of the pocket can guide the rational design of new drugs or optimization of lead compounds by providing insights into the preferred protein interactions. One approach is to estimate the probability of interactions to occur based on statistical information derived from experimental X-ray structures. SuperStar [25] is one of these empirical methods, and it uses

the library of intermolecular interactions derived from the CSD and PDB IsoStar, to identify regions within a protein or around small molecules where particular functional groups (probes) are likely to interact favorably. The interaction maps are generated, and the propensity is reported based on how often the interaction has been observed in crystal structures (Figure 18.2b). If functional groups of a selected ligand, which are of the same type as those used as a probe (for example H-bond acceptors) are located within the higher propensity regions of these maps, then one may infer that such groups make favorable interactions with the protein. However, if these maps are not fully satisfied, drug design can be guided toward changes that will satisfy more or all of the predicted interactions [25, 53]. While SuperStar provides a picture of small functional group's interactions in defined cavities within proteins, such an approach for apo proteins (proteins without ligands) can be quite challenging because it is not possible to prioritize cavities for the sampling, and therefore the derived interaction maps can be noisy (Figure 18.2c).

In 2016, Radoux et al. developed a fragment hotspot map tool to identify hotspot regions in proteins by prioritizing SuperStar's cavities located in hydrophobic regions that are large enough to accommodate a fragment in its binding mode [29]. Fragment hotspot maps highlight fragment-binding sites and their corresponding pharmacophores by sampling the cavities with three pseudomolecular probes that represent an aromatic fragment with an H-bond acceptor, an H-bond donor, or an apolar functional group. The algorithm outputs a set of three maps, one for each interaction probe type, with a relative score, where highly scoring points denote areas where a fragment is likely to form this type of interaction (Figure 18.2c).

A research team at the University of Cambridge has used fragment hotspot mapping to understand the nature of the catalytic core of their cryo-EM structure of the DNA-dependent protein kinase catalytic subunit (DNA-PKcs). DNA-PKcs is of great importance in repairing pathological DNA double-strand breaks (DSBs), making DNA-PKcs inhibitors attractive therapeutic agents for cancer in combination with DSB-inducing radiotherapy and chemotherapy. Fragment hotspot maps revealed interesting binding regions and helped in interpreting and rationalizing the selectivity of known inhibitors [54]. A recent fragment hotspot algorithm optimization resulted in automatization, implementation of different functional probes [55], and ability to work with multiple protein conformations [56]. By comparing two or more ensemble maps, it is possible to identify conserved interactions (i.e. pharmacophoric regions) and to derive a hotspot selectivity map, providing insights into structural differences determining the selectivity of a compound for one protein over another similar protein domain family [56].

A different, interesting approach to target identification has been used by researchers from the N. D. Zelinsky Institute of Organic Chemistry. Here, Elinson et al., used available protein–ligand complexes derived from the PDB to find potential protein target(s) for their scaffold – an unsymmetric spirobarbituric dihydrofuran – a promising new scaffold with many potential pharmaceutical applications [57]. They performed a scaffold-based search using CSD-CrossMiner, a pharmacophore-based tool that enables the mining of crystallographic data from both the CSD and the PDB [26]. In CSD-CrossMiner, pharmacophore queries can

be generated using feature definitions that describe the ensemble of electronic and steric properties of molecules. Pharmacophore queries can describe protein–ligand interactions, ligand scaffolds, or protein environments. Pharmacophores are intuitive, abstract, and transferable; therefore, pharmacophore-based approaches are usually preferred among other scaffold-hopping methods.

In their approach, Elinson et al. generated a pharmacophore query described in Figure 18.1 to search the PDB subset of CSD-CrossMiner for protein pockets interacting with a similar scaffold. Among the hits retrieved, the human aldose reductase in complex with a small molecule (PDB code 2NVD), revealed a great similarity to the co-crystallized ligand and the spirobarbituric dihydrofuran compound. CSD-CrossMiner also helped to highlight the role of hydrogen bonds and π – π interactions in protein–ligand binding. Further molecular docking studies supported the scaffold search findings, suggesting the promising application of spirobarbituric dihydrofurans as aldose-reductase inhibitors for aldose reductase-mediated diseases and diabetes in particular [58] (Figure 18.3).

18.3.2 Hit Identification

Hit identification is the second key step in the drug discovery pipeline. Here, computational approaches such as *in silico* high-throughput screening (HTS) have been

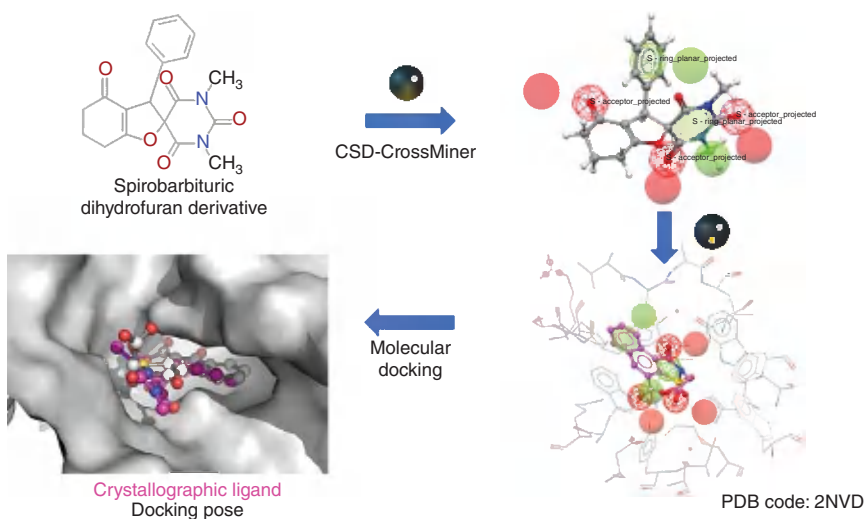


Figure 18.3 The 3D structure of the designed spirobarbituric dihydrofuran derivative was used to generate a pharmacophore query in CSD-CrossMiner. The pharmacophore query consists of two planar rings projected (green spheres) and three hydrogen bond acceptors (red spheres). CSD-CrossMiner is then used to search the protein–ligand binding site dataset for protein pockets binding to similar scaffolds. Among the matching hits, the protein–ligand complex (PDB code 2NVD) was selected. The hit matching the pharmacophore query is displayed. Molecular docking investigations confirmed the similarity of the binding of the native crystallographic ligand (magenta) with the spirobarbituric dihydrofuran derivative docking pose (gray).

successfully used to identify drug candidates, to guide experiments, and accelerate drug design processes [59]. When the 3D structure of the target is available, computational methods such as molecular docking are usually applied to predict the position, orientation, and conformation of a small molecule when bound to the target [60–62]. There are many programs available in this space like Schrodinger-GLIDE [63], OpenEye-FRED [64], CCG-MOE [65], ICM-PRO [66], AutoDock [67], GOLD [68], and several others, which are used for molecular docking studies toward screening of large datasets or focused libraries of ligands [69].

Alternatively, when a 3D structure of the protein target is not available, a ligand-based approach that relies on knowledge of molecules that bind to the biological target of interest rather than the structure of a target protein can be performed. Here, 3D-QSAR and pharmacophore modeling methods are generally used to guide the virtual screening of compound databases [70]. These methods correlate molecular structure with properties (e.g. *in vitro* or *in vivo* biological activity) and assume that the structure of a molecule contains the features responsible for its physical, chemical, and biological activities. They can provide predictive models suitable for hit identification and lead optimization [71].

To be active, a potential drug needs to access to the conformation required for binding to the target. Therefore, the prediction of energetically accessible molecular geometries is of central importance in rational drug design. Force fields and quantum mechanical calculations are routinely used; however, not all the possible conformations are accessible or experimentally observed [2, 72]. Additionally, when looking at protein–ligand complexes, ligand conformations in PDB should be treated with caution because of the limitation of the crystallographic resolution. On occasion, the quality of ligand models in the PDB can be poor, with unfeasible and/or strained conformations [73]. It is not always clear if a ligand structure containing highly strained chemical groups is real or is an artifact of a lack of experimental data or the protocols used by the crystallographer [1, 73].

Small-molecule crystal structures, like the ones in the CSD, are generally far higher in resolution and so not subject to the same challenges observed in proteins. Moreover, structural knowledge stored in the CSD provides a big advantage in assessing real, experimentally accessible conformations in a condensed phase that resemble the extended conformations that ligands tend to adopt when interacting with a protein to maximize interactions with the binding site. This fact is also supported by the observation that favorable interactions, such as H-bonds, occur preminently as intermolecular rather than intramolecular among the CSD. These data suggested that, although small molecule crystals can adopt an unlike conformation compared to the hydrated protein–ligand complexes, the CSD crystal structures can be used to fit the ligand and identify the best model for the binding pose, particularly when the protein–ligand complexes have been solved with low resolution and the conformation within the electron density can be wrong [2, 73, 74].

The geometric library derived from the CSD, Mogul, plays a central role in evaluating the conformation of ligands in protein–ligand crystallographic complexes, and the Mogul report is included in the wwPDB validation pipeline [75]. For each bond length and bond angle in the ligand, a search is performed for small-molecule crystal

structures in the CSD that have a similar chemical environment, proving an overall assessment of the ligand refinement.

In 2018, Cole et al. used the wealth of structural data in the CSD to develop the CSD conformer generator, a knowledge-based approach that uses the diversity of bond, angle, torsion, and ring information in the CSD to compute an ensemble of realistic low energy conformations [24]. The CSD conformer generator, combined with the CSD ligand overlay and field-based ligand screener, can address key steps in a ligand-based screening workflow [27].

The CSD conformer generator has been successfully used to aid in the design of new bioinspired phosphoenolpyruvate (PEP) antibiotic competitive inhibitors against two of the main enzymes involved in the shikimate pathway: deoxy-D-arabino-heptulosonate-7-phosphate synthase (DAHPS) and 5-enolpyruvylshikimate 3-phosphate synthase (EPSPS) [76]. De Oliveira et al. used the CSD Conformer Generator to generate low-energy conformations of 28 PEP derivatives retrieved from the literature and then used the CSD Ligand Overlay program to derive the pharmacophore model. Combining the pharmacophoric prediction with other computational approaches, De Oliveira et al. identified a potential multi-target inhibitor for both DAHPS and EPSPS that could be used as starting point for more potent drug candidates [76].

Another way to perform hit identification is by searching small molecules and protein–ligand structural databases. An example of such an approach has been reported by Manetti et al. in the design of new nicotinic acetylcholine receptors (nAChRs) inhibitors with improved activity or selectivity. The requirements for the nicotinic pharmacophore, shown in Figure 18.4, were derived from the pyrido[3,4-b]homotropene (PHT) ligand and transformed into a 3D geometric query to search the CSD using Conquest (the CSD search system) with the aim of finding new chemotypes for nAChRs [77, 78]. Among the retrieved hits, the CSD entries with refcodes VIXLAX and LOYMOB inspired two new scaffolds. The VIXLAX structure was further simplified to a family of quinoline derivatives, while the LOYMOB structure was simplified to a family of indole derivatives.

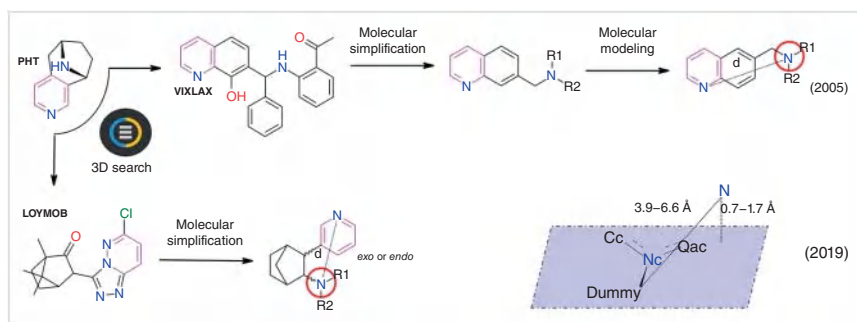


Figure 18.4 Identification of novel chemotypes vs nAChRs using CSD 3D database searching engine. In the bottom right, a schematic representation of the queries used by Manetti et al. [77] to search structures into the CSD (Qa = N, C: subscript c = atoms belonging to ring. The distances of N to the plane and N to the dummy are in the range of 0.7–1.7 and 3.9–6.6 Å respectively). Adapted from Manetti et al. [77].

thought to be a more suitable lead compound in terms of synthetic accessibility, chemical stability, molecular weight, and overall lipophilicity [78]. While the LOYMOB structure suggested the use of a simple bicyclo[2.2.1]heptane moiety as a spacer between the pharmacophore features, leading to the design of a new potential scaffold (Figure 18.4) [77]. The nicotinic activity of the new scaffold was assessed with both computational and *in vitro* experiments, such as radioligand binding affinity on $\alpha 4\beta 2$, and $\alpha 7$ nicotinic receptors of the rat brain. The novel compounds with high binding affinities were also tested by using SHSY5Y cells and on the heterologously expressed individual $\alpha 4\beta 2$, $\alpha 7$, and $\alpha 3\beta 2$ subtypes. This research led to the identification and synthesis of two new chemotypes for several subtypes of neuronal nAChRs. While promising, the authors noted the selectivity would need to be improved [77].

Another example of mining crystallographic data to inspire the design of new scaffolds is provided by researchers from Santen Pharmaceuticals. Iwamura et al. used Conquest in combination with structure–activity relationship data to design new selective non-prostanoid EP2 receptor agonists for the treatment of glaucoma [79]. The statistical analysis of a selective EP2 receptor agonist CP-533536 [80] retrieved from Conquest, provided detailed information on preferred geometric conformations, including a clear preference for a planar (0° or 180°) conformation of the torsion angle $\text{CH}=\text{C}-\text{O}-\text{CH}_2$ in the phenoxyacetic acid moiety. These analyses, combined with medicinal chemistry efforts, led to the development of a prodrug named Omidenepag isopropyl (OMDI), in which a (pyridin-2-ylamino)-acetic acid moiety replaces the phenoxyacetic group. OMDI is now in clinical trials for the treatment of glaucoma, see Figure 18.5 (<https://clinicaltrials.gov/ct2/results?term=Omidenepag+isopropyl>).

18.4 Hit-to-Lead

Following the hit identification (e.g. from an HTS and/or virtual screening), further characterization in terms of intrinsic potency, selectivity, off-target activity, synthesis, and patentability should be evaluated [59]. After analyzing hit molecules and clustering based on chemical similarities, the overall goal is to optimize the most promising series and explore the diversity beyond a particular scaffold.

Computational methods such as molecular docking are very popular at this stage because they can provide a detailed picture of the ligand binding pose and protein–ligand interactions [62, 81]. Conformer generator tools, such as the CSD Conformer Generator, can be used in tandem with docking to generate initial low-energy conformations for ligands, and database mining software, such as CSD-CrossMiner, can be used to further explore protein–ligand complexes (e.g. docking solutions), and to expand diversity over a particular scaffold (i.e. scaffold hopping) [26].

With the aim of exploring an alternate scaffold, a group at the University of Groningen used a pharmacophore approach to simultaneously search in the CSD

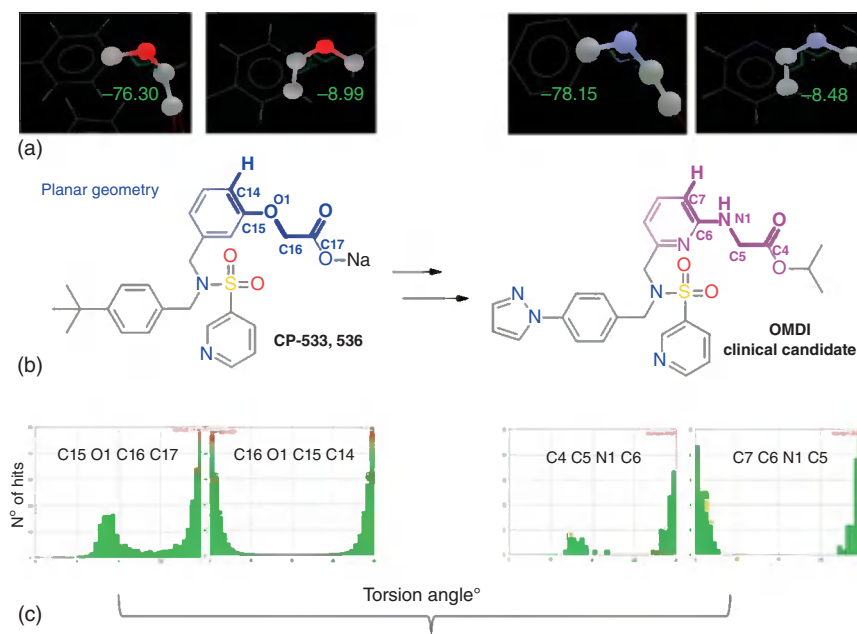


Figure 18.5 Representation of geometric conformations. (a) Planar geometry angles. The similarity of torsion angle values between CP-533, 536, and OMDI compounds. (b) 2D chemical structures of CP-533, 536, OMDI, and structure optimization. (c) Representation of torsion distribution retrieved from the CSD database.

and PDB databases [82]. Starting from a designed library of drug-like molecules, Konstantinidou et al. solved the structure of one molecule to confirm its structural properties and then used CSD-CrossMiner [26] to generate a pharmacophore model searching against the CSD and PDB databases. The pharmacophore template was generated from a reference PDB structure (PDB code 4R3M) that showed a very similar binding pattern to their scaffold. The search returned 37 co-crystal structures with similar structural motifs, providing potential future scaffold hopping alternatives [82].

In the hit-to-lead stage, knowledge-based tools such as Mogul, IsoStar, and SuperStar provide details of usual/unusual conformations and preferred interactions to further help to characterise the lead compound.

Additionally, molecular docking experiments can suggest possible conformational binding poses of potential lead compound(s) and elucidate protein–ligand binding interactions (e.g. are all the possible hydrogen bonds formed?). Among the docking tools, GOLD docking software can be enhanced to use CSD-derived torsion libraries to restrict the ligand conformational space sampled by the searching algorithm [68]. The use of the torsion angle distributions increases both efficiency and effectiveness of the algorithm, improving the chances of GOLD finding the correct answer by biasing the search toward ligand torsion-angle values that are commonly observed in crystal structures [83].

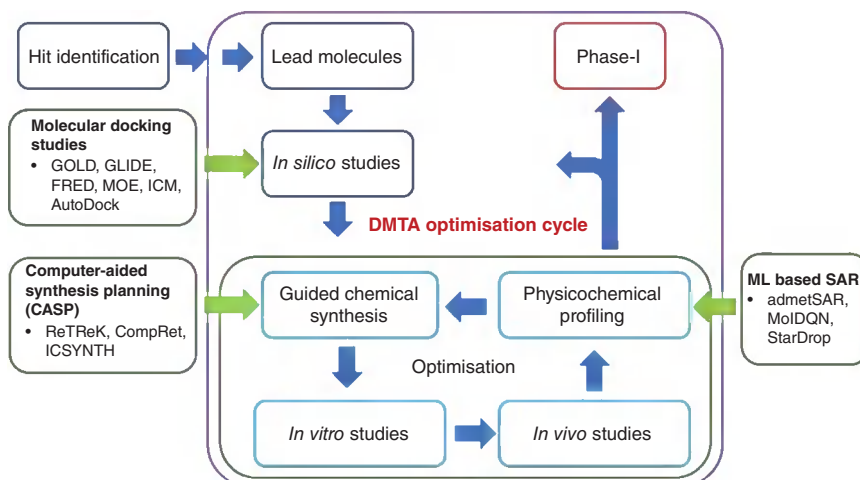


Figure 18.6 Lead optimization process in a drug discovery pipeline with the role of various predictive and knowledge-based tools.

18.4.1 Lead Optimization

Lead optimization phase of drug discovery aims to optimize the lead compound, improving not only potency, if necessary, but also the absorption, distribution, metabolism, excretion, and toxicity (ADMET) profiles. Here, a deep analysis of the ligand binding mode, together with the information obtained by co-crystal complexes, modeling and experimental studies, can be significant in the optimization of selected lead molecules [84].

On a structural basis, it is possible to modify functional groups of the lead molecule by bioisosteric replacement or by designing novel scaffold(s) focusing on retaining or increasing interactions (i.e. H-bonds) with the binding site residues while improving the ADMET profile. In theory, lead optimization consists of several iteration cycles: Design, Make, Test, Analyze (DMTA) cycle (Figure 18.6) [85, 86] in which the lead molecule is modified, novel compounds are synthesized, and they are further tested *in vitro* and *in vivo* to evaluate the efficacy, selectivity, and ADMET profile. Compounds with better safety and potency profiles are selected and subjected to further DMTA cycles until an optimum overall result is achieved to take forward to Phase-I studies [86].

Most, if not all, the computational methods used in the hit-to-lead stage can be employed in the lead optimization. Mogul, for example, is generally used to assess the quality of the set of lead molecules based on preferred geometries and conformations in complexes with protein targets [73]. Moreover, IsoStar can be used to guide bioisosteric replacement by helping the identification of functional groups that share similar interaction patterns [87]. Also, molecular docking helps to characterize binding modes of the lead molecule in the protein target or toward an ensemble of protein conformations [88]. Interaction maps (such as those generated by SuperStar) can be

used to guide so-called “biased docking”, a docking experiment “biased” towards solutions with specific, defined features (e.g. hydrogen bond(s)).

Crystallographic databases and tools mining structural databases can also play a crucial part in the lead optimization process of nucleic acids modulators. In particular, CSD-CrossMiner explores the Nucleic Acid database derived from the PDB, performing molecular similarity and pharmacophoric searches that could lead to the identification of new nucleic acid modulators. One example comes from Axford et al., where they used structural information from the CSD to guide the optimization of the lead molecule branaplam [89]. Branaplam and its analogs have promising applications in the treatment of spinal muscular atrophy (SMA) by acting as survival motor neuron (SMN2) splicing modulators. Compound **1** was obtained from an HTS against the Neuroblastoma Spinal Cord-34 (NSC34) motor neuron cell line, which is an experimental model for SMA drug development. This hit molecule was then chemically modified and optimized to improve its efficacy. Experimental data suggested the design of a planar conformation between the phenol and the proximal nitrogen of pyridazine leading to compound **2** that showed better potency (ELISA EC_{50} = 33 nM), pharmacokinetics (C_{max} improved to 362 nM, its $t_{1/2}$ improved to 5.1 h), and CNS distribution to provide six-fold increase in the brain levels of survival motor neuron (SMN) protein (Figure 18.7). The nearly coplanar conformation of the pyridazine-phenyl system is held by a hydrogen bond between the pyridazine and phenol, and it is important for its biological activity. To improve the biological properties and oral bioavailability, the central pyridazine was also replaced with a thiadiazole to take advantage of the conformational constraints from ring nitrogen and sulfur. These insights were informed by CSD data, where Axford et al., found several 2-fluorophenyl moieties in combination with thiazoles, thiophenes, and thiadiazoles forming $S \cdots X$ ($X = F, Cl$) interactions. These observations suggested a replacement of the initial oxygen-based conformationally restrained (-OH, -OMe) compounds (see **2** in Figure 18.7) with fluorine or chlorine, where intramolecular sulfur-halogen chalcogen bonding provided an attractive alternative to hydrogen-bonding with $S \cdots O$ interactions. Accordingly, several derivatives of compound **3** were synthesized with potency, pharmacokinetics, and central nervous system distribution sufficient to provide a 50% increase in the production of full-length SMN protein in a mouse model of SMA.

Modification of functional groups over a lead molecule is also a common step in the optimization process. There are several examples based on the use of bioisosteres for lead optimization. Bioisosteres are molecules that have chemical and physical similarities producing similar biological effects. Bioisosteric replacements affect several parameters of the lead molecule; these could be size, conformation, inductive effect, mesomeric effects, hydrogen bond formation capacity, solubility, reactivity, stability, and the pK_a [90]. Bioisosteric replacements can help in optimizing ligands by improving selectivity, lowering side effects, reducing toxicity, and improving the pharmacokinetic properties and stability [91].

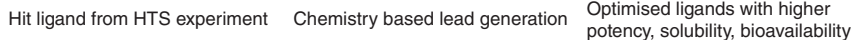


Figure 18.7 CSD guides the lead optimization of pyridazine compounds to coplanar thiadiazoles with improved biological and physicochemical properties.

18.5 Challenges in Drug Development

Drug products are made of two key elements: the active pharmaceutical ingredient (API), which is the biologically active component of the drug, and excipients, which are chemically inactive substances that deliver the medication into the body.

Once the lead compound has been identified and optimized through the process of drug discovery, it is of crucial importance to develop an API that will be stable and safe in a pharmaceutical formulation.

Post-clinical drug development involves the selection of the best-suited form of an API, optimization of its form (if required), development of the API formulation, and administration of the drug molecule. There are several challenges in drug development that are in fact related to the physical form and chemical nature of the API [92]. The specific solid form chosen can affect many important physical properties, such as solubility, bioavailability, and stability. The selection of API's physical and chemical forms is well assisted by the advancement of computational tools, including knowledge-based modeling and databases with information relating to crystal structures and diffraction patterns [2].

When talking about crystal forms and crystallized molecules, drug polymorphism is a major issue because it can result in diverse chemical and physical properties that can affect the stability, solubility, compressibility, and pharmacological efficacy of the API [93].

A good example of drug polymorphism is loratadine form II. Loratadine is a commercially available drug used for the treatment of allergies. It has been crystallized under low temperatures and the X-ray structure is available; however, knowledge about its polymorphic structures and disordered nature is not known. Woollam et al. reported a metastable polymorph of loratadine (form II) identified by 3D electron diffraction studies [94]. They crystallized this form and performed powder X-ray diffraction, transmission electron microscope (TEM), and 3D electron diffraction (3D ED) data collection. The analysis suggested two different conformations for the loratadine form II 7-membered rings, referred to as type 1 and type 2 conformations, respectively.

Woollam et al. used Mercury software [16] to evaluate the quality of the structure by comparison with known structures in the CSD and by calculating structural

overlay and crystal packing similarity [94]. A search for the seven-membered ring torsion via Mogul in the CSD showed the flipping of the cyclopropane bridge connected to aromatic rings, and the overlay of these conformations highlighted the disordered form of loratadine form II. Similar findings assisted by the CSD database could help in-depth understanding of drug polymorphism and its impact on formulation development.

Early characterization of these polymorphs is therefore of great value, and several computational tools are available, including assessment of complete molecular geometries [24], binding hotspots around molecules (full interaction maps, FIMs) [95], and analysis of complete H-bonding networks (hydrogen bond propensities, HBPs) [95–97]. The FIMs method is closely related to SuperStar and provides information on the preferred interactions of a target compound, offering a clear visualization of those interactions that deviate from their ideal geometry (Figure 18.1d). The HBP method uses CSD data to estimate the likelihood of an H-bond between each possible combination of the donor and acceptor groups in the target molecule. This helps identify polymorphic forms that may be metastable because they contain low-probability configurations of H-bonds.

Most major pharmaceutical companies now apply these sophisticated solid-form informatics methods together, on their drug development candidates, to produce a full picture of the risk of polymorphism.

Dávila-Miliani et al., for example, have employed the HBP tool in combination with Hirshfeld surface analysis [98] and energy framework calculations [99] to investigate the nonsteroidal anti-inflammatory drug Flunixin and its close relative Clonixin. The HBP calculation successfully ranked the two polymorphs of Flunixin well and indicated the possible existence of other, yet unobserved, polymorphs. The complementary Hirshfeld surface analysis and energy framework calculations show the differences between the two polymorphs of Flunixin and their similarities with polymorphs of the related Clonixin [100].

18.5.1 How CSD Can Further Impact Drug Discovery

Here we have described how the wealth of information in the CSD, accumulated over many years, has been used as a data source for knowledge-based predictive algorithms.

At a time when scientific “big data,” machine learning (ML) and artificial intelligence (AI) methods are growing in popularity, the quality of scientific data is more important than ever and, the CSD with its high-resolution curated small molecule crystal structures can play a pivotal role. The high quality of molecular structures not only provides a clear picture of the conformations and interactions of small molecules but also has the additional advantage of applying this information for the generation of large, high-quality, unbiased data for drug discovery. The CSD tools and the database have been impactful in various areas of drug discovery, and now it looks forward to the development of ML-based methods for drug discovery and development.

Similarly, we can see that the future lies in the integration of knowledge bases like the CSD into customized ML pipelines and their integration for the purpose of open-source drug discovery programs. Additionally, database mining tools such as CSD-CrossMiner provide access to 3D structure databases such as the PDB, CSD, and Nucleic acid databases. Soon, along with the 3D data currently provided it will be possible the mining of external open-access ligand databases and proprietary databases for performing ligand screening, thus speeding up the hit identification process.

References

- 1 Groom CR, Cole JC. The use of small-molecule structures to complement protein-ligand crystal structures in drug discovery. *Acta Crystallogr. D Struct. Biol.* 2017 73 (Pt 3): 240-245.
- 2 Taylor R, Wood PA. *A Million Crystal Structures: The Whole Is Greater than the Sum of its Parts*. *Chem Rev.* 119 (16): 9427-9477 2019.
- 3 Allen, F.H. and Taylor, R. (2004). Research applications of the Cambridge Structural Database (CSD). *Chem. Soc. Rev.* 33 (8): 463–475.
- 4 Cole, J.C., Wiggin, S., and Stanzione, F. (2019). New insights and innovation from a million crystal structures in the Cambridge Structural Database. *Struct. Dyn.* 6 (5): 054301.
- 5 Groom, C.R., Bruno, I.J., Lightfoot, M.P., and Ward, S.C. (2016). The Cambridge Structural Database. *Acta Crystallogr. B Struct. Sci. Cryst. Eng. Mater.* 72 (Pt 2): 171–179.
- 6 Ferrence, G.M., Tovee, C.A., Holgate, S.J.W. et al. (2023). CSD Communications of the Cambridge Structural Database. *IUCrJ.* 10 (Pt 1): 6–15.
- 7 Li, A., Perez, R.B., Wiggin, S. et al. (2021). The launch of a freely accessible MOF CIF collection from the CSD. *Matter* 4: 1–2.
- 8 Bryant, M.J., Black, S.N., Blade, H. et al. (2019). The CSD drug subset: the changing chemistry and crystallography of small molecule pharmaceuticals. *J. Pharm. Sci.* 108 (5): 1655–1662.
- 9 Wishart, D.S., Knox, C., Guo, A.C. et al. (2006). DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* 34 (Database issue): D668–D672.
- 10 Moghadam PZ, Li A, Liu XW, Bueno-Perez R, Wang SD, Wiggin SB, et al. *Targeted Classification of Metal-Organic Frameworks in the Cambridge Structural Database (CSD)*. 2020 [cited 2023 Jan 15]; Available from: http://aam.ceb.cam.ac.uk/mof-explorer/CSD_MOF_subset.
- 11 Samuel Motherwell, W.D. (2008). The CSD-450,000 answers... But what are the questions? *Crystallogr. Rev.* 14 (2): 97–116.
- 12 Groom, C.R. and Allen, F.H. (2014). The Cambridge structural database in retrospect and prospect. *Angew. Chem. Int. Ed.* 14 (2): 97–116.

- 13 Bruno, I.J., Cole, J.C., Edgington, P.R. et al. (2002). New software for searching the Cambridge Structural Database and visualizing crystal structures. *Acta Crystallogr. D: Struct. Biol.* 58 (Pt 3 Pt 1): 389–397.
- 14 Thomas, I.R., Bruno, I.J., Cole, J.C. et al. (2010). WebCSD: the online portal to the Cambridge Structural Database. *J. Appl. Crystallogr.* 43 (Pt 2): 362–366.
- 15 Battle, G.M. (2011). WebCSD: bringing the Cambridge structural database to undergraduate teaching. *Acta Crystallogr. D: Struct. Biol.* 67: C209–C209.
- 16 Macrae, C.F., Sovago, I., Cottrell, S.J. et al. (2020). Mercury 4.0: from visualization to analysis, design and prediction. *J. Appl. Crystallogr.* 53 (1): 226–235.
- 17 Macrae, C.F., Bruno, I.J., Chisholm, J.A. et al. (2008). Mercury CSD 2.0 – new features for the visualization and investigation of crystal structures. *J. Appl. Crystallogr.* 41 (2): 466–470.
- 18 Macrae, C.F., Edgington, P.R., McCabe, P. et al. (2006). Mercury: visualization and analysis of crystal structures. *J. Appl. Crystallogr.* 39 (3): 453–457.
- 19 Bruno, I.J., Cole, J.C., Kessler, M. et al. (2004). Retrieval of crystallographically-derived molecular geometry information. *J. Chem. Inf. Comput. Sci.* 44 (6): 2133–2144.
- 20 Bruno, I.J., Cole, J.C., Lommerse, J.P.M. et al. (1997). IsoStar: a library of information about nonbonded interactions. *J. Comput. Aided Mol. Des.* 11 (6): 525–537.
- 21 Berman, H.M., Westbrook, J., Feng, Z. et al. (2000). The Protein Data Bank. *Nucleic Acids Res.* 28 (1): 235–242.
- 22 Cottrell, S.J., Olsson, T.S.G., Taylor, R. et al. (2012). Validating and understanding ring conformations using small molecule crystallographic data. *J. Chem. Inf. Model.* 52 (4): 956–962.
- 23 Taylor, R., Cole, J., Korb, O., and McCabe, P. (2014). Knowledge-based libraries for predicting the geometric preferences of druglike molecules. *J. Chem. Inf. Model.* 54 (9): 2500–2514.
- 24 Cole JC, Korb O, McCabe P, Read MG, Taylor R. *Knowledge-Based Conformer Generation Using the Cambridge Structural Database*. 2018 58 (3): 615–629.
- 25 Verdonk ML, Cole JC, Taylor R. *SuperStar: A Knowledge-based Approach for Identifying Interaction Sites in Proteins*. 1999 289 (4): 1093–1108.
- 26 Korb, O., Kuhn, B., Hert, J. et al. (2016). Interactive and versatile navigation of structural databases. *J. Med. Chem.* 59 (9): 4257–4266.
- 27 Giangreco, I., Mukhopadhyay, A., and Cole, J.C. (2021). Validation of a field-based ligand screener using a novel benchmarking data set for assessing 3D-based virtual screening methods. *J. Chem. Inf. Model.* 61 (12): 5841–5852.
- 28 Velec, H.F.G., Gohlke, H., and Klebe, G. (2005). DrugScoreCSDKnowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. *J. Med. Chem.* 48 (20): 6296–6303.
- 29 Radoux, C.J., Olsson, T.S.G., Pitt, W.R. et al. (2016). Identifying interactions that determine fragment binding at protein hotspots. *J. Med. Chem.* 59 (9): 4314–4325.

- 30 Vriza, A., Sovago, I., Widdowson, D. et al. (2022). Molecular set transformer: attending to the co-crystals in the Cambridge structural database. *Digital. Discovery* 1: 834–850.
- 31 Shevchenko, A.P., Eremin, R.A., and Blatov, V.A. (2020). The CSD and knowledge databases: from answers to questions. *CrystEngComm* 22 (43): 7298–7307.
- 32 Brameld, K.A., Kuhn, B., Reuter, D.C., and Stahl, M. (2008). Small molecule conformational preferences derived from crystal structure data. A medicinal chemistry focused analysis. *J. Chem. Inf. Model.* 48 (1): 1–24.
- 33 Bunnage, M.E. (2011). Getting pharmaceutical R&D back on target. *Nat. Chem. Biol.* 7 (6): 335–339.
- 34 Emmerich, C.H., Gamboa, L.M., Hofmann, M.C.J. et al. (2021). Improving target assessment in biomedical research: the GOT-IT recommendations. *Nat. Rev. Drug Discov.* 20 (1): 64–81.
- 35 Schenone, M., Dančik, V., Wagner, B.K., and Clemons, P.A. (2013). Target identification and mechanism of action in chemical biology and drug discovery. *Nat. Chem. Biol.* 9 (4): 232–240.
- 36 Agoni, C., Olotu, F.A., Ramharack, P., and Soliman, M.E. (2020). Druggability and drug-likeness concepts in drug design: are biomodelling and predictive tools having their say? *J. Mol. Model.* 26 (6): 120.
- 37 Hajduk, P.J., Huth, J.R., and Tse, C. (2005). Predicting protein druggability. *Drug Discov. Today* 10 (23–24): 1675–1682.
- 38 Owens, J. (2007). Determining druggability. *Nat. Rev. Drug Discov.* 6: 187.
- 39 Du, X., Li, Y., Xia, Y.L. et al. (2016). Insights into protein-ligand interactions: mechanisms, models, and methods. *Int. J. Mol. Sci.* 17 (2).
- 40 Zhao, J., Cao, Y., and Zhang, L. (2020). Exploring the computational methods for protein-ligand binding site prediction. *Comput. Struct. Biotechnol. J.* 18: 417–426.
- 41 Fauman, E.B., Rai, B.K., and Huang, E.S. (2011). Structure-based druggability assessment—identifying suitable targets for small molecule therapeutics. *Curr. Opin. Chem. Biol.* 15 (4): 463–468.
- 42 Dias, S., Simões, T., Fernandes, F. et al. (2019). CavBench: a benchmark for protein cavity detection methods. *PLoS One* 14 (10): e0223596.
- 43 Xu, Y., Wang, S., Hu, Q. et al. (2018). CavityPlus: a web server for protein cavity detection with pharmacophore modelling, allosteric site identification and covalent ligand binding ability prediction. *Nucleic Acids Res.* 46 (W1): W374–W379.
- 44 Halgren, T.A. (2009). Identifying and characterizing binding sites and assessing druggability. *J. Chem. Inf. Model.* 49 (2): 377–389.
- 45 Hendlich, M., Rippmann, F., and Barnickel, G. (1997). LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graph. Model.* 15 (6): 359–363.
- 46 Kawabata, T. (2010). Detection of multiscale pockets on protein surfaces using mathematical morphology. *Proteins* 78 (5): 1195–1211.
- 47 Simões, T.M.C. and Gomes, A.J.P. (2019). CavVis—A field-of-view geometric algorithm for protein cavity detection. *J. Chem. Inf. Model.* 59 (2): 786–796.

- 48 Krotzky, T., Fober, T., Hullermeier, E., and Klebe, G. (2014). Extended graph-based models for enhanced similarity search in cavbase. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 11 (5): 878–890.
- 49 Krotzky, T., Fober, T., Mernberger, M. et al. (2013). Extended graph-based models for enhanced similarity retrieval in Cavbase. *J. ChemInform.* 11 (5): 878–890.
- 50 Krotzky, T., Grunwald, C., Egerland, U., and Klebe, G. (2015). Large-scale mining for similar protein binding pockets: with RAPMAD retrieval on the fly becomes real. *J. Chem. Inf. Model.* 55 (1): 165–179.
- 51 Kuhn, D., Weskamp, N., Schmitt, S. et al. (2006). From the similarity analysis of protein cavities to the functional classification of protein families using cavbase. *J. Mol. Biol.* 359 (4): 1023–1044.
- 52 Eguida, M. and Rognan, D. (2020). A computer vision approach to align and compare protein cavities: application to fragment-based drug design. *J. Med. Chem.* 63 (13): 7127–7142.
- 53 Ruf, S., Buning, C., Schreuder, H. et al. (2012). Novel β -amino acid derivatives as inhibitors of cathepsin A. *J. Med. Chem.* 55 (17): 7636–7649.
- 54 Liang, S., Thomas, S.E., Chaplin, A.K. et al. (2022). Structural insights into inhibitor regulation of the DNA repair protein DNA-PKcs. *Nature* 601 (7894): 643–648.
- 55 Curran, P.R., Radoux, C.J., Smilova, M.D. et al. (2020). Hotspots API: a Python package for the detection of small molecule binding hotspots and application to structure-based drug design. *J. Chem. Inf. Model.* 60 (4): 1911–1916.
- 56 Smilova, M.D., Curran, P.R., Radoux, C.J. et al. (2022). Fragment hotspot mapping to identify selectivity-determining regions between related proteins. *J. Chem. Inf. Model.* 62 (2): 284–294.
- 57 Duan J, Jiang B, Chen L, Lu Z, Barbosa J PW. US Pat. Appl. 0229084. 2003.
- 58 Elinson, M.N., Ryzhkova, Y.E., Vereshchagin, A.N. et al. (2021). Electrocatalytic multicomponent one-pot approach to tetrahydro-2'H,4H-spiro[benzofuran-2,5'-pyrimidine] scaffold. *J. Heterocyclic Chem.* 58 (7): 1484–1495.
- 59 Hughes, J.P., Rees, S., Kalindjian, S.B., and Philpott, K.L. (2011). Principles of early drug discovery. *Br. J. Pharmacol.* 162 (6): 1239–1249.
- 60 Pinzi, L. and Rastelli, G. (2019). Molecular docking: shifting paradigms in drug discovery. *Int. J. Mol. Sci.* 20 (18): 4331.
- 61 Bender, B.J., Gahbauer, S., Luttens, A. et al. (2021). A practical guide to large-scale docking. *Nat Protoc . Nat. Res.* 16: 4799–4832.
- 62 Stanzione, F., Giangreco, I., and Cole, J.C. (2021). *Chapter Four - Use of Molecular Docking Computational Tools in Drug Discovery* (ed. D.R. Witty and B.B.T.P. Cox) in MC, editors, 273–343. Elsevier.
- 63 Friesner, R.A., Banks, J.L., Murphy, R.B. et al. (2004). Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* 47 (7): 1739–1749.
- 64 McGann, M. (2011). FRED pose prediction and virtual screening accuracy. *J. Chem. Inf. Model.* 51 (3): 578–596.

- 65 Chemical Computing Group ULC, 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7 2022. Molecular Operating Environment (MOE), 2022.02. 2022.
- 66 Abagyan, R., Totrov, M., and Kuznetsov, D. (1994). ICM—a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.* 15 (5): 488–506.
- 67 Morris, G.M., Goodsell, D.S., Huey, R., and Olson, A.J. (1996). Distributed automated docking of flexible ligands to proteins: parallel applications of AutoDock 2.4. *J. Comput. Aided Mol. Des.* 10 (4): 293–304.
- 68 Jones, G., Willett, P., Glen, R.C. et al. (1997). Development and validation of a genetic algorithm for flexible docking1. *J. Mol. Biol.* 267 (3): 727–748.
- 69 Pagadala, N.S., Syed, K., and Tuszynski, J. (2017). Software for molecular docking: a review. *Biophys. Rev.* 9 (2): 91–102.
- 70 McInnes, C. (2007). Virtual screening strategies in drug discovery. *Curr. Opin. Chem. Biol.* 11 (5): 494–502.
- 71 Acharya, C., Coop, A., Polli, J.E., and Mackerell, A.D.J. (2011). Recent advances in ligand-based drug design: relevance and utility of the conformationally sampled pharmacophore approach. *Curr. Comput. Aided Drug Des.* 7 (1): 10–22.
- 72 Diller, D.J. and Merz, K.M.J. (2002). Can we separate active from inactive conformations? *J. Comput. Aided Mol. Des.* 16 (2): 105–112.
- 73 Liebeschuetz, J.W. (2021). The good, the bad, and the twisted revisited: an analysis of ligand geometry in highly resolved protein-ligand X-ray structures. *J. Med. Chem.* 64 (11): 7533–7543.
- 74 Chen, I.J. and Foloppe, N. (2010). Drug-like bioactive structures and conformational coverage with the LigPrep/ConfGen suite: comparison to programs MOE and catalyst. *J. Chem. Inf. Model.* 50 (5): 822–839.
- 75 Smart, O.S., Horský, V., Gore, S. et al. (2018). Validation of ligands in macromolecular structures determined by X-ray crystallography. *Acta Crystallogr. D: Struct. Biol.* 74: 228–236.
- 76 de Oliveira, M.D., de Araújo, J.O., JMP, G. et al. (2020). Targeting shikimate pathway: in silico analysis of phosphoenolpyruvate derivatives as inhibitors of EPSP synthase and DAHP synthase. *J. Mol. Graph. Model.* 101: 107735.
- 77 Manetti D, Garifulina A, Bartolucci G, Bazzicalupi C, Bellucci C, Chiaramonte N, et al. *New Rigid Nicotine Analogues, Carrying a Norbornane Moiety, Are Potent Agonists of $\alpha 7$ and $\alpha 3^*$ Nicotinic Receptors.* 2019 62 (4): 1887-1901.
- 78 Guandalini, L., Martini, E., Dei, S. et al. (2005). Design of novel nicotinic ligands through 3D database searching. *Bioorg. Med. Chem.* 13 (3): 799–807.
- 79 Iwamura, R., Tanaka, M., Okanari, E. et al. (2018). Identification of a selective, non-Prostanoid EP2 receptor agonist for the treatment of Glaucoma: Omidenepag and its prodrug Omidenepag isopropyl. *J. Med. Chem.* 61 (15): 6869–6891.
- 80 Paralkar, V.M., Borovecki, F., Ke, H.Z. et al. (2003). An EP2 receptor-selective prostaglandin E_2 agonist induces bone healing. *Proc. Natl. Acad. Sci.* 100 (11): 6736–6740.

- 81 Leelananda, S.P. and Lindert, S. (2016). Computational methods in drug discovery. *Beilstein J. Org. Chem.* 12: 2694–2718.
- 82 Konstantinidou, M., Boiarska, Z., Butera, R. et al. (2020). Diaminoimidazopyrimidines: access via the Groebke-Blackburn-Bienaymé reaction and structural data mining. *Eur. J. Org. Chem.* 2020: 5601–5605.
- 83 Vazquez, J., Deplano, A., Herrero, A. et al. (2020). Assessing the performance of mixed strategies to combine lipophilic molecular similarity and docking in virtual screening. *J. Chem. Inf. Model.* 60 (9): 4231–4245.
- 84 Ashenden, S.K. (2021). Chapter 6 - Lead optimization. In: *The Era of Artificial Intelligence, Machine Learning, and Data Science in the Pharmaceutical Industry* (ed. S.K. Ashenden), 103–117. Academic Press.
- 85 Ashenden, S.K. (2021). *Chapter 6: Lead Optimization - the Era of Artificial Intelligence, Machine Learning, and Data Science in the Pharmaceutical Industry*. Oreilly.
- 86 Wesolowski, S.S. and Brown, D.G. (2016). The strategies and politics of successful design, make, test, and analyze (DMTA). In: *Cycles in Lead Generation*, 487–512.
- 87 Subbaiah, M.A.M. and Meanwell, N.A. (2021). Bioisosteres of the phenyl ring: recent strategic applications in Lead optimization and drug design. *J. Med. Chem.* 64 (19): 14046–14128.
- 88 Knegtel, R.M.A., Kuntz, I.D., and Oshiro, C.M. (1997). Molecular docking to ensembles of protein structures1. *J. Mol. Biol.* 266 (2): 424–440.
- 89 Axford, J., Sung, M.J., Manchester, J. et al. (2021). Use of intramolecular 1,5-sulfur–oxygen and 1,5-sulfur–halogen interactions in the design of N-methyl-5-aryl-N-(2,2,6,6-tetramethylpiperidin-4-yl)-1,3,4-thiadiazol-2-amine SMN2 splicing modulators. *J. Med. Chem.* 64 (8): 4744–4761.
- 90 Dennis, A. and Smith, D.S.M. (2012). *Bioisosteres in Medicinal Chemistry*. John Wiley and Sons.
- 91 Thornber, C.W. (1979). Isosterism and molecular modification in drug design. *Chem. Soc. Rev.* 8 (4): 563–580.
- 92 Elliott, J. and Hancock, B. (2006). Pharmaceutical materials science: an active new frontier in materials research. *MRS Bull.* 31 (11): 869–873.
- 93 Chistyakov, D. and Sergeev, G. (2020). The polymorphism of drugs: new approaches to the synthesis of nanostructured polymorphs. *Pharmaceutics* 12 (1): 34.
- 94 Woollam, G.R., Das, P.P., Mugnaioli, E. et al. (2020). Structural analysis of metastable pharmaceutical loratadine form II_{,} by 3D electron diffraction and DFT+D energy minimisation. *CrystEngComm* 22 (43): 7490–7499.
- 95 Wood, P.A., Olsson, T.S.G., Cole, J.C. et al. (2012). Evaluation of molecular crystal structures using full interaction maps. *CrystEngComm* 15 (1): 65–72.
- 96 Galek, P.T.A., Allen, F.H., Fábíán, L., and Feeder, N. (2009). Knowledge-based H-bond prediction to aid experimental polymorph screening. *CrystEngComm* 11 (12): 2634–2639.

- 97 Galek, P.T.A., Chisholm, J.A., Pidcock, E., and Wood, P.A. (2014). Hydrogen-bond coordination in organic crystal structures: statistics, predictions and applications. *Acta Crystallogr. Sect. B: Struct. Sci. Cryst. Eng. Mater.* 70 (Pt 1): 91–105.
- 98 Spackman, M.A. and Jayatilaka, D. (2009). Hirshfeld surface analysis. *CrystEngComm* 11 (1): 19–32.
- 99 Mackenzie, C.F., Spackman, P.R., Jayatilaka, D., and Spackman, M.A. (2017). *CrystalExplorer* model energies and energy frameworks: extension to metal coordination compounds, organic salts, solvates and open-shell systems. *IUCrJ*. 4 (5): 575–587.
- 100 Dávila-Miliani, M.C., Dugarte-Dugarte, A., Toro, R.A. et al. (2020). Polymorphism in the anti-inflammatory drug flunixin and its relationship with Clonixin. *Cryst. Growth Des.* 20 (7): 4657–4666.

Part V

Structure-Based Virtual Screening Using Docking

19

Structure-Based Ultra-Large Virtual Screenings*Christoph Gorgulla^{1,2,3,4}*

¹Harvard University, Harvard Medical School, Biological Chemistry and Molecular Pharmacology,
240 Longwood Ave, Boston, MA 02115, USA

²Harvard University, Physics Department, 17 Oxford Street, Cambridge, MA 02138, USA

³Dana-Farber Cancer Institute, Department of Cancer Biology, 360 Longwood Ave, Boston, MA 02215, USA

⁴St. Jude Children's Research Hospital, Structural Biology Department, 262 Danny Thomas Place, Memphis,
TN 38105, USA

19.1 Introduction

A pivotal task in early-stage drug discovery is the identification and optimization of hit and lead molecules. A hit compound is a molecule that showed a sufficiently strong (experimental or predicted) binding affinity in the screen, while a lead compound is a molecule that binds to the target protein, which is chosen to be further optimized, e.g. via medicinal chemistry. Experimental high-throughput screenings (HTSs), in which typically a few hundred thousand molecules are screened, have been the workhorse for discovering initial hit compounds in the past few decades. Despite its central role in the past, this technique has substantial disadvantages and limitations. HTSs are not cheap due to the reagents, supplies, and sophisticated machines required to test a large number of ligands in an automated way. Despite the automation with robots, HTSs are often still time-intensive due to the necessary preparation of the ligand libraries as well as the binding assays. Another problem is that the potency of initial hit compounds discovered by HTSs is often not high, implying that substantial medicinal chemistry is needed to improve the binding strengths of initial hits. If initial hits are discovered with HTS, the number of hits/scaffolds is often quite limited. For challenging target sites such as flat binding surfaces of protein–protein interactions, it is often not possible to find sufficiently strong initial hits at all with HTSs. Another challenge is the circumstance that initial HTS hits might be binding to other target proteins as well because in HTS assays, compounds are typically not tested for specificity. Last but not least, HTSs do not provide direct mechanistic insight on how or to which target receptor the hit compounds actually bind.

One approach that can solve most of these problems is the structure-based ultra-large virtual screening (ULVS) approach. ULVSs are virtual screens in which 100 million or more ligands are screened computationally. ULVSs can be

relatively time- and cost-efficient compared to experimental HTSs when using high-performance computer clusters or the cloud, e.g. within a few days when sufficient resources are available. ULVS are also able to discover highly potent initial hit compounds, often in the submicromolar range for most types of target sites, and the hits contain a relatively large amount of compounds with a wide choice of different scaffolds and thus backup molecules. The high potency in return allows targeting more challenging target sites, such as protein-protein interactions. In addition, structure-based ULVSs provide mechanistic predictions on how the ligand is binding to the target protein.

In this chapter, Section 19.2 provides an introduction to the fundamentals of structure-based virtual screens. Section 19.3 contains an overview of currently available ultra-large ligand libraries. Subsequently, different flavors of ULVSs are explored. Section 19.4 describes standard/docking-based ULVSs, Section 19.5 synthon-based ULVS, and Section 19.6 machine learning-based ULVS. Finally, we briefly review some other acceleration methods that can be used in combination with these approaches in Section 19.7. An overview of the topics in this book and their relations is shown in Figure 19.1.

19.2 Fundamentals

In this section, a brief introduction to several fundamental concepts that are important in structure-based virtual screenings is provided. Among them are virtual screenings themselves (Section 19.2.4), molecular dockings (Section 19.2.3), receptor preparation (Section 19.2.1), and the preparation of ligand libraries (Section 19.2.2).

19.2.1 Receptor Structures and Preparation

Before structure-based virtual screens can be carried out, the receptor and ligand structures have to be prepared. Regarding the receptor structure, the coordinates are typically obtained via structure databases, such as the Protein Data Bank (PDB, <http://www.wwpdb.org/>) or the AlphaFold Structure Database [1, 2]. The PDB traditionally contains all the experimentally determined protein structures. The AlphaFold Structure Database contains predicted structures by AlphaFold [3, 4] of almost any known protein. An alternative to these two sources is to use a homology model. After a starting structure is obtained, conformational sampling can be carried out via molecular dynamics (MD) or Monte Carlo (MC) simulations if desired to obtain information on the dynamics of the receptor, possibly physiologically more relevant coordinates of the target structure, or multiple conformations of the receptor that can be used in ensemble docking procedures. More details on receptor preparation can be found in [5, 6].

19.2.2 Ligand Preparation and Ligand Libraries

In virtual screenings, large collections of ligands are screened, called ligand libraries. Most docking programs require that the ligand be in a ready-to-dock format.

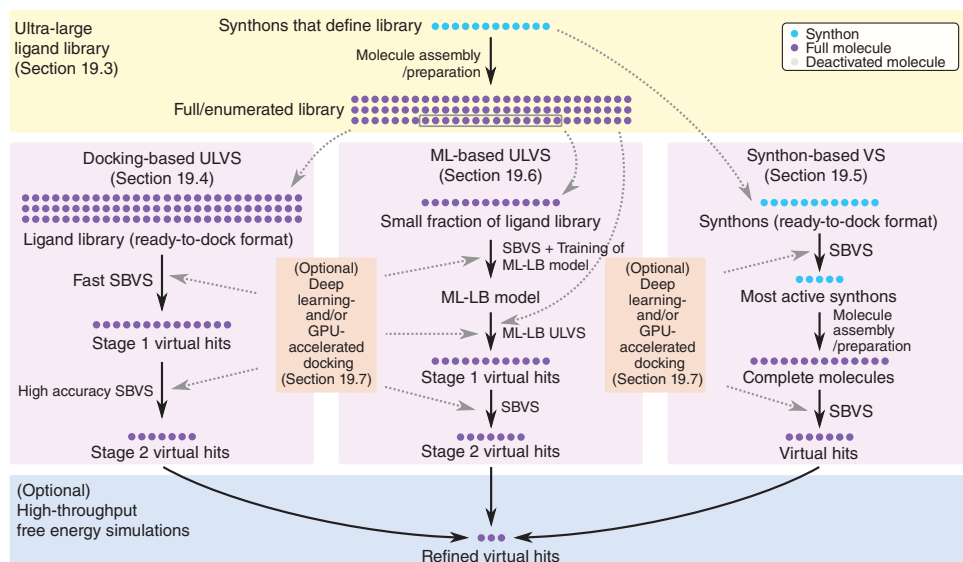


Figure 19.1 Conceptual overview of several ultra-large virtual screening approaches. Ultra-large on-demand ligand libraries based on combinatorial chemistry are particularly attractive due to the commercial availability of the compounds and can be used by most ULVS approaches. In this chapter four types of ULVS are reviewed: docking-based (Section 19.4), synthon-based (Section 19.5), as well as ML-based (Section 19.6). To accelerate the dockings themselves, deep learning-based dockings and/or GPU-accelerated dockings can be used in concert with any of the four ULVS approaches. At the end, the virtual screening results can in principle be refined by carrying out high-throughput free energy simulations (optional).

Typically, this means that molecular properties such as the protonation states, tautomerization states, stereoisomers, and 3D conformations of the ligands have to be predicted. The chemical file format in which the ligand has to be stored depends on the docking program that will use the ligands. The most common formats are the MOL2, SDF, PDB, and PDBQT formats. In contrast to structure-based virtual screenings, ligand-based virtual screenings often do not require a 3D structure of the ligand, but rather in-line notation such as the SMILES or SELFIES formats. Some ligand-based approaches use 3D pharmacophore models, and therefore these methods still require the ligand in 3D format. For reviews that discuss ligand preparation in more detail, see [5–7].

19.2.3 Molecular Docking

Molecular docking programs try to predict how strongly a ligand is binding to a target receptor and in which binding mode [8]. To achieve this, docking programs assign a docking score to each ligand and provide the predicted binding poses. The basic input data the docking programs need are typically the receptor structure, the ligand, as well as additional parameters that specify how the docking should be carried out. These parameters can include options such as the exhaustiveness of the conformational search, the size and location of the targeted docking area on the receptor, or the identity of the residues that should be allowed to be flexible. Hundreds of docking programs have been developed in the past [8–11]. Most docking programs have two parts, a search algorithm, and a scoring function. The search algorithm explores the space of possible conformations of the ligand and the protein. The scoring function assigns a score, which correlates with the predicted binding strength of the ligand to the protein. The docking pose with the best docking score is the final predicted docking pose. There exist different types of docking programs. The four primary classes are physics-based, knowledge-based (also called mean force-based), empirical (also called regression-based), and machine learning-based (also called descriptor-based) [12–15]. Physics-based docking programs can use classical force fields, but could in principle also use quantum mechanical molecular models. A subclass of machine learning-based docking methods are deep learning-based docking methods, which we will touch upon in Section 19.7.

19.2.4 Virtual Screenings

Virtual screenings are computational screens in which a large number of ligands are processed, and for each ligand, it is estimated either if the compound binds to a given target receptor (binary classification), or how strongly the ligand is estimated to bind to the target receptor (regression-based task, in which a score is calculated). Receptor structures are most commonly proteins but can be any type of biological macromolecule, such as RNA or DNA. Virtual screens can be ligand-based or structure-based. In ligand-based virtual screens, known binders to the target receptor are used to decide either if and sometimes how strongly newly screened ligands are likely to bind as well. In structure-based virtual screens, the receptor structure is

used instead, and molecular dockings are carried out to assess the binding strength of the screened ligands. Structure-based methods are preferred if reliable structures are available for a given receptor, as they are independent of any known inhibitors, can be more accurate, and are not biased toward known compounds.

Multiple virtual screenings can be combined in a staged manner, where the best virtual hits of the previous screen are screened again in the next stage with higher accuracy or with a different method to improve the reliability of the results. The primary advantage of multi-staged virtual screens is that they can reduce computational costs substantially when compared to screening all compounds with the accuracy of the final stage. Multistaging can be employed in standard/docking-based ULVSs and is an integral part of synthon-based ULVSs as well as ML-based SB-ULVSs.

19.3 Ultra-Large Ligand Libraries

Ultra-large ligand libraries are libraries that contain over 100 million ligands. They can be classified as being public, commercial, or proprietary [16]. Public ligand libraries are freely available, but the molecules are not directly commercially available in general. Commercial libraries on the other hand are ligand libraries of chemical compound vendors that can be purchased from these companies. Commercial libraries are mostly freely available as well. Proprietary libraries are generally not freely available but are kept and used internally by the companies that possess them. Examples of proprietary libraries are Merck's MASSIV [17] and Pfizer's PGVL [18], and additional examples can be found in [16]. The term virtual library can refer to any library that only exists digitally, but not physically. In this section, an overview of currently available ultra-large public and commercial libraries is provided.

19.3.1 Commercial Libraries

One type of freely available ligand library are commercial libraries. Commercial libraries of ultra-large size are based on a smaller number of fragments (typically between 1000 and 1 000 000) that are combined via combinatorial chemistry approaches into full molecules. The compounds mostly do not exist yet and are only synthesized by the chemical suppliers when purchased by the customers. Therefore, these ligand libraries are also called on-demand libraries.

Since combinatorial on-demand libraries are based on a smaller number of fragments, the question arises whether these libraries lead to a chemically diverse space of molecules. Several studies have looked into this aspect and concluded that the resulting chemical spaces are quite diverse [19–21].

19.3.1.1 REAL Database and REAL Space

Enamine was likely the first compound vendor that offered ultra-large on-demand libraries for drug discovery. The REAL Database reached over 100 million molecules

approximately in the year 2014. The initial version in 2007 contained approximately 29 million molecules [22]. In the meantime, the REAL Database contains over five billion molecules [23]. The REAL Database satisfies Veber's rule and Lipinski's rule of five. The chance of successful synthesis is approximately 80%, and the compounds require approximately three to four weeks to be synthesized after ordering from Enamine.

Enamine has a second on-demand ligand library, the REAL Space, that is even larger than the REAL Database, containing over 30 billion molecules [24]. It is very similar in concept and design to the REAL Space. Due to historical and technical reasons, Enamine keeps these two libraries separate. The REAL Space contains druglike molecules with a maximum size of 450 daltons. The synthesis success rate and the shipping time are approximately 80% and three to four weeks, respectively, and thus the same as for the REAL Database. One difference from the REAL Database is that the REAL Space does not strictly comply with Lipinski's rule of five, but still to a large extent [25]. The compliance of a library with the rule of five is not necessarily an advantage, since a substantial number of approved drugs are not fully compliant with the rule of five [26]. The Enamine REAL Space and the REAL Database are not strictly disjoint, with 50%–70% of the REAL Database being contained in the REAL Space. The REAL Space is available via an enumerated SMILES version and BioSolveIT's *infiniSee* software [27]. In addition, it was prepared into a ready-to-dock format by the VirtualFlow team and made freely available (see Section 19.3.1.7).

19.3.1.2 CHEMriya

Otava Chemicals provides an on-demand ultra-large ligand library called CHEMriya [28], containing over 12 billion molecules. It was released in the year 2021 and is based on 33 000 building blocks. The primary means of accessing the library is via BioSolveIT's *infiniSee* software [27], which allows searching the library for similar compounds based on query molecules. The library is currently not available in a ready-to-dock format. The synthesis and shipping time are four to six weeks after ordering.

19.3.1.3 GalaXi

Another company, WuXi AppTec, has the GalaXi library as one of their products [29]. It is an on-demand library with approximately 8 billion molecules [27]. The success rate of synthesis is between 60% and 80% and takes between four and eight weeks. Similarly to CHEMriya, GalaXi can also be explored and searched via *infiniSee* from BioSolveIT [27].

The ultra-large on-demand libraries from Enamine, Otava, and WuXi AppTec contain billions of compounds. A natural question that arises is whether and how much these libraries overlap and contain identical molecules. This question was investigated by Bellmann et al. [30], and this study found that with less than 1% identical molecules, there is no significant overlap between REAL Space from Enamine, CHEMriya from Otava, and GalaXi from WuXi AppTec.

19.3.1.4 eXplore

The largest on-demand space currently available is eXplore from eMolecules, with over 2.8 trillion molecules [27, 31]. The library is based upon readily available building blocks from other compound vendors. In total, a number of 40 proven reactions are used to combine the building blocks into full molecules. The majority of the molecules require only one to two synthesis steps, allowing for efficient production. The synthesis can be carried out in two ways. Either the customer purchases the building blocks from eMolecules and carries out the synthesis themselves. Or eMolecules takes care of both the purchase of the building blocks and the synthesis as well. Similar to the previous on-demand libraries, eXplore can also be accessed via *infiniSee* from BioSolveIT [27].

19.3.1.5 Freedom Space

Another recent on-demand space is the Freedom Space from Chem-Space, located in Ukraine [27, 32]. This library is based on fragments from various compound vendors as well as known reactions. The library consists of 201 million compounds, 74% of which comply with Lipinski's rule of five, and is based on fragments from various compound vendors. The synthesis is carried out by contract research organizations, and the delivery time is approximately four weeks. The Freedom Space can be explored via *infiniSee* from BioSolveIT [27] and can be downloaded in SMILES format from the webpage of Chem-Space (<https://chem-space.com/compounds/freedom-space>).

19.3.1.6 ZINC Libraries

There exist a large number of compound vendors that offer smaller libraries for drug discovery, including HTS libraries. The sizes of these libraries are mostly in the range of 100 000–5 000 000 molecules. Dealing with all these libraries individually is very complicated and time-consuming, e.g. because each of them can be in a different format and due to the large number of them that are provided by different entities. The ZINC library addresses this problem by providing a unified platform that provides the molecules of hundreds of compound vendors by aggregating the individual vendor libraries. Furthermore, it provides many of these molecules in a ready-to-dock format for virtual screening purposes. For each molecule, the available compound vendors are listed, among other properties. The first version of ZINC became available in 2005, and subsequent versions followed in 2012, 2015, and 2020 [19, 33–35]. The 2015 version was the first to exceed 100 million molecules, thus reaching the ultra-large scale [35]. The latest version (2020) contains approximately 1.5 billion molecules, out of which approximately half are available in a ready-to-dock format at the time of writing (November 2022).

19.3.1.7 VirtualFlow Libraries

Many of the libraries described earlier in this section are not, or only partially, available in a ready-to-dock format. The VirtualFlow open-source project makes some of them available in a ready-to-dock format, and in addition, provides an ULVS platform. The prepared libraries can be accessed via the project homepage,

<https://virtual-flow.org/> [36]. One of the available libraries is the REAL Database from Enamine of the year 2018, consisting of over 1.4 billion molecules in a ready-to-dock format. This library was historically the first ready-to-dock library freely available, with over one billion molecules. The VirtualFlow Project also provides the ZINC15 library (version of the year 2018), containing over 1.4 billion molecules, in a ready-to-dock format. Most recently, the team made available the REAL Space of Enamine (version of 2022), containing over 68 billion stereoisomers. This is the largest ready-to-dock library available to date by an order of magnitude.

19.3.2 Public Virtual Libraries

Another type of freely available ultra-large ligand libraries are publicly available libraries, which are not generally commercially available. The main advantage of these libraries is that they allow consideration of the chemical space beyond what is currently commercially available, and thus remove the limits on what can be explored via commercial libraries. The major disadvantage is that these compounds cannot be readily purchased in general in a time- and cost-efficient way, even though custom synthesis can still be possible for many of them. Custom synthesis however is in most cases considerably more demanding in time and cost compared to synthesizing/purchasing compounds from commercial libraries. Two public ligand libraries of ultra-large size are described below.

19.3.2.1 Generated Databases (GDBs)

The GDBs were likely the first public ultra-large ligand libraries [37, 38]. The different versions of the GDBs are characterized by the number of certain heavy atoms present in the molecules of the corresponding library. The GDB-11, for instance, consists only of molecules with a maximum of 11 atoms of types C, O, N, and F. Besides the GDB-11, there exist the GDB-13 and the GDB-17 libraries. During the generation of these libraries, certain rules were followed regarding the chemical stability, the valency of the atoms, and synthetic feasibility. The GDB-11 contains approximately 110 million molecules (counting stereoisomers), while the GDB-13 contains approximately 977 million compounds, and the GDB-17 counts 166 billion molecules [39, 40]. These libraries are available in the SMILES format, and thus not in a ready-to-dock format for structure-based virtual screenings. To address the issue that the compounds in the GDBs are not commercially available and can be difficult to synthesize, the GDBChEMBL was made available [41]. The GDBChEMBL contains only molecules of the GDB-17 that are similar to compounds in ChEMBL. However, it only contains 10 million molecules and is therefore not of ultra-large size. Another library, the GDBMedChem, was designed, which consists of 10 million molecules that have favorable medicinal chemistry properties [42]. The GDBs have recently been explored by deep learning approaches [43].

19.3.2.2 KnowledgeSpace

The KnowledgeSpace is similar in concept to the eXplore library from eMolecules and the Freedom Space by Enamine, in that it is based on commercially available

building blocks [27, 44]. However, there are fewer restrictions on the chemical reactions required to synthesize the molecules, as well as fewer restrictions on the availability of the building blocks. On the one hand, this results in a vastly bigger space, reaching $2.9 * 10^{14}$ molecules, and with that becoming the largest freely available library for drug discovery. On the other side, the compounds are not readily commercially available and can be hard to synthesize. The library can be explored and searched via *infiniSee* from *BioSolveIT* [27].

19.4 Docking-Based Ultra-Large Virtual Screenings

When screening ultra-large ligand libraries, the most direct way is to dock all ligands that should be screened. This approach, referred to as docking-based ULVSs in this text, was historically the first type of ULVS that was carried out [36, 45–47]. Its main advantages are simplicity and thoroughness, while its major disadvantage is the high computational cost.

ULVSs became possible due to advancements in different fields and technologies. First, commercially available ligand libraries became large enough to reach the ultra-large scale, a development that was led by Enamine with the *REAL Database*. Second, computational resources became powerful and widely available enough to scientists working in drug discovery, in the form of on-premises computer clusters as well as via the cloud (e.g. the Google Cloud or AWS Cloud). Third, the required software platforms became available, such as the open-source virtual screening platform *VirtualFlow* [36]. Last but not least, an abundance of new protein structures has appeared, driven by cryo-electron microscopy (cryo-EM) as well as the *AlphaFold* structure prediction technique.

19.4.1 Success Stories

Docking-based ULVSs had remarkable success stories in the past few years. In this section, we summarize most of these success stories (an overview can be found in Table 19.1).

19.4.1.1 Gorgulla (2018)

The first ultra-large virtual screens were likely reported in [48], where six ultra-large virtual screens were described. Three of these targeted EBP1, a protein that plays a role in certain types of cancer, each involving 130 million ligands. Another target was the peptidoglycan, which was targeted with 130 million ligands as well. The KIX domain of SREBP, a protein involved in cancer and obesity, was targeted with 180 million compounds. Finally, the protein KEAP1 (Kelch-like ECH-associated protein 1) was targeted with 300 million molecules (for more details, see further below). There was no experimental validation involved at this point, except for minimal validation of the screens related to EBP1. To make these screens possible, the author developed the virtual screening platform *VirtualFlow*, the first platform of this kind. *VirtualFlow* was used at first in this thesis to prepare the entire

Table 19.1 Overview of ultra-large virtual screens with experimental validation.

Target	Target site	Compounds screened	Compounds tested	True hit rate	Reference
EBP1	Protein–protein interfaces	3 × 130 million	12	33%	Gorgulla [48]
Dopamine receptor D4	Orthosteric site	138 million	549	24%	Lyu et al. [47]
AmpC β -lactamases	Active site	99 million	51	11%	Lyu et al. [47]
MT ₁ receptor	Orthosteric site	150 million	38	39%	Stein et al. [49]
KEAP1	NRF2 interaction site	1.3 billion	590	11%	Gorgulla et al. [36]
Sigma-2 Receptor	Membrane ligand pocket	490 million	484	26%	Alon et al. [50]
5-HT _{2A}	Orthosteric site	75 million	17	24%	Kaplan et al. [51]
α 2A adrenergic receptor	Orthosteric site	301 million	48	63%	Fink et al. [52]
M _{pro}	Active site	235 million	100	19%	Luttens et al. [53]
SARS-CoV-2 Macrodomain	Active site	400 million	124	40%	Gahbauer et al. [54]

ZINC15 library into a ready-to-dock format, resulting in the first ultra-large libraries ready for structure-based virtual screens. Two versions of the ZINC15 library were prepared: the 2014 pre-publication version containing 130 million ligands and the 2016 version containing approximately 300 million ligands. These two libraries were subsequently deployed in the above-mentioned screenings. For more details on VirtualFlow, see Section 19.4.2.2. In this dissertation, it was also shown that the true hit rate (experimentally confirmed hits divided by the total number of compounds tested) increases with the scale of the screen. This is significant since virtual screenings have historically suffered from low true hit rates.

19.4.1.2 Lyu et al. (2019)

The first peer-reviewed publication about ultra-large virtual screens that was published appeared in 2019 [47]. In this study, two proteins were targeted, using the docking program DOCK [55]. The first target was the enzyme AmpC β -lactamase, which was targeted with 99 million compounds from the ZINC library. The second target was the D4 dopamine receptor, which is a G-protein-coupled receptor (GPCR) that was targeted with 138 million ligands from the ZINC library. The study contained thorough experimental validation, and the authors have shown that highly potent hit and lead compounds were discovered, up to the picomolar range for the D4 receptor. For AmpC, the initial hits were optimized by searching the library for similar compounds, docking them, and subsequently testing the best analogs, resulting in significantly improved binding affinities. In this study, it was

moreover experimentally shown that the true hit rate improves with the docking score (and thus the scale of the screen). This observation confirms the earlier theoretical prediction of this circumstance that the author reported in [48].

19.4.1.3 Stein et al. (2020)

The second ultra-large virtual screen involving extensive experimental validation was published in 2020 [49], involving animal studies for the first time in the context of ultra-large virtual screens. In this study, the Melatonin receptor MT₁ was targeted with approximately 150 million molecules obtained from the ZINC 15 library using the program DOCK [55]. MT₁ is a GPCR that plays a role in the circadian rhythm. The best experimental compounds had binding affinities in the picomolar range, which is remarkable. The best compounds were optimized, and two inverse agonists were discovered. Animal studies involving mice have demonstrated that the compounds exhibit biological activity by altering the circadian rhythm of the mice and that the compounds exert their activity via a target-specific mechanism.

19.4.1.4 Gorgulla et al. (2020)

The first virtual screening platform dedicated to ultra-large virtual screens, VirtualFlow, was published in 2020 [36]. In this study, the first ready-to-dock library with over 1 billion compounds (the REAL Database of Enamine of the year 2018) was made freely available, containing over 1.4 billion molecules (see also Section 19.3.2). More details about VirtualFlow can be found in Section 19.4.2.2. VirtualFlow was applied to the KEAP1-NRF2 interaction, which was the first protein–protein interaction successfully targeted with ultra-large virtual screens. This screen, which involved 1.3 billion molecules, was also the first screen to exceed one billion molecules. The screen was carried out in a multi-staged way, in which 1.3 billion molecules were screened in the first stage with a rigid protein structure, and the highest-scoring three million compounds of stage 1 were rescreened in a second stage where the amino acid residues were allowed to be flexible at the target site. Several hundred virtual hits were experimentally verified, and the most potent compounds had binding affinities in the low nanomolar range and had the ability to displace the NRF₂ peptide from KEAP₁.

19.4.1.5 Alon et al. (2021)

Another GPCR, the σ^2 receptor was targeted in a drug discovery project involving ultra-large virtual screens [50]. 490 million ligands were screened using DOCK [55]. Experimental validation of 484 molecules led to 127 hits in the nanomolar range, 31 of which had a binding affinity below 50 nM. Several compounds were selected for optimization regarding the binding affinity and specificity, which led to substantial further improvements. The observation in [47] that the experimental true hit rate improves with the docking score was confirmed in this study.

19.4.1.6 Other Success Stories

In [51], with the 5-HT_{2A} receptor one more GPCR was successfully targeted in an ULVS campaign. In this study, 75 million compounds from the ZINC library were screened using the program DOCK [55]. Experimental validation led to hits in the

low micromolar range. The compounds were further optimized, and cryo-EM structures of the ligand–protein complexes were reported that showed that the predicted binding poses by the program DOCK were accurate.

The α_{2A} adrenergic receptor was targeted with 301 million compounds from the ZINC20 library with the program DOCK [52]. Experimental validation led to 12 molecules with binding affinities in the submicromolar range. Cryo-EM structures confirmed also in this case that the predicted binding poses of two compounds were roughly correct. These complex structures were subsequently used in an optimization campaign to improve the compounds further.

Two SARS-CoV-2 proteins, M_{pro} and the nsp3 macrodomain were also successfully targeted with ULVSs, both using DOCK 3.7. Approximately 235 million compounds were docked against the nsp3 macrodomain, and 100 compounds were experimentally validated, leading to 19 confirmed hits, corresponding to a 19% true hit rate [53]. The experimental validation included cell-based assays and cocrystal structures of protein–ligand complexes, matching the predicted docking poses. The hits were subsequently optimized, leading to inhibitors with submicromolar potency.

The nsp3 macrodomain was targeted with approximately 400 million compounds [54]. 124 compounds were experimentally validated, leading to 50 experimentally confirmed compounds. For 47 of these compounds, co-crystal structures were obtained.

19.4.2 Available Software for ULVSs

While conceptually simple, docking-based ULVS pose a substantial challenge regarding their actual application in drug discovery projects because of the extremely large number of ligands that have to be docked. This means that an enormous number of input and output files are involved and need to be handled in an automatic manner. In addition, the computations have to be parallelized on a large number of instances, typically using central processing units (CPUs), but since recently also graphics processing units (GPUs) can be used (see also Section 19.7).

To carry out docking-based ULVSs, two possible ways are to use docking programs in an ad hoc manner by writing some custom scripts that try to deal with the workflow and the parallelization. However, this is typically time-consuming, labor-intensive, and complex and requires advanced programming skills. Alternatively, one can use virtual screening platforms that are dedicated to ultra-large virtual screens.

19.4.2.1 UCSF DOCK

The program DOCK is one of the docking programs out there with the longest history. DOCK is developed and maintained at USCF and has many advanced features. It is primarily a docking program rather than a platform for ULVSs, and thus using DOCK in ULVSs directly can be challenging. However, it is possible, and several of the early success stories of ULVSs that were mentioned above in Section 19.4.1 used DOCK 3.7. Furthermore, a detailed protocol was published that describes how DOCK can be used in large-scale docking campaigns [56].

19.4.2.2 VirtualFlow

The first software platform specialized in ULVSs was VirtualFlow [36, 48], which is freely available to anyone. The project is an active open-source project (project website: <https://virtual-flow.org/>), and anyone is welcome to participate in its further development. VirtualFlow consists of two modules. The first module is called VFPL (VirtualFlow for Ligand Preparation), which is dedicated to preparing ultra-large ligand libraries. The second module, VFVS (VirtualFlow for Virtual Screening), is dedicated to carrying out the virtual screening procedure itself. In addition, the project provides via its homepage several ligand libraries that were prepared with VFPL, which can be readily used with VFVS. For details on the available libraries, see Section 19.3.2.

VFPL includes a variety of preparation steps for each ligand, including desalting, neutralization, tautomerization, protonation, stereoisomer enumeration, 3D coordinate calculation, and target-format conversion.

VFVS has many options and is highly flexible in how it can be used. It supports a large number of docking programs (over 40) and can be deployed in single-stage or multi-staged screening campaigns. VFVS can carry out ensemble dockings as well as consensus dockings using multiple docking programs and scoring functions. Protein flexibility is typically included in the second stage, either via ensemble dockings or via side-chain flexibility modeled by the docking programs. Among the supported docking programs are AutoDock Vina 1.12 [57], AutoDock Vina 1.2 [58], Smina [59], QuickVina 2 [60], QuickVina-W [61], Vina-Carb [62], VinaXB [63], GWOVina [64], and PLANTS [65]. Many of these docking programs have special features that can be used within VFVS.

An overview of the workflow with VirtualFlow can be seen in Figure 19.2. In order that VirtualFlow is able to process billions of ligands efficiently, it is able to massively parallelize the calculations using CPUs or GPUs. VirtualFlow uses a perfectly parallel (also called embarrassingly parallel) parallelization strategy to allow a linear scaling behavior with respect to the number of CPUs or GPUs used and was demonstrated to exhibit a linear scaling behavior up to 5.7 million CPUs in the AWS Cloud.

19.5 Synthon-Based Virtual Screenings

One of the major disadvantages of docking-based, ultra-large virtual screens is that they require a large number of computational resources. Approaches that reduce the computational costs while still allowing the exploration of ultra-large compound libraries are therefore desirable. One approach that allows doing that is the synthon-based virtual screening paradigm (see also Figure 19.1). Here, the synthons themselves, which make up the molecules in the library, are prepared for molecular docking and then docked against the target protein. Subsequently, a set of synthons is selected based on the docking scores and possibly other criteria to assemble complete molecules. These complete molecules are then prepared again for molecular docking and docked to the target site/receptor (typically a

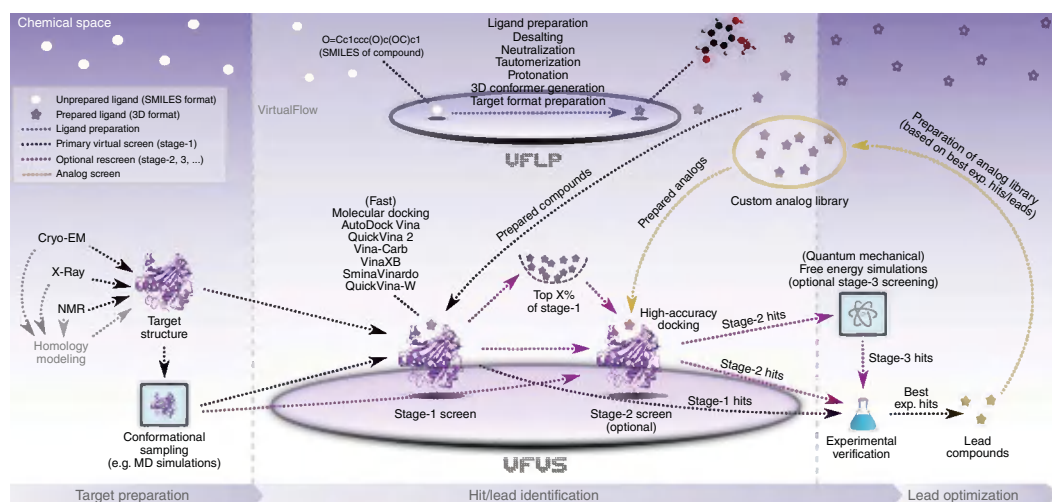


Figure 19.2 Conceptual architecture of the VirtualFlow platform. VirtualFlow consists of two modules : VirtualFlow for Ligand Preparation (VFLP) as well as VirtualFlow for Virtual Screening (VFVS). VFLP is dedicated to preparing ultra-large ligand libraries into a ready-to-dock format. VFVS is specialized in carrying out structure-based virtual screenings and can use the ligand libraries prepared by VFLP. VFVS has a flexible design, can be used to carry out single- as well as multi-staged virtual screenings, and can also be used to optimize hit and lead compounds by screening custom analog libraries.

few million). Not only does this result in a reduction in the computational costs of approximately two orders of magnitude but also reduces the storage requirements by roughly the same factor. The reason is that in synthon-based virtual screening approaches, the library does not have to be completely in a ready-to-dock format, which requires that each molecule is explicitly enumerated and stored in a ready-to-dock format (such as the PDBQT format). Instead, the synthons and the reactions that make up the library are used directly, and only the molecules of interest are enumerated and prepared for molecular docking. This approach is possible with all on-demand libraries based on combinatorial chemistry, such as the REAL Database, the REAL Space, the GalaXi Space, CHEMriya, Freedom Space, and eXplore.

One synthon-based approach is called V-SYNTHES, which was the first to explore the REAL Space of Enamine with experimental validation [66]. In this approach, after the synthons have been docked, the synthons that are selected for the assembly stage are chosen based on their docking scores as well as on a diversity criterion. The diversity criterion allows the inclusion of synthons that use a specific type of reaction only up to 20%. The authors applied this method to discover novel hit compounds to the cannabinoid receptors CB1 and CB2. Subsequent experimental validation led to 14 compounds in the nanomolar range. The best hits were optimized by further virtual screenings, leading to a compound with subnanomolar binding affinity.

A similar approach to V-SYNTHES is Chemical Space Docking (see also Figure 19.3) [67]. Here, the fragments were selected based on the docking score. In their study, the authors used the REAL Space of Enamine (2019 version) to experimentally demonstrate their method and applied it to the protein ROCK1. After the docking of the synthons, the best 500 of them were chosen and assembled into complete molecules. 13 compounds (corresponding to 19%) out of the 69 experimentally tested compounds had submicromolar potencies, with the best compound having a potency of 39 nM.

19.6 Machine Learning-Based Virtual Screenings

An alternative to synthon-based approaches to reduce the costs of ultra-large virtual screens are machine learning-based approaches that combine ligand-based and structure-based virtual screening paradigms. In this context, a relatively small portion of the entire ligand library is initially screened against the target protein using molecular dockings. Based on the docking results, a ligand-based machine learning model is trained on the ligands and the docking scores. This machine learning model subsequently is then used to screen the entire library to identify promising molecules. Finally, a structure-based virtual screen is carried out to dock the best hits of the ligand-based approach to obtain docking poses as well as more reliable docking scores. This approach is considerably faster than docking-based ULVSs because ligand-based screens are orders of magnitude faster. Below, we outline several of the reported machine learning-based screening techniques that were demonstrated on the ultra-large scale.

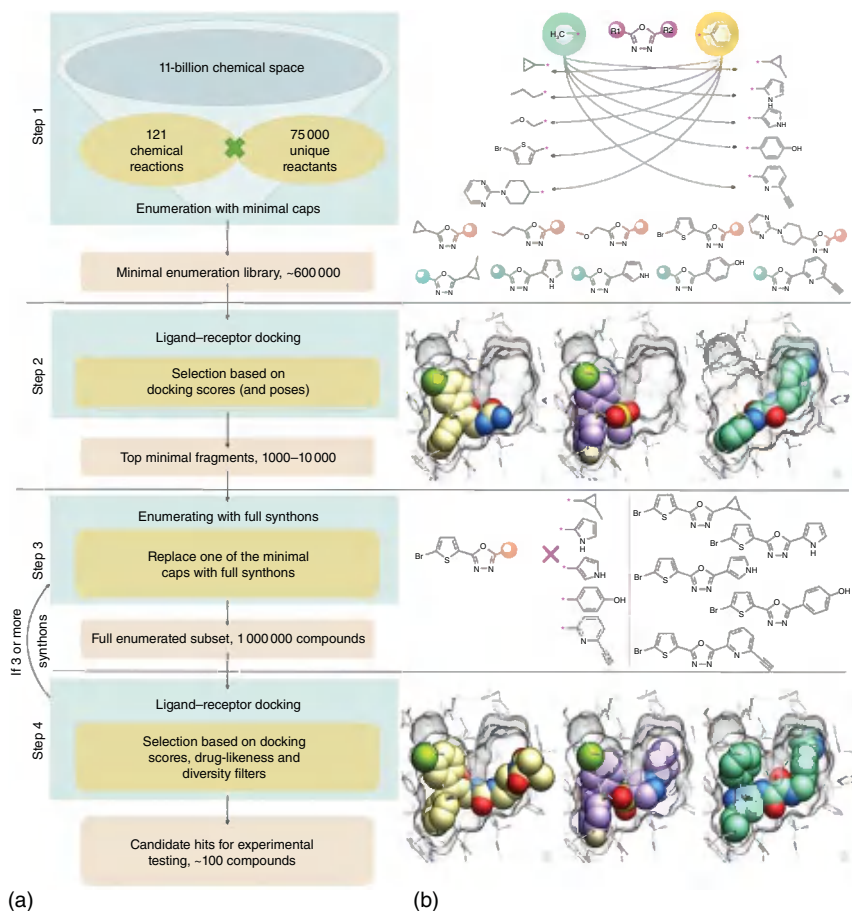


Figure 19.3 Overview and workflow of the V-SYNTHES approach (a), together with examples (b). Source: [66] © 2021, Springer Nature Limited.

19.6.1 Deep Docking

One machine learning-based approach that was carried out on the ultra-large scale is Deep Docking [68]. This method uses a deep learned-based quantitative structure-activity relationship (QSAR) model as the ligand-based machine learning model. To train the model, approximately 1% of the ligand library that is to be screened is used for molecular docking to generate the training data. After the training, the entire library is screened with the deep learning model, and the model is iteratively refined during the screening procedure (see also Figure 19.4). The speedup is approximately 100 compared to standard docking-based ULVSs. A detailed protocol on how to use the method is available in a separate publication [69]. In the original paper, the method was demonstrated with 12 different drug targets. Later, the method was applied to target M_{pro} , the main protease of the SARS-CoV-2 virus, against which over 40 billion compounds were screened [70]. Deep Docking is freely available via

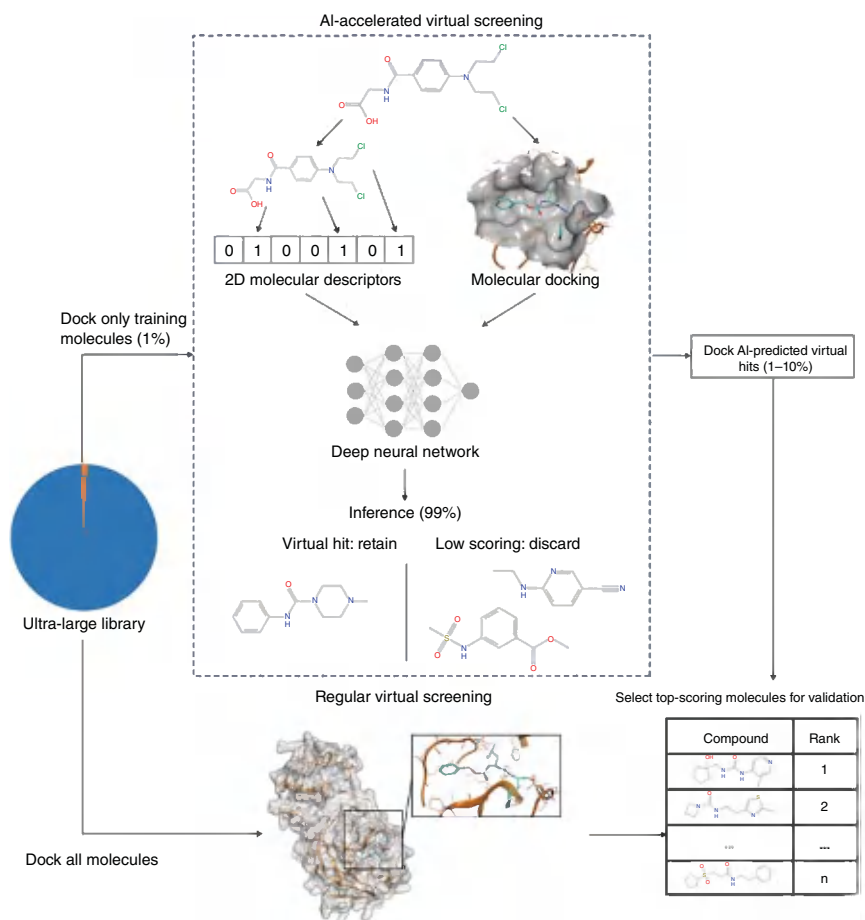


Figure 19.4 Overview of the Deep Docking approach and comparison with docking-based ULVS. Source: [69] © 2022, Springer Nature Limited.

GitHub (<https://github.com/vibudh2209/D2>). A graphical user interface for Deep Docking was made available recently [71].

19.6.2 MolPAL

A different deep learning approach was taken by a method called *Molecular Pool-based Active Learning* (MolPal) [72]. In this method (see also Figure 19.5), molecular fingerprints, Bayesian optimization, and surrogate models are utilized. Surrogate models of different types can be used, and the authors have demonstrated MolPal with random forest, directed-message-passing neural networks, as well as feedforward neural networks. The model is trained in an iterative fashion by docking approximately 2.5% of the ligands of the library that is to be screened. MolPal was applied to the D₄ dopamine receptor as well as AmpC β -lactamase with a ligand library containing 100 million molecules. MolPal was able to recover

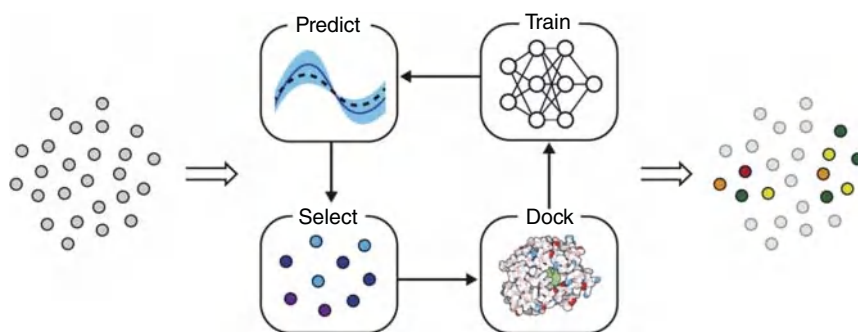


Figure 19.5 Conceptual overview of the MolPal approach. Source: [72] © 2021, Royal Society of Chemistry. CC BY 3.0.

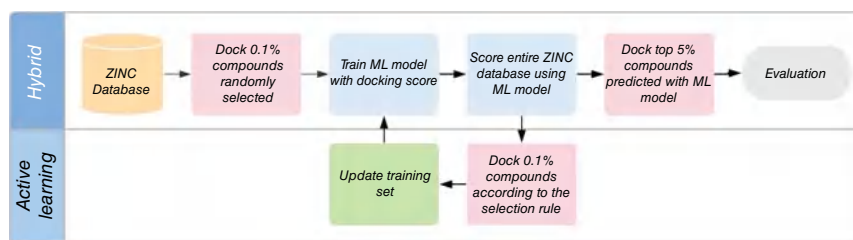


Figure 19.6 Conceptual overview of the AutoQSAR/DeepChem (AQ/DC) approach. Source: [73] © 2021, American Chemical Society.

approximately 90–95% of the top 50,000 virtual hits when docking 2.5% of the entire 100 million compound ligand libraries, resulting in a speedup of approximately 40.

Another machine learning-based approach that is similar to DeepDocking is AutoQSAR/DeepChem (AQ/DC) [73]. Also here, a fraction of the ligand library is initially docked to train a deep learning-based QSAR model utilizing molecular fingerprints (see also Figure 19.6). The deep learning model is based on graph-convolutional neural networks. Active learning can be used as an option to iteratively refine the model during the screening procedure. Overall, approximately 5% of the library is docked, resulting in computational costs of 15–20% relative to docking-based ultra-large virtual screens. The method was demonstrated on AmpC β -lactamase as well as the D4 receptor that was previously targeted by docking-based ultra-large virtual screens (see also Section 19.4.1) [47]. Approximately 80% of the previously experimentally confirmed hit compounds were recovered.

19.7 Other Acceleration Techniques

In this chapter, we have reviewed docking-based ULVSs, as well as several approaches to how ultra-large virtual screens can be accelerated (synthon-based screens and machine learning-based approaches). These acceleration methods take place on the virtual screening level and affect the entire screening workflow. An

alternative to accelerate ULVSs is to accelerate the dockings themselves, either via algorithmic approaches or hardware acceleration. Using accelerated docking methods can not only be used in combination with docking-based ULVSs but also with any of the accelerated virtual screening approaches that we have seen in this chapter (synthon-based screens and machine learning-based approaches), since all of these approaches use external docking programs and can in principle use any available docking method.

19.7.1 Deep Learning Approaches to Molecular Docking

On the algorithmic level, deep learning approaches to molecular docking are very promising due to their fast speed. In addition, deep learning models have the potential to increase the accuracy of the predictions, particularly in the future. Deep learning can be used for different tasks within molecular docking procedures. Deep learning models can be used as scoring functions to predict the binding strength between the ligand and the target protein, which is a regression task. In the context of virtual screenings, scoring functions are sometimes replaced by classification functions, which predict whether a ligand will bind to a protein (sufficiently strong) or not. Most work so far regarding the development of deep learning models for molecular docking has focused on scoring and classification functions. Among the scoring functions are NNScore 2.0 [74], Pafnucy [75], DeepAffinity [76], OnionNet [77], PotentialNet [78], K_{DEEP} [79], DeepAtom [80], DeepBindRG [81], TopologyNet [82], and Math-DL [83], Erdas-Cicek2019 [84], Atomic CNN [85], Francoeur2020 [86], Cang2018 [87], and Zhu2020 [88]. Among the classification functions are NNScore [89], DeepVS [90], Ragoza2017 [91], AtomNet [92], DenseFS [93], Lim2019 [94], Torgn2019 [95], Tanebe2019 [96], Tsubaki2019 [97], Morrone2020 [98], Li2019 [99], Sato2019 [100], Sato2010 [101], BindScope [102], and Lim2019 [94]. The second task that docking programs carry out is pose prediction. First pose prediction methods were published only very recently. Among them are PoseNetDiMa [103], Masters2022 [104], DeepDock [105], EquiBind [106], TANKbind [107], and DiffDock [108]. Regarding the names of all the above methods, the name of the first author was used followed, by the publication year in case no proper name was provided by the authors.

Deep learning-based pose prediction in particular has the potential to reduce the computational costs of molecular dockings because it makes the compute-intensive sampling algorithms obsolete. The resulting acceleration was demonstrated, for example, by TANKbind [107] and EquiBind [106], two approaches that carry out blind dockings. Both of these methods required only a fraction of a second, a one to two orders of magnitude speedup compared to traditional blind docking programs such as QuickVina-W.

19.7.2 GPU Acceleration of Molecular Dockings

Modern GPUs have 10–100 more computational power than CPUs. One common performance measure is floating point operations per second (FLOPS). Traditional docking programs have only used CPUs, but the first docking programs using GPUs have appeared in the past few years. Among them are GNINA [91], MedusaDock

GPU [109], AutoDock GPU [110], or Vina GPU [111]. While these programs exhibit a clear speedup on GPUs, it is not so clear how large the effective cost savings are when using them with GPUs compared to using the CPUs versions. Factors that play a role are the prices of GPUs, as well as the question of how many GPU-based docking instances can be run in parallel per GPU.

19.8 Quality of Ultra-Large Virtual Screening Results

ULVSs represent a major advance in the broader field of computational drug discovery. One interesting aspect regarding ULVSs is how they compare to smaller-scale (traditional) virtual screenings in terms of the novelty of the hits discovered, their diversity, and their true hit rates.

Regarding the novelty of the compounds, ULVSs generally provide equal or higher novelty than traditional virtual screens because they require ultra-large ligand libraries. Such libraries were developed and became only in the past few years, and they provide access to chemical space that was previously mostly unexplored. Traditional virtual screens that screen a smaller library can screen traditional ligand libraries (e.g. experimental HTS libraries) or relatively novel libraries (either smaller novel libraries, or a small part of the novel ultra-large libraries).

Regarding the diversity of the hits, the situation is similar. Small-scale screens can provide diverse results, but ultra-large ligand libraries are able to provide an equal or large diversity, for example, because more scaffolds are needed to build ultra-large libraries. Two recent papers have looked into the diversity of ultra-large ligand libraries and found them to be highly diverse [19–21]. Several recent papers involving ultra-large virtual screens discovered a large number of diverse scaffolds [36, 47, 49–52, 54].

The true hit rates improve with the scale of the screening, as shown theoretically by Gorgulla *et al.* [36, 48] and confirmed experimentally [47, 50]. A recent paper provides additional insights on the effect of the library size on virtual screenings [112]. Several recent papers (see Section 19.4.1) have shown that ULVSs can achieve relatively high true hit rates, mostly between 10% and 40% [36, 47, 49–51, 54], and sometimes even above 60% [52]. These studies have targeted different types of target sites, including active sites, GPCR orthosteric sites, and protein–protein interactions.

19.9 Conclusion and Outlook

ULVSs have only recently been reported, but they have already demonstrated several remarkable successes by identifying highly potent hit and lead compounds. Yet the chemical space that they explore, on the order of billions of molecules, is still vanishingly small when compared to the space of druglike molecules, a space that is estimated to contain more than 10^{60} molecules. ULVSs, in particular their accelerated versions, therefore have much potential to be further improved and will likely play a key role in transforming how drug discovery will be carried out in the future.

References

- 1 Tunyasuvunakool, K., Adler, J., Wu, Z. et al. (2021). Highly accurate protein structure prediction for the human proteome. *Nature* 596: 590–596.
- 2 Varadi, M., Anyango, S., Deshpande, M. et al. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research* 50 (D1): D439–D444.
- 3 Jumper, J., Evans, R., Pritzel, A. et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596 (7873): 583–589.
- 4 Jumper, J. and Hassabis, D. (2022). Protein structure predictions to atomic accuracy with AlphaFold. *Nature Methods* 19 (1): 11–12.
- 5 Madhavi Sastry, G., Adzhigirey, M., Day, T. et al. (2013). Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *Journal of Computer-Aided Molecular Design* 27 (3): 221–234.
- 6 Muegge, I. and Rarey, M. (2001). Small molecule docking and scoring. *Reviews in Computational Chemistry* 17: 1–60.
- 7 DesJarlais, R.L., Cummings, M.D., and Gibbs, A.C. (2007). Virtual docking: how are we doing and how can we improve? In: *Frontiers in Drug Design & Discovery: Structure-Based Drug Design in the 21st Century*, vol. 81, 81–103. Bentham Science Publishers.
- 8 Pagadala, N.S., Syed, K., and Tuszynski, J. (2017). Software for molecular docking: a review. *Biophysical Reviews* 9 (2): 91–102.
- 9 Biesiada, J., Porollo, A., Velayutham, P. et al. (2011). Survey of public domain software for docking simulations and virtual screening. *Human Genomics* 5 (5): 497.
- 10 Fan, J., Fu, A., and Zhang, L. (2019). Progress in molecular docking. *Quantitative Biology* 7 (2): 83–89.
- 11 Sousa, S.F., Fernandes, P.A., and Ramos, M.J. (2006). Protein-ligand docking: current status and future challenges. *Proteins: Structure, Function, and Bioinformatics* 65 (1): 15–26.
- 12 Ain, Q.U., Aleksandrova, A., Roessler, F.D., and Ballester, P.J. (2015). Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. *Wiley Interdisciplinary Reviews: Computational Molecular Science* 5 (6): 405–424.
- 13 Li, J., Fu, A., and Zhang, L. (2019). An overview of scoring functions used for protein–ligand interactions in molecular docking. *Interdisciplinary Sciences: Computational Life Sciences* 11 (2): 320–328.
- 14 Liu, J. and Wang, R. (2015). Classification of current scoring functions. *Journal of Chemical Information and Modeling* 55 (3): 475–482.
- 15 Yang, C., Chen, E.A., and Zhang, Y. (2022). Protein–ligand docking in the machine-learning era. *Molecules* 27 (14): 4568.
- 16 Hoffmann, T. and Gastreich, M. (2019). The next level in chemical space navigation: going far beyond enumerable compound libraries. *Drug Discovery Today* 24 (5): 1148–1156.

- 17 Knehan, T., Klingler, F.-M., Kraut, H. et al. (2017). Merck AcceSSible InVen-tory (MASSIV): in silico synthesis guided by chemical transforms obtained through bootstrapping reaction databases. In: *Abstracts of Papers of the American Chemical Society*, vol. 254. Washington, DC: American Chemical Society.
- 18 Hu, Q., Peng, Z., Sutton, S.C. et al. (2012). Pfizer Global Virtual Library (PGVL): a chemistry design tool powered by experimentally validated parallel synthesis information. *ACS Combinatorial Science* 14 (11): 579–589.
- 19 Irwin, J.J., Tang, K.G., Young, J. et al. (2020). ZINC20 – a free ultralarge-scale chemical database for ligand discovery. *Journal of Chemical Information and Modeling* 60 (12): 6065–6073.
- 20 Tomberg, A. and Boström, J. (2020). Can easy chemistry produce complex, diverse, and novel molecules? *Drug Discovery Today* 25 (12): 1–8.
- 21 Tingle, B., Tang, K., Castanon, J. et al. (2022). Zinc-22 – a free multi-billion-scale database of tangible compounds for ligand discovery. *Journal of Chemical Information and Modeling* 63 (4): 1166–1176.
- 22 Shivanyuk, A.N., Ryabukhin, S.V., Bogolyubsky, A.V. et al. (2007). Enamine real database: making chemical diversity real. *Chemistry Today* 25 (6): 58–59.
- 23 Enamine (2022). REAL Database: the largest enumerated database of syn-thetically feasible molecules. <https://enamine.net/compound-collections/real-compounds/real-database> (accessed 26 August 2023).
- 24 Enamine (2022). REAL Space: billions of make-on-demand molecules. <https://enamine.net/compound-collections/real-compounds/real-space-navigator> (accessed 26 August 2023).
- 25 Grygorenko, O.O., Radchenko, D.S., Dziuba, I. et al. (2020). Generating multi-billion chemical space of readily accessible screening compounds. *iScience* 23 (11): 101681.
- 26 DeGoey, D.A., Chen, H.-J., Cox, P.B., and Wendt, M.D. (2018). Beyond the rule of 5: lessons learned from AbbVie's drugs and compound collection. *Journal of Medicinal Chemistry* 61 (7): 2636–2651. PMID: 28926247.
- 27 BioSolveIT (2022). infiniSee. <https://www.biosolveit.de/infiniSee/>.
- 28 OTAVA (2022). 12 Billion Novel Molecules: CHEMriya – OTAVA's On-Demand Chemical Space. <https://www.otavachemicals.com/products/chemriya> (accessed 28 August 2023).
- 29 WuXi AppTec (2022). GalaXi Space. <https://www.labnetwork.com/frontend-app/p/#/library/virtual> (accessed 26 August 2023).
- 30 Bellmann, L., Penner, P., Gastreich, M., and Rarey, M. (2022). Compar-ison of combinatorial fragment spaces and its application to ultralarge make-on-demand compound catalogs. *Journal of Chemical Information and Modeling* 62 (3): 553–566.
- 31 eMolecules (2022). eXplore. <https://marketing.emolecules.com/explore> (accessed 26 August 2023).
- 32 Chemspace (2022). Freedom Space. <https://chem-space.com/compounds/freedom-space> (accessed 26 August 2023).

- 33 Irwin, J.J. and Shoichet, B.K. (2005). ZINC—a free database of commercially available compounds for virtual screening. *Journal of Chemical Information and Modeling* 45 (1): 177–182.
- 34 Irwin, J.J., Sterling, T., Mysinger, M.M. et al. (2012). ZINC: a free tool to discover chemistry for biology. *Journal of Chemical Information and Modeling* 52 (7): 1757–1768.
- 35 Sterling, T. and Irwin, J.J. (2015). ZINC 15—ligand discovery for everyone. *Journal of Chemical Information and Modeling* 55 (11): 2324–2337.
- 36 Gorgulla, C., Boeszoermenyi, A., Wang, Z.-f. et al. (2020). An open-source drug discovery platform enables ultra-large virtual screens. *Nature* 580 (7805): 663–668.
- 37 Meier, K., Bühlmann, S., Arús-Pous, J., and Reymond, J.-L. (2020). The generated databases (GDBs) as a source of 3D-shaped building blocks for use in medicinal chemistry and drug discovery. *Chimia* 74 (4): 241.
- 38 Reymond, J.L. and Awale, M. (2012). Exploring chemical space for drug discovery using the chemical universe database. *ACS Chemical Neuroscience* 3 (9): 649–657.
- 39 Blum, L.C. and Reymond, J.L. (2009). 970 Million druglike small molecules for virtual screening in the chemical universe database GDB-13. *Journal of the American Chemical Society* 131 (25): 8732–8733.
- 40 Ruddigkeit, L., Van Deursen, R., Blum, L.C., and Reymond, J.L. (2012). Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *Journal of Chemical Information and Modeling* 52 (11): 2864–2875.
- 41 Bühlmann, S. and Reymond, J.-L. (2020). ChEMBL-likeness score and database GDBChEMBL. *Frontiers in Chemistry* 8: 4–10.
- 42 Awale, M., Sirockin, F., Stiefl, N., and Reymond, J.-L. (2019). Medicinal chemistry aware database GDBMedChem. *Molecular Informatics* 38 (8–9): 1900031.
- 43 Arús-Pous, J., Blaschke, T., Ulander, S. et al. (2019). Exploring the GDB-13 chemical space using deep generative models. *Journal of Cheminformatics* 11 (1): 20.
- 44 Detering, C., Claussen, H., Gastreich, M., and Lemmen, C. (2010). KnowledgeSpace – a publicly available virtual chemistry space. *Journal of Cheminformatics* 2 (S1): O9.
- 45 Gorgulla, C. (2022). Recent developments in structure-based virtual screening approaches. *arXiv preprint arXiv:2211.03208*.
- 46 Gorgulla, C., Jayaraj, A., Fackeldey, K., and Arthanari, H. (2022). Emerging frontiers in virtual drug discovery: from quantum mechanical methods to deep learning approaches. *Current Opinion in Chemical Biology* 69: 102156.
- 47 Lyu, J., Wang, S., Balius, T.E. et al. (2019). Ultra-large library docking for discovering new chemotypes. *Nature* 566 (7743): 224–229.
- 48 Gorgulla, C. (2018). Free energy methods involving quantum physics, path integrals, and virtual screenings: development, implementation and application in drug discovery. PhD thesis. Freie Universität Berlin.

- 49 Stein, R.M., Kang, H.J., McCorvy, J.D. et al. (2020). Virtual discovery of melatonin receptor ligands to modulate circadian rhythms. *Nature* 579: 609–614.
- 50 Alon, A., Lyu, J., Braz, J.M. et al. (2021). Structures of the $\sigma 2$ receptor enable docking for bioactive ligand discovery. *Nature* 600 (7890): 759–764.
- 51 Kaplan, A.L., Confair, D.N., Kim, K. et al. (2022). Bespoke library docking for 5-HT_{2A} receptor agonists with antidepressant activity. *Nature* 610 (7932): 582–591.
- 52 Fink, E.A., Xu, J., Hübner, H. et al. (2022). Structure-based discovery of nonopioid analgesics acting through the $\alpha 2A$ -adrenergic receptor. *Science* 377 (6614): eabn7065.
- 53 Lutten, A., Gullberg, H., Abdurakhmanov, E. et al. (2022). Ultralarge virtual screening identifies SARS-CoV-2 main protease inhibitors with broad-spectrum activity against coronaviruses. *Journal of the American Chemical Society* 144 (7): 2905–2920.
- 54 Gahbauer, S., Correy, G.J., Schuller, M. et al. (2023). Iterative computational design and crystallographic screening identifies potent inhibitors targeting the Nsp3 macrodomain of SARS-CoV-2. *Proceedings of the National Academy of Sciences of the United States of America* 120 (2): e2212931120.
- 55 Coleman, R.G., Carchia, M., Sterling, T. et al. (2013). Ligand pose and orientational sampling in molecular docking. *PLoS ONE* 8 (10): 1–19.
- 56 Bender, B.J., Gahbauer, S., Lutten, A. et al. (2021). A practical guide to large-scale docking. *Nature Protocols* 16: 4799–4832.
- 57 Trott, O. and Olson, A.J. (2010). AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry* 31 (2): 455–461.
- 58 Eberhardt, J., Santos-Martins, D., Tillack, A.F., and Forli, S. (2021). AutoDock Vina 1.2. 0: new docking methods, expanded force field, and python bindings. *Journal of Chemical Information and Modeling* 61 (8): 3891–3898.
- 59 Koes, D.R., Baumgartner, M.P., and Camacho, C.J. (2013). Lessons learned in empirical scoring with Smina from the CSAR 2011 benchmarking exercise. *Journal of Chemical Information and Modeling* 53 (8): 1893–1904.
- 60 Alhossary, A., Handoko, S.D., Mu, Y., and Kwoh, C.-K. (2015). Fast, accurate, and reliable molecular docking with QuickVina 2. *Bioinformatics* 31 (13): 2214–2216.
- 61 Hassan, N.M., Alhossary, A.A., Mu, Y., and Kwoh, C.-K. (2017). Protein-ligand blind docking using QuickVina-W with inter-process spatio-temporal integration. *Scientific Reports* 7 (1): 15451.
- 62 Nivedha, A.K., Thieker, D.F., Makeneni, S. et al. (2016). Vina-Carb: improving glycosidic angles during carbohydrate docking. *Journal of Chemical Theory and Computation* 12 (2): 892–901.
- 63 Koebel, M.R., Schmadeke, G., Posner, R.G., and Sirimulla, S. (2016). AutoDock VinaXB: implementation of XBSF, new empirical halogen bond scoring function, into AutoDock Vina. *Journal of Cheminformatics* 8 (1): 27.
- 64 Gorgulla, C., Fackeldey, K., Wagner, G., and Arthanari, H. (2020). Accounting of receptor flexibility in ultra-large virtual screens with VirtualFlow using a

- grey wolf optimization method. *Supercomputing Frontiers and Innovations* 7 (3): 4–12.
- 65 Gorgulla, C., Çınaroğlu, S.S., Fischer, P.D. et al. (2021). VirtualFlow ants—ultra-large virtual screenings with artificial intelligence driven docking algorithm based on ant colony optimization. *International Journal of Molecular Sciences* 22 (11): 5807.
 - 66 Sadybekov, A.A., Sadybekov, A.V., Liu, Y. et al. (2022). Synthon-based ligand discovery in virtual libraries of over 11 billion compounds. *Nature* 601 (7893): 452–459.
 - 67 Beroza, P., Crawford, J.J., Ganichkin, O. et al. (2022). Chemical space docking enables large-scale structure-based virtual screening to discover ROCK1 kinase inhibitors. *Nature Communications* 13 (1): 1–10.
 - 68 Gentile, F., Agrawal, V., Hsing, M. et al. (2020). Deep Docking: a deep learning platform for augmentation of structure based drug discovery. *ACS Central Science* 6 (6): 939–949.
 - 69 Gentile, F., Yaacoub, J.C., Gleave, J. et al. (2022). Artificial intelligence - enabled virtual screening of ultra-large chemical libraries with deep docking. *Nature Protocols* 17 (3): 672–697.
 - 70 Ton, A.-T., Gentile, F., Hsing, M. et al. (2020). Rapid identification of potential inhibitors of SARS-CoV-2 main protease by deep docking of 1.3 billion compounds. *Molecular Informatics* 39 (8): 2000028.
 - 71 Yaacoub, J.C., Gleave, J., Gentile, F. et al. (2021). DD-GUI: a graphical user interface for deep learning-accelerated virtual screening of large chemical libraries (Deep Docking). *Bioinformatics* 38 (4): 1146–1148.
 - 72 Graff, D.E., Shakhnovich, E.I., and Coley, C.W. (2021). Accelerating high-throughput virtual screening through molecular pool-based active learning. *Chemical Science* 12 (22): 7866–7881.
 - 73 Yang, Y., Yao, K., Repasky, M.P. et al. (2021). Efficient exploration of chemical space with docking and deep learning. *Journal of Chemical Theory and Computation* 17 (11): 7106–7119.
 - 74 Durrant, J.D. and McCammon, J.A. (2011). NNScore 2.0: a neural-network receptor–ligand scoring function. *Journal of Chemical Information and Modeling* 51 (11): 2897–2903.
 - 75 Stepniewska-Dziubinska, M.M., Zielenkiewicz, P., and Siedlecki, P. (2018). Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. *Bioinformatics* 34 (21): 3666–3674.
 - 76 Karimi, M., Wu, D., Wang, Z., and Shen, Y. (2019). DeepAffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks. *Bioinformatics* 35 (18): 3329–3338.
 - 77 Zheng, L., Fan, J., and Mu, Y. (2019). OnionNet: a multiple-layer intermolecular-contact-based convolutional neural network for protein–ligand binding affinity prediction. *ACS Omega* 4 (14): 15956–15965.
 - 78 Feinberg, E.N., Sur, D., Wu, Z. et al. (2018). PotentialNet for molecular property prediction. *ACS Central Science* 4 (11): 1520–1530.

- 79 Jiménez, J., Skalic, M., Martinez-Rosell, G., and De Fabritiis, G. (2018). K_DEEP: protein–ligand absolute binding affinity prediction via 3D-convolutional neural networks. *Journal of Chemical Information and Modeling* 58 (2): 287–296.
- 80 Li, Y., Rezaei, M.A., Li, C., and Li, X. (2019). DeepAtom: a framework for protein–ligand binding affinity prediction. *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 303–310. IEEE.
- 81 Zhang, H., Liao, L., Saravanan, K.M. et al. (2019). DeepBindRG: a deep learning based method for estimating effective protein–ligand affinity. *PeerJ* 7: e7362.
- 82 Cang, Z. and Wei, G.-W. (2017). TopologyNet: topology based deep convolutional and multi-task neural networks for biomolecular property predictions. *PLoS Computational Biology* 13 (7): e1005690.
- 83 Nguyen, D.D., Gao, K., Wang, M., and Wei, G.-W. (2020). MathDL: mathematical deep learning for D3R Grand Challenge 4. *Journal of Computer-Aided Molecular Design* 34 (2): 131–147.
- 84 Erdas-Cicek, O., Atac, A.O., Gurkan-Alp, A.S. et al. (2019). Three-dimensional analysis of binding sites for predicting binding affinities in drug design. *Journal of Chemical Information and Modeling* 59 (11): 4654–4662.
- 85 Gomes, J., Ramsundar, B., Feinberg, E.N., and Pande, V.S. (2017). Atomic convolutional networks for predicting protein–ligand binding affinity. *arXiv preprint arXiv:1703.10603*.
- 86 Francoeur, P.G., Masuda, T., Sunseri, J. et al. (2020). Three-dimensional convolutional neural networks and a cross-docked data set for structure-based drug design. *Journal of Chemical Information and Modeling* 60 (9): 4200–4215.
- 87 Cang, Z., Mu, L., and Wei, G.-W. (2018). Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening. *PLoS Computational Biology* 14 (1): e1005929.
- 88 Zhu, F., Zhang, X., Allen, J.E. et al. (2020). Binding affinity prediction by pairwise function based on neural network. *Journal of Chemical Information and Modeling* 60 (6): 2766–2772.
- 89 Durrant, J.D. and McCammon, J.A. (2010). NNScore: a neural-network-based scoring function for the characterization of protein–ligand complexes. *Journal of Chemical Information and Modeling* 50 (10): 1865–1871.
- 90 Pereira, J.C., Caffarena, E.R., and Santos, C.N.D. (2016). Boosting docking-based virtual screening with deep learning. *Journal of Chemical Information and Modeling* 56 (12): 2495–2506.
- 91 Ragoza, M., Hochuli, J., Idrobo, E. et al. (2017). Protein–ligand scoring with convolutional neural networks. *Journal of Chemical Information and Modeling* 57 (4): 942–957.
- 92 Wallach, I., Dzamba, M., and Heifets, A. (2015). AtomNet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. *arXiv preprint arXiv:1510.02855*.
- 93 Imrie, F., Bradley, A.R., van der Schaar, M., and Deane, C.M. (2018). Protein family-specific models using deep neural networks and transfer learning

- improve virtual screening and highlight the need for more data. *Journal of Chemical Information and Modeling* 58 (11): 2319–2330.
- 94 Lim, J., Ryu, S., Park, K. et al. (2019). Predicting drug–target interaction using a novel graph neural network with 3D structure-embedded graph representation. *Journal of Chemical Information and Modeling* 59 (9): 3981–3988.
 - 95 Torng, W. and Altman, R.B. (2019). Graph convolutional neural networks for predicting drug–target interactions. *Journal of Chemical Information and Modeling* 59 (10): 4131–4149.
 - 96 Tanebe, T. and Ishida, T. (2019). End-to-end learning based compound activity prediction using binding pocket information. *International Conference on Intelligent Computing*, 226–234. Springer.
 - 97 Tsubaki, M., Tomii, K., and Sese, J. (2019). Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics* 35 (2): 309–318.
 - 98 Morrone, J.A., Weber, J.K., Huynh, T. et al. (2020). Combining docking pose rank and structure with deep learning improves protein–ligand binding mode prediction over a baseline docking approach. *Journal of Chemical Information and Modeling* 60 (9): 4170–4179.
 - 99 Li, F., Wan, X., Xing, J. et al. (2019). Deep neural network classifier for virtual screening inhibitors of (S)-adenosyl-l-methionine (SAM)-dependent methyltransferase family. *Frontiers in Chemistry* 7: 324.
 - 100 Sato, A., Tanimura, N., Honma, T., and Konagaya, A. (2019). Significance of data selection in deep learning for reliable binding mode prediction of ligands in the active site of CYP3A4. *Chemical and Pharmaceutical Bulletin* 67 (11): 1183–1190.
 - 101 Sato, T., Honma, T., and Yokoyama, S. (2010). Combining machine learning and pharmacophore-based interaction fingerprint for in silico screening. *Journal of Chemical Information and Modeling* 50 (1): 170–185.
 - 102 Skalic, M., Martínez-Rosell, G., Jiménez, J., and De Fabritiis, G. (2019). Play-Molecule BindScope: large scale CNN-based virtual screening on the web. *Bioinformatics* 35 (7): 1237–1238.
 - 103 Mahmoud, A.H., Lill, J.F., and Lill, M.A. (2020). Graph-convolution neural network-based flexible docking utilizing coarse-grained distance matrix. *arXiv preprint arXiv:2008.12027*.
 - 104 Masters, M., Mahmoud, A.H., Wei, Y., and Lill, M.A. (2022). Deep learning model for flexible and efficient protein–ligand docking. *ICLR2022 Machine Learning for Drug Discovery*.
 - 105 Liao, Z., You, R., Huang, X. et al. (2019). DeepDock: enhancing ligand–protein interaction prediction by a combination of ligand and structure information. *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 311–317. IEEE.
 - 106 Stärk, H., Ganea, O., Pattanaik, L. et al. (2022). EquiBind: geometric deep learning for drug binding structure prediction. *International Conference on Machine Learning*, 20503–20521. PMLR.

- 107 Lu, W., Wu, Q., Zhang, J. et al. (2022). TANKbind: trigonometry-aware neural networks for drug-protein binding structure prediction. *bioRxiv*.
- 108 Corso, G., Stärk, H., Jing, B. et al. (2022). DiffDock: diffusion steps, twists, and turns for molecular docking. *arXiv preprint arXiv:2210.01776*.
- 109 Fan, M., Wang, J., Jiang, H. et al. (2021). GPU-accelerated flexible molecular docking. *The Journal of Physical Chemistry B* 125 (4): 1049–1060.
- 110 Santos-Martins, D., Solis-Vasquez, L., Tillack, A.F. et al. (2021). Accelerating AutoDock4 with GPUs and gradient-based local search. *Journal of Chemical Theory and Computation* 17 (2): 1060–1073.
- 111 Tang, S., Chen, R., Lin, M. et al. (2022). Accelerating autodock Vina with GPUs. *Molecules* 27 (9): 3041.
- 112 Lyu, J., Irwin, J.J., and Shoichet, B.K. (2023). Modeling the expansion of virtual screening libraries. *Nature Chemical Biology* 19: 712–718.

20

Community Benchmarking Exercises for Docking and Scoring

Bharti Devi, Anurag TK Baidya, and Rajnish Kumar

Department of Pharmaceutical Engineering & Technology, Indian Institute of Technology (B.H.U.), Varanasi, Uttar Pradesh, India

20.1 Introduction

High throughput screening (HTS) is a traditional approach for identifying hit compounds by screening small molecule libraries in isolated protein- or cell-based assays. However, for screening large libraries of millions of compounds, HTS is time-consuming, cost-intensive, and tends to have a high rate of false positives. Following hit identification, lead optimization is a difficult and time-consuming process where multi-parametric optimization of the compounds is performed through several design-make-test-analyze cycles [1]. There has been a paradigm shift from the traditional drug development course with the arrival of supercomputers and advancements in information technology both in the arena of hardware and software, rendering the drug discovery process effective and technology-oriented in the last couple of decades [2]. Thereby, the time- and resource-intensive assays might be replaced with algorithms for the prediction of the activity of ultra-large libraries containing billions of compounds against a specified target, reducing the overall time and cost of the projects [3].

Molecular docking is a fast and inexpensive computational drug design tool that is being extensively used in academic and industrial drug discovery projects for the identification of hit compounds by prioritizing a subset of highly ranked compounds for further *in vitro* screening [4]. The search algorithm produces different poses of the docked molecule, which are further ranked by the convenient scoring function in order to evaluate the strength of the binding interaction between these two partners [5]. Because of the continuous emergence of new docking and scoring methods and tools, their benchmarking is critical to assess their accuracy on the real-world system so as to provide an assessment of the accuracy of these tools to the end users and the developers for decision-making [6].

The main focus of this chapter is on community benchmarking exercises for docking and scoring. Starting with an overview of docking and scoring, followed by a

short section on the need for benchmarking and various available benchmarking tools, we have concluded this chapter with various open challenges published by academic and industrial researchers in the field.

20.1.1 Overview of Molecular Docking

Various binding orientations are being predicted for every docking run sample by allowing some extent of degree of freedom to the ligands that are to be docked (semi-flexible docking), whereas in some other cases, the protein residues within the ligand binding site are subjected to an additional degree of geometric conformational freedom (fully flexible docking) [7]. While the majority of the docking software packages are able to predict the binding pose when compared to the experimental binding pose with reasonable accuracy, the binding affinity predictions useful for ranking the compounds or predicting the absolute binding affinity of the molecules are still a major challenge [8]. The correct tautomeric/protomeric states for the ligands [9], the inclusion of structural water molecules that serve as a platform for hydrogen bond interactions between the target protein and ligand [10], metal coordination [11], and pKa of the titratable residues [12] are some of the additional challenges that need to be considered. Considering the use of diversified approximations for performing docking and scoring functions, benchmarking of docking tools is a crucial part of structure-based drug design and discovery.

In the case of drug action, biomolecular interactions involving the target and its ligand can be referred to as protein–ligand interactions [13]. In-silico molecular docking process is carried out in order to predict the three-dimensional geometrical conformation and orientation within an active target binding site of a protein, thereby generating the optimal binding pose [14] (Fig. 20.1). It has become an important strategy in the drug discovery paradigm that provides vital information necessary for accurate structural modeling and appropriate activity prediction. The first theory on the binding mechanism was given by Emil Fischer, which represented the “lock-key model” based on the complementarity of the native shape of protein and ligand followed by the introduction of the “induced-fit” theory given by Koshland in 1958 to interpret the various conformational degrees of changes in the protein induced by the ligand while binding [15].

Molecular docking is regarded as a multi-step process [16] that starts with the implementation of search algorithms that analyze the geometrical conformations (pose) of small molecules in the binding cavity of the target protein [17]. Due to multiple conformational degrees of freedom of small molecules, their pose prediction is a challenging task. Precise sampling of such degrees of freedom (3 N-6) should be done with sufficient accuracy to identify the conformations that complement the binding pocket in a fast and efficient manner in order to render the evaluation of hundreds of thousands of molecules in a given docking screen. Scoring functions are used to complement the docking study by ranking the conformations with respect to their differences in binding energy.

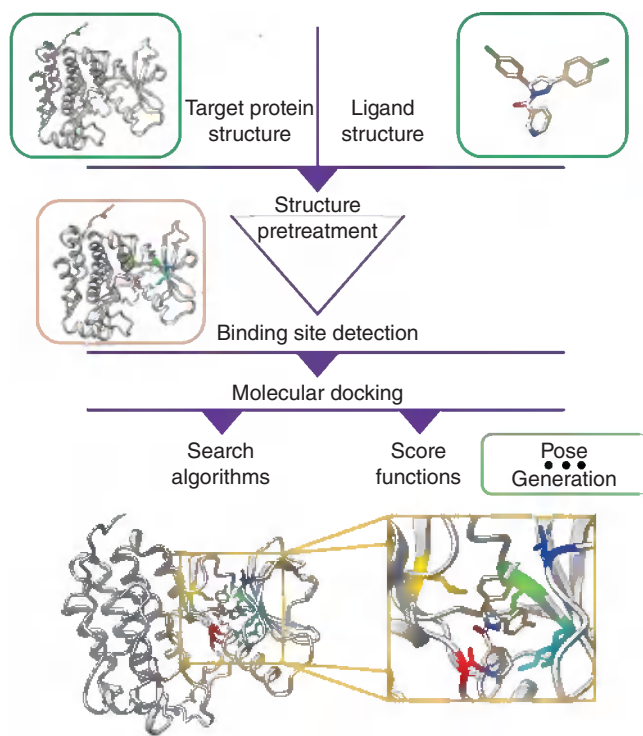


Figure 20.1 Overview of the molecular docking workflow. It generally starts with acquiring the 3D structures of the macromolecular target protein and the ligand to be docked. Followed by the structure pre-treatment in order to make it suitable for the docking workflow. Then the target binding site is detected followed by docking. The calculations are completed in two major steps: posing followed by scoring, thereby generating a list of possible binding modes of the formed complexes between the target protein and ligand.

Inaccurate predictions during the docking of the ligand–protein interaction are assorted as either soft or hard failures. If the search algorithm is not able to pinpoint the global energy minima analogous to the crystal structure, it is regarded as soft failure. Hard failures, on the other hand, are the result of a fault in the energy function due to which the global energy minima is related to the mis-docked pose having energy less than the crystal structure [18, 19].

Docking and scoring have various applications in drug discovery, as follows:

- (a) Structure-based virtual screening (SBVS) for hit identification attempts to predict the compounds with a high likelihood of binding to the target.
- (b) Identifying the best possible mode of interaction between the ligand and the target protein by utilizing scoring functions thereby, playing a major role behind the success of a SBVS tool [20].
- (c) Binding mode prediction: Hypothesis generation for enabling compound design during the lead optimization stage and also providing necessary information for growing fragment hits by predicting the optimal binding pose [21].

- (d) Rationalizing Structure Activity Relationship (SAR) from a structural perspective, predicting the ligand–protein interaction at molecular basis, and thereby laying out the SAR [22].
- (e) Prioritizing ideas for synthesis so that synthetic resources could be focused on the most prospective compounds/ideas.
- (f) To address the selectivity challenges of a molecule for its appropriate molecular target is a critical property to be optimized precisely in the line of small molecule drug discovery and development [23] (structural insights for dialing in selectivity or dialing out the off-target activity).

20.2 Need for Benchmarking

SBVS is generally utilized for identifying molecules with a high likelihood of binding a target to prioritize further evaluation of potential hit compounds using *in vitro* assays, which are generally restricted to a small subset of the chemical database containing numerous thousands of molecules [24]. In order to address such issue of “early recognition,” virtual screening (VS) algorithms must be capable of placing the actives to the top during the ranking of the molecules. Although the performances of these programs vary, their comparison could be beneficial to the researchers and would aid in choosing the best-performing and most suitable tools and protocols as per the needs of drug designing. Benchmarks have been used since the beginning of the docking and VS periods in order to measure the performance of various programs by assessing their capacity to portray the geometric orientation of the crystal structure of the known ligands [25, 26] or by examining how well they can distinguish between active and inactive compounds [27, 28]. Therefore, it plays a key role in evaluating the performance of the docking tool in the VS operation.

A benchmark should contain cutting-edge techniques that can be utilized as standards for other investigations. Finally, to provide details regarding the effectiveness of the tested approaches, the datasets utilized for benchmarking should comprise a variety of targets and tested compounds. Users can choose the best method for their task by using the results of such benchmarking research. The techniques to assess and the datasets on which the methods must be tested are the inputs to the benchmarking process. The process’s final product is an evaluation of the suggested approaches using the specified performance metrics. The three stages of the benchmarking process are data preparation, database ranking (screening), and evaluation (Fig. 20.2).

The benchmarking datasets are prepared at the initial step and are difficult, laborious, and time-consuming to prepare in order to fulfill the criteria for benchmarks and obtain a high-quality dataset [29]. The dataset should be able to explain the relevant characteristics of the property under investigation [30]. The dataset should be of sufficient size to allow statistical studies, cover the maximum number of known relevant events, and most importantly, be free of any unnecessary overlapping events. Therefore, a fair benchmarking dataset has become crucial for comprehensive quality evaluation [31].

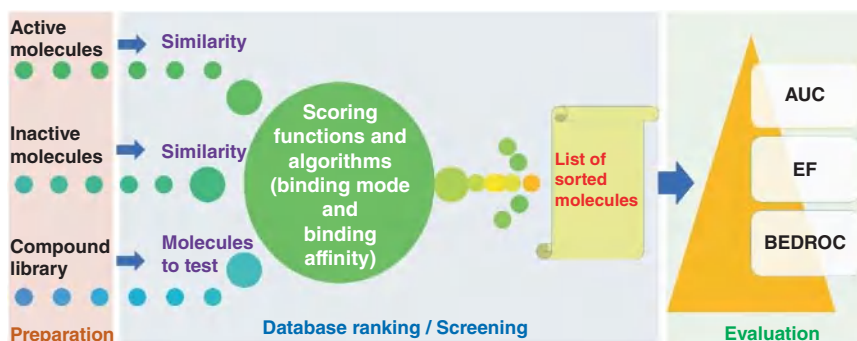


Figure 20.2 General workflow for assessing and benchmarking a structure-based virtual screening paradigm. It involves the preparation of the benchmarking dataset (that includes actives and inactives), which is assessed for the similarity of shape, structure, and interactions of docking poses, and finally a testing dataset (for which actual screening is carried out). For a good benchmarking process, actives and inactives should possess chemical similarity in order to minimize the biasness. Further after preparing datasets, a scoring method is used to predict binding poses. Lastly, evaluation metrics such as area under curve (AUC) and Enrichment Factor are utilized to analyze the tested methods and benchmarking datasets.

20.2.1 Benchmarking Datasets

Several benchmarking datasets have been developed and are available in the public domain. These datasets can be divided into datasets for pose prediction and datasets for VS. In general, these datasets contain information about the target protein, ligands for docking, experimental binding affinity, experimentally identified binding conformation (co-crystal structure), and the labeled actives or inactives [32].

The utility of benchmarking datasets has been improved over time with the development of newer improved benchmarking datasets; however, there are a few issues related to the choice of (a) the active compounds, like the significance of the manual curation of the data as well as the pharmacological profile of the active compounds; (b) the decoy selection, especially the significance of switching from presumptive to true decoys; and (c) the relative target structure must be pre-processed and corrected in order to make sure a high-quality assessment of the method is being carried out.

20.2.1.1 Benchmarking Sets for Pose Prediction

The success of a docking algorithm for a given dataset relies upon its ability to appropriately predict a ligand's known binding position. Thus, enabling it to be fine-tuned and parameterized in such a manner that the predictions can better represent reality. Accordingly, these above-mentioned models should be checked for their performance on aforesaid data. The demand for such benchmarking datasets that primarily intend to group the maximum number of high-quality data together, of which one such extensively used is PDBBind. There are further refined groups of complexes and a core obtained from them, that are being utilized as a standard set in benchmarking scoring functions (SFs).

There are benchmarking databases as well that are specifically meant for certain purposes, like protein–protein, membrane protein–protein complexes, and a PDBBind-derived blind set for measurement of machine learning scoring functions. The predicted orientation could be evaluated for accuracy with the help of root mean square deviation (RMSD) by correlating the predicted pose with the experimental one.

Cross Docking Benchmark Dataset There have been humongous efforts made during the last decade in order to improve the molecular docking approach to precisely anticipate the binding pose and the ranking affinities of the molecules. One of the major shortcomings is the availability of standard measures for assessing docking accuracy and the presence of universally accepted datasets in order to benchmark and compare numerous different docking algorithms throughout. A cross-docking benchmark server was created to overcome the above-stated issue. It consists of a dataset of approximately 4399 protein–ligand complexes of about 95 different target proteins, designed to be delivered as benchmarking set and gold standard for pose prediction and ranking for docking targets. The subset used for the benchmarking was designed from the target described in DUD-E (where DUD stands for Directory of Useful Decoys and E stands for enhanced), since it includes functionally diversified groups of target proteins like kinases, proteases, and other enzymes [33].

Astex It is a diverse dataset that is primarily anticipated for empowering the users with a highly valuable set to test algorithms and aid in the construction and establishment of newer and enhanced scoring functions. The preliminary stage for the preparation of the validation set includes sequential protein analysis registered in the Protein Data Bank (PDB) in order to group the analogous amino acid sequences that represent the aforesaid protein system, like the cdk2 kinase, HIV protease, and more. Thus, the final version of its diverse data set comprises 85 assorted, relevant protein–ligand complexes that have been pre-processed in a suitable format in order to carry out docking studies, which is readily available online (<http://www.ccdc.cam.ac.uk>) to the entire research community [34].

PDBbind PDBbind database, originally released in 2004, aims to serve as an exhaustive collection of experimentally quantified data for the binding affinity of all the deposited protein complexes in the PDB as a molecular recognition process among the biological macromolecules and organic small molecules that are of huge importance in various life processes [35]. It thus provides us with the data for linking the energetic properties of biomolecular complexes with their structural information, which is advantageous for in-silico analysis. The database is being updated annually in order to maintain balance with the growth of Protein Data Bank. It is publicly available on the PDBbind-CN Web server at <http://www.pdbbind-cn.org>.

Among various applications of PDBbind, it is also utilized for validating and developing docking/scoring methods. Like, AutoDock Vina was developed by Olson *et al.* using PDBbind as a refined set for calibrating the scoring function [36]. Which resulted in a nearly 100-fold increase in speed and better binding orientation prediction capabilities as compared to the earlier AutoDock 4 [37].

20.2.1.2 Benchmarking Sets for Virtual Screening

DUD and DUD-E The primary measure to quantify molecular docking depends upon the ligand enrichment with challenging decoys. DUD [32] was developed to improve the selection of the decoys in order to meet the benchmarking needs and immediately matured to be the widely accepted benchmark for assessing various virtual ligand screening (VLS) techniques. Decoys are selected in such a way that they display identical physicochemical attributes with regard to the active molecules and further have diverse structural characteristics in an effort to eliminate selection bias and choose substances that could fulfill their role as negative control and not actually bind to the target under study. Intense use of the DUD database had indicated some weaknesses in both the ligands and decoys that resulted in some serious liabilities, such as the absence of diverse ligands, inadequate property matching with respect to its net charge, as well as there is a noteworthy number of fake decoys [38].

The DUD-E database, an improved version of the DUD where 'E' stands for 'enhanced', was suggested in 2012 [38]. The DUD-E was designed to overcome the shortcomings in the original DUD directory with regard to the molecules (ligands and decoys) and thus provide a significant amount of data sets (102 targets separated into 8 different protein categories). The DUD-E decoys creation protocol differs somewhat from the DUD protocol and is much more compelling, which resulted in the local chemical space being narrowed or widened around six different properties of the given active moiety (molecular weight, hydrogen bond acceptors and donors, approximated water-octanol partition coefficient, rotatable bonds, and the net charge) and was used to extract 3000 to 9000 DUD-E decoys for each protonation state in the pH range of 6–8 for each active molecule from the ZINC database. Just the first 25% of the most contradictory ZINC compounds were preserved after a more rigorous topological dissimilarity screening utilizing the ECFP4 fingerprints and Tanimoto coefficient. Each and every decoy corresponds to one and only one ligand, in order to make it easier to remove the others if the study requires filtering a particular ligand. The DUD-E benchmarking set is available at <http://dude.docking.org> for the global research community.

Lit-PCBA Using the PubChem BioAssay database (PCBA), a new dataset has been created, LIT-PCBA, which is unbiasedly prepared considering machine learning as well as VS applications and contains pre-processed input files (ligands and targets) for direct application. The data set was generated using 149 dose-response PCBAs, which have clear definitions for active and inactive compounds. Prominently, a thorough analysis of the metadata made it possible to exclude assay artifacts, frequent hitters, and false positives, thereby able to keep the active and inactive compounds having the same molecular property range. In Sybyl-X 2.1.1, target set preparation was done, and if there were higher than 20 ligand-bound structures, the protein–ligand complexes were clustered based on the variety of their interaction patterns to get the graphs, which were then computed with IChem. GRIMscore metrics have been used for validating the similarity matrices. An agglomerative nesting clustering was performed using each matrix by utilizing ward clustering method, the Euclidean distance matrix, and the agnes function in R version 3.5.2,

and a maximum of 15 clusters were obtained. The highest resolution PDB entry for each cluster served as the protein–ligand PDB template to the associated target set. Preliminary VS attempts using cutting-edge techniques indicated that the data set was quite demanding, particularly as a result of the biases in the potency distribution of labeled active molecules were not present. Users can access the LIT-PCBA data set for free at <http://drugdesign.unistra.fr/LIT-PCBA>.

MUV MUV (Maximum Unbiased Validation) dataset is an alternative to DUD-E with advantages like its public access, and the dataset is designed using drug-like actives taken from PCBA and 3D structures of the targets (HIV RT-RNase, Cathepsin G, SF1, HSP90, PKA, FAK, and FXIa) from the PDB database. This dataset can be readily used for the validation of SBVS methods [39].

20.2.2 Evaluation Metrics

The evaluation phase involves computing performance metrics. The assessment of the benchmarking datasets for the active molecules and the precision for binding mode prediction are the two key criteria used to evaluate VS technologies retroactively. Receiver operator characteristics curve (ROC), area under the ROC curve (AUC), Enrichment Factor (EF), robust initial enhancement (RIE), or Boltzmann-enhanced discrimination of ROC (BEDROC) are all metrics that can be used in VS, with AUC and EF being the two that are most frequently used [40].

20.2.2.1 Enrichment Factor

EF is a parametric measure that expresses the enrichment over random selection for the top λ % of the database. It is calculated by dividing the actual number of actives by the predicted number of actives [29]. The formula for calculating this EF is

$$EF_{subset} = \frac{\left\{ \frac{Ligands_{selected}}{N_{subset}} \right\}}{\left\{ \frac{Ligands_{total}}{N_{total}} \right\}}$$

It thereby, helps to reflect the overall competence of the docking estimation in order to identify the true positives all over the background database as compared to the random selection. For example, for a particular target protein along with 100 annotated ligands ($ligands_{total}$) within the database consisting of 98,000 molecules (N_{total}), then only a known ligand ($ligands_{selected}$) probably will be found in a chosen subset of 980 compounds (N_{subset}) by the means of random selection, which corresponds to an EF of 1 [32]. Enrichment studies are generally utilized in SBVS process in order to pick a particular subset from the library, which is enriched with compounds analogous to the whole cluster and also has the ideally required potency toward an appropriate target of interest. The EF can be generally used to quantify the success rate of prediction of the percentage of active compounds from the screening dataset [30].

20.2.2.2 Docking Enrichment (DE)

DE is a perceptive parameter, which is quite easily apprehended [31] and is especially useful for VS, where the results can be phenomenally influenced by the choice of background datasets, like the decoys used for the screening. It denotes the percentage of true positives that are found (h) within all the actives (n) for a certain percentile of the top-ranked molecules ($x\%$) of the chemical dataset.

$$DE_{x\%} = \frac{h}{n}$$

The decoy molecules in relation to the ligands play a critical role in such EFs. Docking-based scoring functions can particularly depend on the molecular size. If the difference is significant with respect to its size distribution between the decoys and ligands then the docking enrichment (DE) can perform well, likewise for other physical properties [41]. The decoy dataset must feature the overall physical characteristics of the ligands sufficiently, so as to ensure that the enrichment is not just a simple separation based on trivial physical features. Furthermore, the decoys should possess chemically distinct characteristics with respect to ligands so that they will be non-binders to the target protein under investigation [42].

20.2.2.3 BEDROC Score

Boltzmann-enhanced discrimination of receiver operating characteristic (BEDROC) score [31] resulted as a derivative of the generalization of ROC that can address the ‘early recognition’ as well as RIE. It plays a considerable role in SBVS, where it is involved in the estimation of binding affinity of the ligand onto a receptor through a suitable scoring function after the successful docking of the ligand candidate onto a target protein [43]. The success rate of VS lies in the synchronous consolidation of a docking protocol with a scoring function used to rank the actives at the beginning of the screening in a larger set of compounds, and such a combination can be referred to as the ranking method [44, 45].

It is more complex and less intuitive compared to DE; it takes account of all the atoms and is not bound by the size of the chemical database for a given protein. But, this score might get altered due to the weight provided for the top-ranked molecules by parameter α .

$$BEDROC = \frac{\sum_{i=1}^n e^{-\alpha r_i/N}}{R_\alpha \left(\frac{1-e^{-\alpha}}{e^{\alpha/N}-1} \right)} \times \frac{R_\alpha \sinh(\alpha/2)}{\cosh(\alpha/2) - \cosh(\alpha/2 - \alpha R_\alpha)} + \frac{1}{1 - e^{\alpha(1-R_\alpha)}}$$

The number of actives is represented by “ n ”, the ratio of actives in the chemical database is denoted by “ R_α ” ($R_\alpha = n/N$), “ N ” is the total number of compounds, and “ r_i ” is indicated by the program as the rank of the i th active. For $\alpha = 80.5$, 80% of the BEDROC score is accounted for by the 2% top-ranked molecules. The α value may be altered so that 0.5% ($\alpha = 321.9$) or 8% ($\alpha = 20.0$) among the top-ranked molecules can account for about 80% of the score. The adaptation to the early problem recognition can result in modulation of the weight of the top-ranked molecules by utilizing the parameter α , which represents the degree of “early recognition” [46].

20.2.2.4 Receiver Operating Characteristic Curve

There are some ways by which the quality of generated ranking can be assessed, termed as metrics or figures of merit (FoM) out of which the recall or retrieval rate (RTR) and receiver operating characteristic curve (ROC) are common. The ROC curve method is a well-established metric implemented in VS that can be adopted objectively to quantify the capability of a particular test set to distinguish between two populations. It is important to use validation sets containing actives and decoys that will help in estimating the competence of a VS workflow in selecting active molecules as well as the ability to dispose of the inactive ones; the ROC curve is primarily well accustomed for such critical decision making [47].

It practically aids in evaluation of the success of a given docking or VS protocol, thereby assisting in critical decision-making and eliminating the risk of any errors in judgment that could have serious consequences like new hit prediction, compound library screening, docking predictions, and so on. Thus, it could be greatly beneficial in the field of drug design and discovery for critical assessment and decision-making.

20.2.2.5 Root Mean Square Deviation (RMSD)

RMSD is a binding mode prediction method to assess the overall aspect of the imitation of a known (i.e. crystallographic) binding orientation with the help of a computational method, like docking. RMSD values are calculated for a ligand when it displays different poses at a specific binding site cluster of a protein [48]. It is utilized for directly differentiating the near-native conformations with respect to the highly deviating ones having high RMSD values. RMSD-based methods can boost 120% improvement on docking outcomes when compared to their counterparts trained for binding affinity prediction [49]. Accordingly, in the case of re-docking, a low RMSD with respect to the original binding pose is good (preferably less than 1.5 Å, or even better when it is less than 1 Å). This signifies good reproduction of the docked and correct pose.

20.2.2.6 Real-Space *R* Values

Real-space *R* values (RSR) have been developed to be a technique to estimate how well a crystallographic model can fit in the experimental electron density for binding mode prediction [50]. RSR values can also be applied to evaluate the overall aspect of small components in the model, such as particular individual ligands, in contrast to conventional *R*-values, which provides the magnitude to which it fits the model's overall structure with regard to the crystallographic data. RSR values can compare the theoretical density (ρ_{calc}) computed precisely from the atomic model with an experimental density (ρ_{obs}) spacing on the grid is commonly set to one-third of the actual resolution, so both densities are calculated. In order to determine the RSR values, grid points that are near the scrutinized entity (e.g. a residue or ligand) are the only ones that are generally considered. The RSR value can be calculated as

$$RSR = \frac{\sum |\rho_{obs} - \rho_{calc}|}{\sum |\rho_{obs} + \rho_{calc}|}$$

Both the sums expand over the grid points near residue or ligand. Such as where, according to the RMSD criterion, a docking position could be wrongly labeled as a docking failure.

20.2.2.7 Root Mean Squared Error (RMSE)

RMSE can be explained as the mean of the square of all the errors that have been square-rooted. It is utilized as a standard statistical metric to quantify model performance in terms of absolute binding affinity for the docking study. RMSE is commonly applied and acknowledged as a superior all-purpose error metric in the field of numerical predictions.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=0}^n (f_i - o_i)^2}$$

Where $\sum$ is the summation of all values; f is the predicted value; o is the observed or actual value; $(f_i - o_i)^2$ are the differences between predicted and observed values and squared; and n is the total sample size. RMSE is generally considered as a good extent of accuracy, but as it is scale-dependent, it can only compare the forecasting errors of various models or model configurations for a particular given variable and not between different variables [51].

20.2.2.8 Spearman's Rank-Order Correlation

Spearman's rank-order correlation is a nonparametric form of Pearson product-moment correlation used for binding affinity ranking in docking methods. Spearman's correlation coefficient is utilized to measure the strength and direction of the relationship among the two ranking variables (ρ , also signified by r_s). Correlation coefficients have constant numerical values between -1.0 and 1.0 , and values closer to -1.0 or 1.0 suggest a substantial linear association.

It is useful where one needs to determine the relationship between the two sets of statistical data. It has benefits over the Pearson's correlation, which is more suitable when continuous data is available for a pair of variables and the relationship follows a linear line, whereas if both the criteria are not met by the dataset, then Spearman's rank-order correlation is a better choice [52].

20.2.2.9 Kendall Rank Correlation

Also commonly known as "Kendall's tau coefficient." Kendall rank correlation is generally utilized for testing the similarities among the ordering of the data when ranked by quantities in order to rank the binding affinity prediction. It uses a pairs of observations for determining the strength of association depending upon the pattern of concordance and discordance among the pairs [53, 54].

- **Concordant:** Identical manner of order (consistency). Observation pairs are considered concordant if $(x_2 - x_1)$ and $(y_2 - y_1)$ have identical signs.
- **Discordant:** Differently ordered (inconsistency). Observation Pairs are considered as discordant if $(x_2 - x_1)$ and $(y_2 - y_1)$ have contradictory signs.

Spearman's rho correlation values are generally larger than Kendall's Tau values. The calculations are typically based on the concordant and discordant pairs. It is especially helpful in scenarios where the data to be worked upon has already been ineffective in one or more assumptions during the test. *P* values are more precise and indifferent from errors when the size of the sample is smaller and is the best alternative to Spearman's correlation.

20.3 Community Benchmarking Exercises

In recent years, crowdsourcing competitions have become popular, and several industries/academic groups have launched open competitions challenging the scientific community to come up with computational tools/techniques, which can predict the protein–ligand binding affinity accurately and efficiently. The following section covers these community challenges.

20.3.1 Statistical Assessment of Proteins and Ligands (SAMPL) Challenge

Started in 2008, Statistical Assessment of Proteins and Ligands Challenge is a sequence of open competitions with participation from various scientific areas including academic, industrial, and research institutions. In these challenges, participants are given tasks to predict various parameters like binding affinities, binding poses, partition coefficient, and *pK*_a predictions, which can be later used in the drug discovery process. SAMPL3 challenge was for fragment-based screening and binding affinity prediction against trypsin. In continuation, SAMPL4 challenge was focused on small molecule VS, binding pose, and binding affinity prediction with regard to the core catalytic domain of HIV integrase. SAMPL5 challenge was held in 2015–2016, in which a total of 76 prediction sets were submitted by all over 18 participating groups. Various error metrics were utilized for predictions so as to analyze the statistics along with the assessment of individual submissions [55].

The latest addition to such computational chemistry-based blind trial challenges that focus on binding affinity study are SAMPL7 and SAMPL9 challenges. SAMPL7 focuses primarily on the fragment-based screening, binding orientation prediction, and binding affinity prediction against its pharmacologically admissible bromodomain target, which is PHIP2. There has been evidence that PHIP has a crucial role in cancer metastasis. The challenge involved X-ray crystallographic fragment-based screening experiment that led to the 3D structures of multiple hits. The binding pose as well as the binding energy for the ligands were to be approximated, and the top ligands with high binding affinity would be taken for experimental validation. The SAMPL7 initiative also encompassed another challenge, which focused more on *pK*_a, partition, and permeability prediction as well [56]. SAMPL9 focused on the identification of molecules that can bind to nanoluciferase. A library of compounds was supplied to the participants, and they had to predict which among the library of compounds inhibited nanoluciferase and which did not. In the second phase of the

challenge, the participants have to predict the IC_{50} values of the active compounds. This also includes ranking the molecules based on their affinity for the target [57].

20.3.2 Drug Design Data Resource (D3R) Challenges

The drug design data resource (D3R) promotes the exchange of high-quality protein–ligand datasets and workflows, as well as community-wide, blinded prediction challenges, with the primary objective of advancing the field of computer-aided drug designing and development. The D3R, a tool supported by the National Institutes of Health, intends to enhance the techniques by developing better and more robust methodologies in ligand docking as well as scoring. It primarily focuses on the forecasting or ranking of the ligand–protein binding efficiency and evaluates computer-aided drug discovery methodologies by gathering protein–ligand datasets that have not been reported previously. A general procedure or method followed in every D3R grand challenge (GC) is provided with one or more specific targets and is divided into two phases. The first phase involves pose prediction testing and compound ranking ability based on affinity, where participants were given the freedom to select a PDB structure, which helps to test the docking as well as scoring methods, along with the methodology used for the purpose of assigning various protein–ligand binding interactions, selection of the correct protein structure, and flexibility of the binding site. Followed by the second phase, which involves the ranking methodology testing for compounds based on the awareness of some ligand–protein poses, which means participants have to repeat the task with available ligand pose information. It also contains a sub-challenge for predicting alchemical calculations of relative binding free energy. According to Spearman's rho rank correlation and Kendall's tau coefficients, predicted affinity rankings were assessed. Both are ranked perfectly (1 = perfect ranking, -1 = perfectly reversed), on a scale of 1 to -1 and the centered RMSDs (kcal/mol) were used to evaluate the free energy calculations. Affinity predictions are evaluated in terms of the ranking statistics. Kendall's τ [58, 59], Spearman's ρ [60], and pose predictions were evaluated in terms of the symmetry-corrected RMSD between predicted and crystallographic poses. The centered RMSE used for assessing the free energy sets was recomputed in 10,000 rounds of resampling with replacement to generate error bars based on experimental uncertainty in all D3R challenges (Table 20.1).

20.3.3 Critical Assessment of Computational Hit-Finding Experiments

The critical assessment of computational hit-finding experiments (CACHE) is a worldwide community benchmarking initiative that compares and improves various small-molecule hit identification protocols and algorithms through various stages of predictions and experimental validations (Fig. 20.3), which was launched in December 2021 with the first challenge in its series. All computational chemists and AI-savvy researchers from biotech, pharma, and academia are given the chance to take part in the CACHE Challenges, which aim to predict small compounds (hits) that bind to protein targets linked with diseases. Every CACHE challenge

Table 20.1 List of D3R GC challenges sponsored by the National Institutes of Health that have been successfully hosted in the past.

Challenge no.	Target	No. of predictions	Dataset of compounds	Method/tools
D3R GC 2015 [61]	HSP90 MAP4K4	350 prediction	200 HSP90 30 MAP4K4	AutoDock Vina [62], rDOCK [63] or variant (Surflex [64] and Gold) and some combined tools RosettaLigand-Omega-ROCS, Glide-Prime-Desmond-Qsite, and Surflex-Grim.
D3R GC2 2016-2017 [58]	Farnesoid X receptor (FXR)	262	102 (IC ₅₀) 36 cocrystal structures 18 and 15 for FE	Gold [59], Glide [60], Vina, IChem-GRIM, HYDE, and Smina.
D3R GC3 2017-2018 [65]	CatS VEGFR2, JAK2, p38-alpha, TIE2 and ABL1	375	136 CatS 85 VEGFR2 89 JAK2 72 p38- α 18 TIE2 2 ABL1	ICM [66], Glide [67], POSIT [68], LeadFinder [69], and SMINA
D3R GC4 [70]	CatS and Beta secretase 1 (BACE 1)	407	154 BACE 1 20 pose prediction 34 FE 459 CatS 39-FE	AutoDock Vina [62], Glide [67], ICM [66], Corina [41], Gold [42], CACTVS [71], Rosetta [72], and EFindSite [73]

specifies a particular protein target with biological or pharmacological relevance. Participants predict hits computationally, and the CACHE team experimentally validates the hits. Each competition consists of a hit-finding and a hit-expansion round of experimental testing and prediction, following which the generated data, inclusive of the chemical structures, are to be made freely accessible to the public. Every contest lasts roughly 20 months and is simple to join; participants only fill out an online form to start the research process.

In the first CACHE Challenge, which was supported by the Michael J. Fox Foundation, the target was Leucine-rich repeat serine/threonine-protein kinase 2 (LRRK2) (PDB ID: 7LHT), a commonly mutated gene, genetically belonging to Parkinson's disease [46]. Participants were asked to find hits for the WD40 repeat (WDR) (PDBID 6DLO) domain of LRRK2. Kinase domain-targeting molecules from different industries like Genentech, Pfizer, Novartis, and Merck are in preclinical or early stage clinical trials. Recent studies have discovered that the closed form of

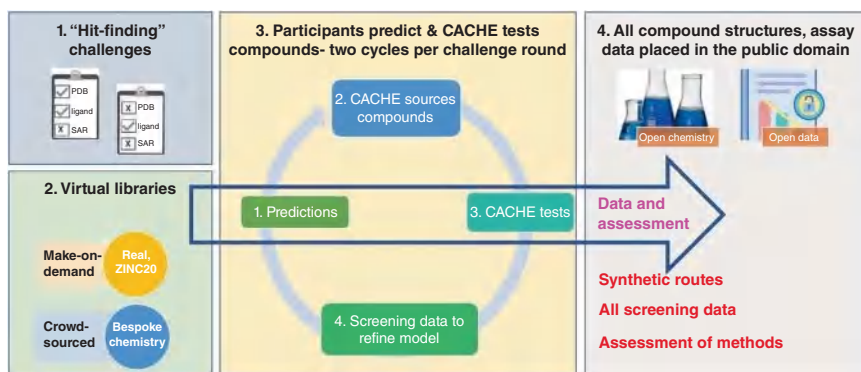


Figure 20.3 Workflow of CACHE Challenges, which consists of sequential steps 1. “Hit-finding” challenges, 2. Virtual libraries, 3. Participants utilize in-silico methodologies to anticipate the chemically most suitable hits and furthermore, CACHE experimentally tests the compounds, and 4. The generated data is made public to the researchers.

LRRK2 is poised by Kinase inhibitors, while the open form of LRRK2 is responsible for the formation of pathogenic LRRK2 filaments within the cells, which is thought to be inhibited by targeting the WDR domain of LRRK2 that is juxtaposed to the kinase domain. Thus, making it a novel approach to target this protein, and this will also be an allosteric mechanism of inhibiting LRRK2. All the data including prediction methods are made public at <https://cache-challenge.org/challenge-1/computational-methods>. The target for the second CACHE challenge is NSP13 helicase of SARS-CoV-2, which is an RNA binding site. In this challenge, participants have to find such type of ligands that can compete with RNA with the help of structure-based drug design, and they have provided the structure of helicase, i.e. PDB IDs: 5RLH, 5RLZ, 5RML, and 5RMM (Challenge #2 | CACHE [<http://cache-challenge.org>]). Applications have been opened for CACHE challenge 3, to identify ligands that can target the macro domain of SARS-CoV-2 NSP3 to bind at the ADPr site where participants can use any of the approaches from ligand-based or structure-based. Challenge has been specified to find such types of ligands that lack carboxylic acid and be able to compete with the substrate, i.e. ADP-ribose.

20.3.4 Continuous Evaluation of Ligand Protein Predictions (CELPP)

Since docking calculations are helping in the advancement of drug discovery and development processes by aiding in the prediction of the ligand binding pose against the targeted protein. CELPP a new blinded prediction contest that gets around addressing issues like which docking protocol works the best, protein and ligand pre-processing criteria that are to be taken care of, and the components that call for urgent need of improvement in the entire process of molecular docking [74]. To provide a weekly pose prediction task similar to the Continuous Automated Model Evaluation for the protein structure prediction challenge, which aided as its inspiration, CELPP employs a weekly PDB release with an inventory of structures scheduled for forthcoming publication, and its workflow is presented in Fig. 20.4.

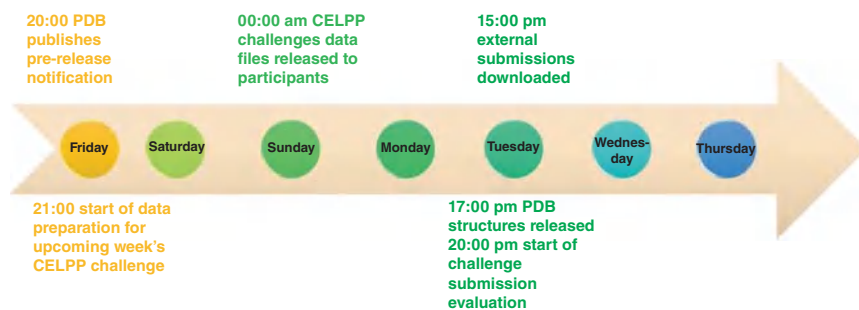


Figure 20.4 Workflow showing the timeline for CELPP weekly challenge.

In-house CELPP scripts detect new PDB entries having co-crystallized structures of proteins with small molecules appropriate for computerized docking calculations each week by downloading the series of new entries from PDB to be published after 5 days (<https://github.com/drugdata/D3R>).

Target complexes consist of the amino acid sequences of the protein, the identified ligand, and the mother liquor of the known pH of the crystal structure. The PDB is searched for the crystallographic structures with respect to each protein target using additional scripts (STAR Methods, which stands for Selection of Target Complexes and Receptor Structures), which can subsequently select nearly about five structures that are suitable for carrying out the docking calculation studies. Furthermore, the selected structures are added to the weekly CELPP data package, consisting of ligand information, pH values during crystallization, as well as other details of STAR Methods. The deadline for the CELPP is just before the publication of new PDB entries comprising the crystallographic orientation of binding poses, and participants can download the provided package data, execute the workflows exclusively designed by the participants in order to anticipate the ligand binding geometrical pose and must submit the predicted version of the docking pose before the due date on a web directory that is secure and password-protected. After the deadline, D3R scripts analyze the predicted data that were submitted, transmit the decision of the findings to every participant, and also update the results in the ongoing statistics that are accessible online. During the period of 2017–2018, a total of 1989 targets were selected and provided to the participants for ligand pose prediction.

20.4 Lessons Learned from the Benchmarking Exercises

A number of issues have been raised and addressed in the different benchmarking exercises [48], and some key lessons learned from them are summarized here.

20.4.1 Quality of Crystal Structures

The focus should be on the true electron density associated with the binding site of the crystallographic structure while selecting one. The real-space correlation

coefficient (RSCC) for the ligand must be less than 0.9, together with a well-resolved density, which is one of the most crucial characteristics of a good structure [49]. In order to prevent any unreal modification in the ligand coordinates, the ligand structures have to be free of any kind of interaction with the crystal additives, like the water molecules or crystal symmetrical packing [48].

20.4.2 Need of Sufficient Metrics to Assess Docking and Scoring

After the selection of a good crystal structure, the next step is the assessment of the docking method. The most common metric used for this analysis is the calculation of the RMSD value of the ligand, which compares the atom positions from the docked pose to the crystallized pose and should be less than 2.0 Å (Fig. 20.5A, B). But this is not the best method for assessment in case of cross-docking ligands or when considering protein flexibility. Structural alignment can be used either by superimposing the backbone of a protein or its binding site residues (that includes careful selection of residues). For instance, Damm et. al., used the Gaussian-weighted method (wRMSD with the help of the MOE software package) for the superimposition of protein structures. Another alternative to the RMSD metric can be the evaluation of protein–ligand contacts, which are associated with the choice of residues and the cut-offs used for the measurement of contacts. For example, Damm et. al., used a standard value cut-off of ≤ 3.5 Å for hydrogen-bond and electrostatic interactions (first-row atoms N, O, and F), while for larger atoms (S, Br, Cl, etc.), the cut-off was a little higher at ≤ 3.8 Å. For ligands coordinated to metal ions, they suggested ≤ 2.8 Å as a contact cut-off, nonetheless, these values probably depend on the system. The interactions for docked poses can be compared to those of crystal or NMR structures (Fig. 20.5C, D). Further, for comparison of experimental affinities with scoring or ranking of ligands, four values were used, namely R^2 , Pearson R , Kendall τ , and Spearman ρ . For assessment of the ability of a particular method to differentiate actives from inactives ROC were used and validated through AUCs. For this purpose, enrichment rates are also quite useful, at which all the docking and scoring methods are good but difficulties arise while ranking many related actives.

20.4.3 Usefulness of Statistics, Error Bars, and Confidence Intervals (CI)

The standard deviations or confidence intervals could be simply approximated with the help of bootstrapping process, which involves random selection of data sets with replacement. Sample size can be varied but random selection is a must while choosing a subset of data and recalculating the metric, for instance, AUC. The procedure to estimate error for AUC, if there are N ligands assessed by the given ranking algorithm, is to choose 90% of those N scores randomly and repeat the process 10,000 times, providing a distribution of about 10,000 AUC values. For any given scoring or ranking mechanism, its distribution provides the mean, median, standard deviation, and 95% CI of the AUC. The 95% confidence interval gives the order of values from 2.5% to 97.5% to the distribution, leaving the rest of the extreme 5% of the AUC values split among the two ends of the distribution. When such an approach is used

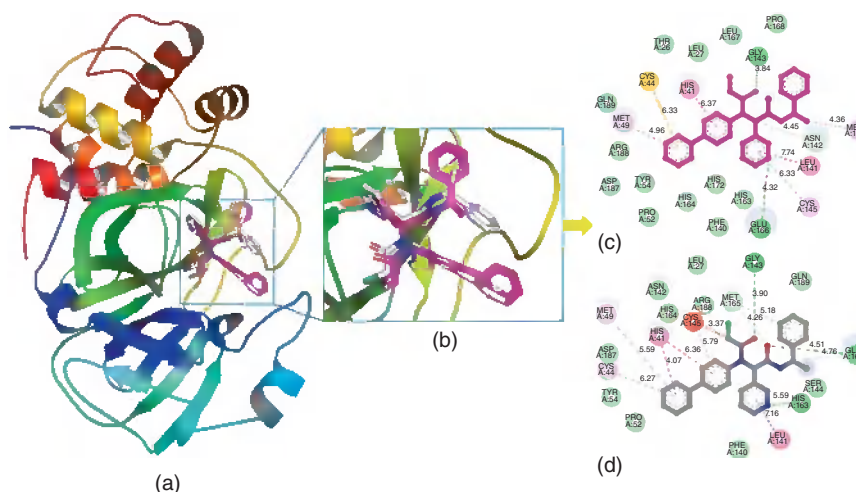


Figure 20.5 Example of protein–ligand RMSD metric and comparison of protein–ligand contacts with measured distances (Å) pre and post docking. (A) Showing the position of ligand at the binding site of SARS-CoV-2 Mpro (PDB ID: 7RN1). The 3D figure was prepared with the help of PyMOL 2.5 [75]. (B) Enlarged view of superimposed structure of co-crystallized ligand before (pink) and after (gray) docking holding an RMSD value of 1.8 Å. (C) Protein–ligand interaction of co-crystallized structure before docking, and (D) protein–ligand interaction of co-crystallized structure after docking. The 2D docking image was made with the help of BIOVIA Discovery Studio Visualizer [76].

on a new data set, the 95% CI is an estimation of the range of AUC values that will probably occur 95% of the time.

20.4.4 Requirement of Good Data Set

A dataset of ligands should have a large range of physical properties like the number of hydrogen bond groups, rotatable bonds, molecular weight, etc., which help in the examination of several factors that can affect scoring. Moreover, inactive compounds that had their chemical similarity tested and verified experimentally should also be included instead of assumed inactives. The affinities for active compounds should evenly span ≥ 4 orders of magnitude to minimize the consequences as a result of experimental error. Kramer et al. [77] ingeniously deduced that the range of data and experimental accuracy determine the maximum Pearson R for a data set. When a model performs better than $R_{\max}$ at fitting the data, it was probably overfitted with too many parameters; however, it could also have been luck. A high 95% CI would be present in the Pearson R from that lucky/overfit model to denote a low level of statistical significance. Additionally, Kramer et al., examined ChEMBL20 with a focus on compounds whose affinities were determined by various, independent labs. They discovered that over a subgroup of ChEMBL, the standard deviation was $\sigma_{\text{expt}} = 0.54 \text{ p}K_i$. According to anecdotal evidence, researchers deem threefold variations in individually measured K_i values (or K_d or IC_{50}) to be in excellent accord, which amounts to a fairly similar value of $0.48 \text{ p}K_i$. Due to this, the variation

in the error between laboratories is 0.5 pK_i, which is greater than the standard error bars claimed by the literature (which are measured within one lab).

20.5 Summary

Drug discovery is an extensive and challenging pathway that needs a humungous amount of funding as well as concerted efforts from interdisciplinary experts. Through advancements in computational methods and techniques, this process has become a little bit more efficient, less expensive, less time-consuming and has a higher success rate. Molecular docking as a computational tool helps in accelerating the hit identification through VS of large chemical libraries. However, before using molecular docking, its proper validation using different tools is important in order to use the methods efficiently. These validation methods, like the use of benchmarking sets, RMSD, and ROC calculations, can be carried out before proceeding to find hit through VS. Lately, several community challenges focused on the identification of hit compounds or prediction of binding affinity/binding free energy have succeeded in achieving their objectives. These benchmarking and validation tools should be implemented rigorously in such challenges to increase the reliability and reproducibility of the results from these screens.

Ultimately, this can aid in the improvement of methods and techniques used in drug discovery since people with different expertise, like computational biologists and medicinal chemists do explore and research together so as to obtain reliable data and increase the likelihood that a molecule will develop into a medicine. These activities will aid the CADD community in benchmarking and prospectively validating the current docking and scoring programs. Understanding and guiding future technical improvement based on the lessons acquired through such exercises will all be advantageous to the community.

References

- 1 Yang, C., Chen, E.A., and Zhang, Y. (2022). Protein–ligand docking in the machine-learning era. *Molecules* 27 (14): 4568.
- 2 Mohan, A., Banerjee, S., and Sekar, K. (2021). Role of advanced computing in the drug discovery process. In: *Innovations and Implementations of Computer Aided Drug Discovery Strategies in Rational Drug Design*, 59–90. Springer.
- 3 Sliwoski, G., Kothiwale, S., Meiler, J., and Lowe, E.W. (2014). Computational methods in drug discovery. *Pharmacol. Rev.* 66 (1): 334–395.
- 4 Pinzi, L. and Rastelli, G. (2019). Molecular docking and scoring: shifting paradigms in drug discovery. *Int. J. Mol. Sci.* 20 (18): 4331.
- 5 Yadava, U. (2018). Search algorithms and scoring methods in protein–ligand docking. *Endocrinol. Int. J.* 6 (6): 359–367.
- 6 Hahn, D., Bayly, C., Boby, M.L. et al. (2022). Best practices for constructing, preparing, and evaluating protein–ligand binding affinity benchmarks [article v1.0]. *Living J. Comput. Mol. Sci.* 4 (1): 1497.

- 7 Neves, M.A., Totrov, M., and Abagyan, R. (2012). Docking and scoring with ICM: the benchmarking results and strategies for improvement. *J. Comput. Mol. Design* 26 (6): 675–686.
- 8 Moitessier, N., Englebienne, P., Lee, D. et al. (2008). Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. *British journal of pharmacology*. 153 (S1): S7–S26.
- 9 Milletti, F. and Vulpetti, A. (2010). Tautomer preference in PDB complexes and its impact on structure-based drug discovery. *Journal of chemical information and modeling*. 50 (6): 1062–1074.
- 10 Roberts, B.C. and Mancera, R.L. (2008). Ligand– protein docking with water molecules. *Journal of chemical information and modeling*. 48 (2): 397–408.
- 11 Kirton, S.B., Murray, C.W., Verdonk, M.L., and Taylor, R.D. (2005). Prediction of binding modes for ligands in the cytochromes P450 and other heme-containing proteins. *Proteins: Structure, Function, and Bioinformatics*. 58 (4): 836–844.
- 12 Ten Brink, T. and Exner, T.E. (2010). pKa based protonation states and microspecies for protein–ligand docking. *Journal of computer-aided molecular design*. 24 (11): 935–942.
- 13 Meng, X.-Y., Zhang, H.-X., Mezei, M., and Cui, M. (2011). Molecular docking and scoring: a powerful approach for structure-based drug discovery. *Current computer-aided drug design*. 7 (2): 146–157.
- 14 Fan, J., Fu, A., and Zhang, L. (2019). Progress in molecular docking. *Quantitative Biology*. 7 (2): 83–89.
- 15 Koshland, D.E. Jr., (1995). The key-lock theory and the induced fit theory. *Angewandte Chemie International Edition in English*. 33 (23-24): 2375–2378.
- 16 Brooijmans, N. and Kuntz, I.D. (2003). Molecular recognition and docking algorithms. *Annual review of biophysics and biomolecular structure*. 32 (1): 335–373.
- 17 Kitchen, D.B., Decornez, H., Furr, J.R., and Bajorath, J. (2004). Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature reviews Drug discovery*. 3 (11): 935–949.
- 18 Verkhivker, G.M., Bouzida, D., Gehlhaar, D.K. et al. (2000). Deciphering common failures in molecular docking of ligand-protein complexes. *Journal of computer-aided molecular design*. 14 (8): 731–751.
- 19 Ramírez, D. and Caballero, J. (2018). Is it reliable to take the molecular docking top scoring position as the best solution without considering available structural data? *Molecules*. 23 (5): 1038.
- 20 Maia, E.H.B., Assis, L.C., De Oliveira, T.A. et al. (2020). Structure-based virtual screening: from classical to artificial intelligence. *Frontiers in chemistry*. 8: 343.
- 21 Jacquemard C, Drwal MN, Desaphy J, Kellenberger E. Binding mode information improves fragment docking. *Journal of cheminformatics*. 2019; 11(1): 1-15.
- 22 Danao, K., Nandurkar, D., Rokde, V. et al. Molecular docking and scoring: metamorphosis in drug discovery. *Molecular Docking-Recent Advances*. .
- 23 Johnson, D.K. and Karanicolas, J. (2015). Selectivity by small-molecule inhibitors of protein interactions can be driven by protein surface fluctuations. *PLoS computational biology*. 11 (2): e1004081.

- 24 Lagarde, N., Zagury, J.-F., and Montes, M. (2015). Benchmarking data sets for the evaluation of virtual ligand screening methods: review and perspectives. *Journal of chemical information and modeling*. 55 (7): 1297–1307.
- 25 Plewczynski, D., Łaźniewski, M., Augustyniak, R., and Ginalski, K. (2011). Can we trust docking results? Evaluation of seven commonly used programs on PDBbind database. *Journal of computational chemistry*. 32 (4): 742–755.
- 26 Kellenberger, E., Rodrigo, J., Muller, P., and Rognan, D. (2004). Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins: Structure, Function, and Bioinformatics*. 57 (2): 225–242.
- 27 Charifson, P.S., Corkery, J.J., Murcko, M.A., and Walters, W.P. (1999). Consensus scoring: a method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *Journal of medicinal chemistry*. 42 (25): 5100–5109.
- 28 Warren, G.L., Andrews, C.W., Capelli, A.-M. et al. (2006). A critical assessment of docking programs and scoring functions. *Journal of medicinal chemistry*. 49 (20): 5912–5931.
- 29 Škoda P, Hoksza D, editors. Benchmarking platform for ligand-based virtual screening. 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2016: IEEE.
- 30 Chen, H., Lyne, P.D., Giordanetto, F. et al. (2006). On evaluating molecular-docking methods for pose prediction and enrichment factors. *Journal of chemical information and modeling*. 46 (1): 401–415.
- 31 Chaput, L., Martinez-Sanz, J., Saettel, N., and Mouawad, L. (2016). Benchmark of four popular virtual screening programs: construction of the active/decoy dataset remains a major determinant of measured performance. *Journal of cheminformatics*. 8 (1): 1–17.
- 32 Huang, N., Shoichet, B.K., and Irwin, J.J. (2006). Benchmarking sets for molecular docking. *Journal of medicinal chemistry*. 49 (23): 6789–6801.
- 33 Wierbowski, S.D., Wingert, B.M., Zheng, J., and Camacho, C.J. (2020). Cross-docking benchmark for automated pose and ranking prediction of ligand binding. *Protein Science*. 29 (1): 298–305.
- 34 Repasky, M.P., Murphy, R.B., Banks, J.L. et al. (2012). Docking performance of the glide program as evaluated on the Astex and DUD datasets: a complete set of glide SP results and selected results for a new scoring function integrating WaterMap and glide. *Journal of Computer-Aided Molecular Design*. 26: 787–799.
- 35 Wang, R., Fang, X., Lu, Y., and Wang, S. (2004). The PDBbind database: collection of binding affinities for protein? ligand complexes with known three-dimensional structures. *Journal of Medicinal Chemistry*. 47 (12): 72977–72980.
- 36 Trott, O. and Olson, A.J. (2010). AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*. 31 (2): 455–461.
- 37 Eberhardt, J., Santos-Martins, D., Tillack, A.F., and Forli, S. (2021). AutoDock Vina 1.2. 0: new docking methods, expanded force field, and python bindings. *Journal of Chemical Information and Modeling*. 61 (8): 3891–3898.

- 38 Mysinger, M.M., Carchia, M., Irwin, J.J., and Shoichet, B.K. (2012). Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *Journal of medicinal chemistry*. 55 (14): 6582–6594.
- 39 Rohrer, S.G. and Baumann, K. (2009). Maximum unbiased validation (MUV) data sets for virtual screening based on PubChem bioactivity data. *Journal of chemical information and modeling*. 49 (2): 169–184.
- 40 Truchon, J.-F. and Bayly, C.I. (2007). Evaluating virtual screening methods: good and bad metrics for the “early recognition” problem. *Journal of chemical information and modeling*. 47 (2): 488–508.
- 41 Tetko, I.V., Gasteiger, J., Todeschini, R. et al. (2005). Virtual computational chemistry laboratory—design and description. *Journal of computer-aided molecular design*. 19 (6): 453–463.
- 42 Verdonk, M.L., Cole, J.C., Hartshorn, M.J. et al. (2003). Improved protein–ligand docking using GOLD. *Proteins: Structure, Function, and Bioinformatics*. 52 (4): 609–623.
- 43 Zhao, W., Hevener, K.E., White, S.W. et al. (2009). A statistical framework to evaluate virtual screening. *BMC bioinformatics*. 10 (1): 1–13.
- 44 Jain, A.N. (2004). Virtual screening in lead discovery and optimization. *Current opinion in drug discovery & development*. 7 (4): 396–403.
- 45 Doman, T.N., McGovern, S.L., Witherbee, B.J. et al. (2002). Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. *Journal of medicinal chemistry*. 45 (11): 2213–2221.
- 46 Burai-Patrasu M, Nivedha AK, Rostaing O, Chukka P, Moitessier N, Pottel J. *The First CACHE Challenge—Identifying Binders of the WD-Repeat Domain of Leucine-Rich Repeat Kinase 2*. 2022.
- 47 Pathania, S., Randhawa, V., and Bagler, G. (2013). Prospecting for novel plant-derived molecules of Rauvolfia serpentina as inhibitors of aldose reductase, a potent drug target for diabetes and its complications. *PloS one*. 8 (4): e61327.
- 48 Carlson, H.A. (2016). *Lessons Learned over Four Benchmark Exercises from the Community Structure–Activity Resource*, 951–954. ACS Publications.
- 49 Deller, M.C. and Rupp, B. (2015). Models of protein–ligand crystal structures: trust, but verify. *Journal of computer-aided molecular design*. 29 (9): 817–836.
- 50 Yusuf, D., Davis, A.M., Kleywegt, G.J., and Schmitt, S. (2008). An alternative method for the evaluation of docking performance: RSR vs RMSD. *Journal of chemical information and modeling*. 48 (7): 1411–1422.
- 51 van Westen, G.J., Swier, R.F., Cortes-Ciriano, I. et al. (2013). Benchmarking of protein descriptor sets in proteochemometric modeling (part 2): modeling performance of 13 amino acid descriptor sets. *Journal of cheminformatics*. 5 (1): 1–20.
- 52 Ain, Q.U., Aleksandrova, A., Roessler, F.D., and Ballester, P.J. (2015). Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. *Wiley Interdisciplinary Reviews: Computational Molecular Science*. 5 (6): 405–424.

- 53 Kendall, M.G. (1938). A new measure of rank correlation. *Biometrika*. 30 (1/2): 81–93.
- 54 Kendall, M.G. (1945). The treatment of ties in ranking problems. *Biometrika*. 33 (3): 239–251.
- 55 Yin, J., Henriksen, N.M., Slochower, D.R. et al. (2017). Overview of the SAMPL5 host–guest challenge: are we doing better? *Journal of computer-aided molecular design*. 31 (1): 1–19.
- 56 Amezcua M, El Khoury L, Mobley DL. SAMPL7 Host–Guest Challenge Overview: assessing the reliability of polarizable and non-polarizable methods for binding free energy calculations. *Journal of Computer-Aided Molecular Design*. 2021; 35(1): 1-35.
- 57 Deng, C.-L., Cheng, M., Zavalij, P.Y., and Isaacs, L. (2022). Thermodynamics of pillararene-guest complexation: blinded dataset for the SAMPL9 challenge. *New Journal of Chemistry*. 46 (3): 995–1002.
- 58 Gaieb, Z., Liu, S., Gathiaka, S. et al. (2018). D3R grand challenge 2: blind prediction of protein–ligand poses, affinity rankings, and relative binding free energies. *Journal of computer-aided molecular design*. 32 (1): 1–20.
- 59 Verdonk, M.L., Chessari, G., Cole, J.C. et al. (2005). Modeling water molecules in protein–ligand docking using GOLD. *Journal of medicinal chemistry*. 48 (20): 6504–6515.
- 60 Halgren, T.A., Murphy, R.B., Friesner, R.A. et al. (2004). Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *Journal of medicinal chemistry*. 47 (7): 1750–1759.
- 61 Gathiaka, S., Liu, S., Chiu, M. et al. (2016). D3R grand challenge 2015: evaluation of protein–ligand pose and affinity predictions. *Journal of computer-aided molecular design*. 30 (9): 651–668.
- 62 Huey, R., Morris, G.M., and Forli, S. (2012). Using AutoDock 4 and AutoDock vina with AutoDockTools: a tutorial. *The Scripps Research Institute Molecular Graphics Laboratory*. 10550: 92037.
- 63 Ruiz-Carmona, S., Alvarez-Garcia, D., Foloppe, N. et al. (2014). rDock: a fast, versatile and open source program for docking ligands to proteins and nucleic acids. *PLoS computational biology*. 10 (4): e1003571.
- 64 Coelho M, Ishii H, Maes P. Surflex: a programmable surface for the design of tangible interfaces. CHI’08 extended abstracts on Human factors in computing systems2008. p. 3429–34.
- 65 Gaieb, Z., Parks, C.D., Chiu, M. et al. (2019). D3R grand challenge 3: blind prediction of protein–ligand poses and affinity rankings. *Journal of computer-aided molecular design*. 33 (1): 1–18.
- 66 Abagyan, R., Totrov, M., and Kuznetsov, D. (1994). ICM—A new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. *Journal of computational chemistry*. 15 (5): 488–506.
- 67 Friesner, R.A., Banks, J.L., Murphy, R.B. et al. (2004). Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *Journal of medicinal chemistry*. 47 (7): 1739–1749.

- 68 Kelley, B.P., Brown, S.P., Warren, G.L., and Muchmore, S.W. (2015). POSIT: flexible shape-guided docking for pose prediction. *Journal of Chemical Information and Modeling*. 55 (8): 1771–1780.
- 69 Stroganov, O.V., Novikov, F.N., Stroylov, V.S. et al. (2008). Lead finder: an approach to improve accuracy of protein-ligand docking, binding energy estimation, and virtual screening. *J Chem Inf Model*. 48 (12): 2371–2385.
- 70 Parks, C.D., Gaieb, Z., Chiu, M. et al. (2020). D3R grand challenge 4: blind prediction of protein–ligand poses, affinity rankings, and relative binding free energies. *Journal of computer-aided molecular design*. 34 (2): 99–119.
- 71 Ihlenfeldt, W.D., Takahashi, Y., Abe, H., and Sasaki, S.-i. (1994). Computation and management of chemical properties in CACTVS: an extensible networked approach toward modularity and compatibility. *Journal of chemical information and computer sciences*. 34 (1): 109–116.
- 72 Chaudhury, S. and Gray, J.J. (2008). Conformer selection and induced fit in flexible backbone protein–protein docking using computational and NMR ensembles. *Journal of molecular biology*. 381 (4): 1068–1087.
- 73 Feinstein, W.P. and Brylinski, M. (2014). eFindSite: enhanced fingerprint-based virtual screening against predicted ligand binding sites in protein models. *Molecular informatics*. 33 (2): 135–150.
- 74 Wagner JR, Churas CP, Liu S, Swift RV, Chiu M, Shao C, et al. Continuous evaluation of ligand protein predictions: a weekly community challenge for drug docking. *Structure*. 2019; 27(8): 1326–35. e4.
- 75 Yuan, S., Chan, H.S., and Hu, Z. (2017). Using PyMOL as a platform for computational drug design. *Wiley Interdisciplinary Reviews: Computational Molecular Science*. 7 (2): e1298.
- 76 Jejuri, B.L. and Rohane, S.H. (2021). Drug designing in discovery studio. *Asian J Res Chem*. 14 (2): 135–138.
- 77 Kramer, C., Kallioikoski, T., Geddeck, P., and Vulpetti, A. (2012). The experimental uncertainty of heterogeneous public K_i data. *Journal of medicinal chemistry*. 55 (11): 5165–5173.

Part VI

In Silico ADMET Modeling

21

Advances in the Application of In Silico ADMET Models – An Industry Perspective

Wenyi Wang¹, Fjodor Melnikov², Joe Napoli¹, and Prashant Desai¹

¹Drug Metabolism and Pharmacokinetics, Genentech Inc., South San Francisco, CA 94080, United States

²Safety Assessment, Genentech Inc., South San Francisco, CA 94080, United States

21.1 Introduction

Small-molecule drug discovery involves extensive evaluation of molecular physico-chemical properties, pharmacokinetic (PK) properties, potencies, and toxicities. These properties are investigated by way of biochemical assays, cell-based assays, tissue-based assays, and preclinical animal studies. Scientists go through many rounds of selection and filtering to arrive at a candidate structure, from a potentially infinite chemical space. As such, compound selection matters *at all stages of drug discovery*. Early in the research, scientists must select a finite number of compounds for synthesis from the nearly infinite number of virtual possibilities. Subsequently, a select set of compounds or chemical series is evaluated in a variety of in vitro and in vivo screening assays before identifying candidates for clinical trials. At each stage of the process, scientists leverage existing knowledge and expert judgment to advance compounds that are most likely to be efficacious and safe when administered to humans. It is an expensive and challenging ordeal. Computational methods have been demonstrated to significantly improve the efficiency of drug discovery, reduce costs, and provide safer and more efficacious drugs to patients faster. This chapter focuses on how in silico tools are applied to enable, assist, and expedite the drug discovery process in the areas of absorption, distribution, metabolism, excretion, and toxicity (ADMET).

In silico ADMET tools involve diverse computational techniques and strategies that are incorporated in early drug discovery. In early research, generative models can be used to generate ideas for structural modifications or to design novel compounds or chemical series [1–3]. These are typically machine learning models that generate de novo molecular designs by leveraging information from large chemical databases. Another in silico tool is based on matched molecular pairs/series (MMP/MMS) [4, 5], which are derived from systematic analysis of molecular transformations in existing databases and associating those with corresponding changes in the experimental outcome for a given transformation across

multiple examples. MMP/MMS and many emerging artificial intelligence (AI) generative models [1, 6] can take ADMET properties and synthetic feasibility into account. One of the most common in silico models for ADMET are quantitative structure–activity relationship (QSAR) models, which relate compounds' chemical structures to their measured in vitro or in vivo ADMET properties using various machine learning algorithms [7–12]. When scientists or generative models propose new compounds for synthesis, such QSAR models can be used to prioritize the synthesis of compounds with more favorable predicted ADMET profiles. During the lead generation and optimization process, multiple in vitro assays are utilized to further characterize their ADMET profile. Several ADMET in vitro assays are expensive and have longer turn-around times. QSAR models for in vitro ADMET endpoints are often used at this stage to prioritize molecules for in vitro screening to expedite chemical selection and optimize in vitro resource utilization. Subsequently, as projects prepare to advance promising compounds to in vivo studies, understanding in vitro–in vivo correlation (IVIVc) becomes critical [13, 14]. At this stage, mechanistic PK models may be used to extrapolate the in silico or in vitro data to in vivo PK to evaluate in silico–in vitro–in vivo correlation (ISIVIVc) [15, 16]. These models utilize in vivo mechanistic considerations for a given species along with physico–chemical properties and in vitro ADME measurements of a compound to simulate its in vivo PK. Recently, QSAR models for in vivo PK parameters like oral bioavailability, area under the curve (AUC), and concentration–time profile have also been described, which can be used to prioritize compounds for in vivo PK studies [17–21]. Furthermore, investigating the metabolites of compounds is crucial, as these may have ADME properties that differ from those of the parent compound, potentially altering its safety and/or efficacy in vivo. To that end, various in silico approaches have been developed to predict the site of drug metabolism [22].

Following in silico predictions, measured data can be confirmatory or directive in identifying and filling mechanistic gaps when predicting in vivo behavior, either directly or indirectly through the mechanistic IVIVc models. In this chapter, we provide an overview of the various in silico models routinely used during the drug discovery process to design and prioritize compounds with optimal ADMET and PK properties. Our primary focus is on QSAR models for ADMET prediction since these are the most common models used by the industry.

21.2 QSAR Models

The process of discovering a new therapy/drug starting from the biological target identification to the ultimate approval by regulatory authorities for use in patients, is mainly divided into two phases: discovery and development. The goal of the discovery phase, often referred to as the pre-clinical phase, is to identify a promising drug candidate for initiating clinical studies in humans. During the discovery phase, scientists filter and select the most promising compounds from a large number of possibilities and advance them through the drug discovery pipeline with the goal of arriving at an efficacious and safe clinical candidate. During the development

(clinical) stage, a series of clinical studies are conducted to evaluate the efficacy and safety of the candidate. Thus, tools that help to increase the probability of finding compounds with desirable ADMET properties during each learning cycle are key to successful drug discovery. QSAR is one of the most commonly used computational techniques to predict various properties of small molecules [23]. These models have been used routinely across the industry to prioritize compound design, synthesis, and testing [12]. Over the past 15+ years, numerous publications from the industry have reported successful examples of building robust QSAR models for ADMET endpoints, as well as their prospective application in the real-world setting. For example, in a series of publications from Eli Lilly, the authors described a detailed process for data generation and curation, model evaluation, and real-world prospective application for a variety of ADME endpoints like efflux by P-glycoprotein (P-gp) [24], unbound brain-to-plasma partition coefficient ($K_{p,uu}$) [25], uptake by the organic anion-transporting polypeptide 1B1 transporter (OATP) [26], and Cytochrome P450-mediated victim drug–drug interaction [27]. A recent article from Bayer AG [28] presented their platform for building and delivering ADME QSAR models and highlighted the recent impact of deep neural networks (DNNs) using selected application examples. Sheridan et al. from Merck [29] emphasized the importance of regular model updates and shared prediction accuracy of production ADMET models as a function of their versions. Another publication from Pfizer [30] described an interpretable, probability-based confidence metric for continuous QSAR models such as the human liver microsomal (HLM) clearance predictor. Similarly, QSAR models for ADME endpoints have been around for more than a decade at Genentech and are routinely used in design prioritization before synthesis [11, 12, 16, 31–33]. It is apparent from the series of publications from the industry (including the examples listed above) that the development of high-quality and high-impact QSAR models requires careful processing of chemical and biological data, computational algorithm selection, and appropriate considerations for their applicability domain; especially when using the models in a prospective setting. The following sections aim to address these aspects of model building in more detail.

21.2.1 Conventional QSAR Model Development

A typical QSAR model requires three key components: (i) high-quality, consistently measured ADME data; (ii) a set of relevant structural descriptors; and (iii) statistical algorithms to identify the relationship between the experimental outcome and the structural descriptors. Figure 21.1a illustrates the common QSAR modeling workflow, and Figure 21.1b depicts the most commonly measured ADME endpoints, related structural descriptors, and statistical algorithms used in the industry.

21.2.1.1 Data Curation

A high-quality dataset is the foundation of a successful QSAR model. Therefore, data curation is often the most important and time-consuming part of the modeling effort. A series of data curation procedures is typically needed to ensure a clean

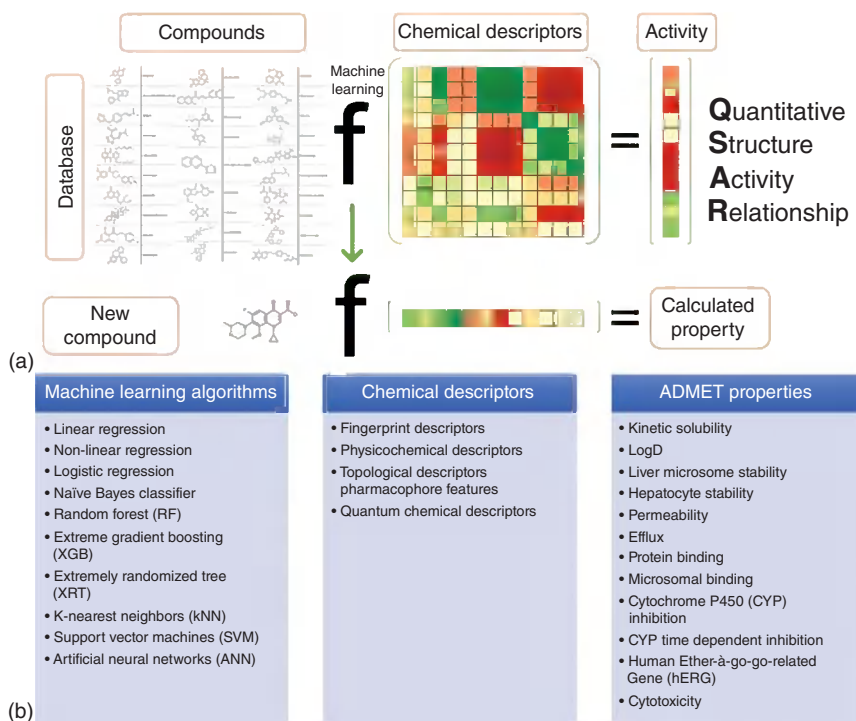


Figure 21.1 QSAR modeling cheatsheet. (a) QSAR modeling workflow and (b) commonly measured ADME endpoints, related structural descriptors, and statistical algorithms used in industry.

and model-ready dataset. There are primarily two aspects of data curation: structural curation and endpoint data curation.

Curating Chemical Structures Using the correct representation of the chemical structure is necessary for building a high-quality QSAR model. This is especially crucial when the data is collected from multiple sources. When collected from public sources, there may be duplicates, structural errors, or lack of concordance in formats. In this case, the first step of data curation would be the curation of the structures. Typically, this is done by neutralizing salts, removing erroneous or undefined molecules, inorganics, or any additional considerations based on the intended domain of chemical space. Additionally, a standardization process for handling tautomer and stereo-chemistry needs to be defined. For building 2D QSAR models, stereoisomers may be treated as equals, while for 3D models that incorporate stereochemistry-specific descriptors, stereoisomers must be retained as separate compounds. Many mature software packages for data curation are available. Examples include ChemAxon, OpenBabel, and MolVS in Python [34, 35]. Most pharmaceutical companies employing QSAR models have developed mature internal structural curation procedures that are integrated into the company's infrastructure. Typically, these procedures ensure all internal structures are curated and processed consistently before they are saved in the internal databases.

Curating Biological Assay Data It is equally important to understand the biological data and their interpretation. Factors like assay sensitivity and reproducibility are critical in informing the best practices for data aggregation and curation to minimize the inclusion of potential false negatives (FNs) and false positives (FPs) for modeling. In order to identify the most appropriate and optimal set of data for modeling purposes, QSAR modelers should work closely with the experimental scientists to develop a detailed understanding of the data, the assay's history, and the assay conditions. As pharmaceutical companies generate data over the decades, experimental protocols/conditions/reagents evolve, sites of experiment change, and data quality varies [36–40]. As a best practice, the process variables of each experiment at the time of the experiment should be captured as metadata. With that information in hand, scientists can identify the potential caveats and make the best judgment as to which portion of the data should be included or excluded when modeling. It is ideal to revisit this in order to keep the modeling data up-to-date with the ever-evolving assays.

A common practice, exemplified by a publication from AstraZeneca, is to exclude measurements that are less than or greater than the specified experimental dynamic range [41]. In their paper describing the QSAR models for P-gp efflux, authors from Eli Lilly systematically analyzed the reproducibility of the data and other factors that may influence the outcome [24]. Based on the experimental variability, the authors decided to exclude compounds with a net efflux ratio (NER) between 2.8 and 3.2 from their categorical model based on NER cut-off of 3. Similarly, they identified and excluded potential false negative outcomes based on a separate measurement of inhibition of P-gp in addition to compounds with very slow passive permeability ($<1 \times 10^{-6}$ cm/s) and very low % cell ($<1\%$).

In another case at Genentech, canine Mdr1 knockout MDCK (gMDCKI) cells were established and used for internal permeability screening in place of the conventional Madin-Darby canine kidney (MDCK) cells around 2018 [42]. Although it was shown that gMDCKI is a more representative cell line for passive permeability screening to predict the intestinal absorption of small-molecule drug candidates [42], when the new cell line was first established, the amount of gMDCKI data was insufficient for building a productive model, compared to the existing model for MDCK cell line. An analysis of 135 Genentech compounds [42] showed that the apparent permeability (Papp) from apical-to-basolateral (A-to-B) values determined in both cell lines generally correlated well, with an $R^2 = 0.86$ for $\log(\text{Papp A-to-B})$. Hence, despite the differences between the two assays, after careful consideration, it was decided to use data from both cell lines for modeling Papp A-to-B. Subsequently, as more data from the gMDCKI cell line is generated, MDCK cell line data could be eliminated from modeling in the future. Another example of data curation is the exclusion of compounds with less than 70% or greater than 130% recovery [42], in order to ensure that the most reliable data are used for modeling.

Whether the data are obtained from the public domain or from internal experiments, duplicate experiments for a single compound are frequently present. In this case, aggregation of the replicates is required. Data could be aggregated using arithmetic or geometric averages, medians, or more rigorous approaches to flag

potential outliers before calculating such parameters. The choice of aggregation strategy should depend on the data in hand and be informed by the data distribution as well as an understanding of the experimental variability. Authors from AstraZeneca presented an exhaustive analysis of the influence of experimental variability on QSAR models for eight ADME endpoints. In this case, molecules with repeat measurements whose standard deviation (StdDev) was greater than twice the typical StdDev for that response variable were removed. Where molecules had acceptable repeat measurements, the mean value was considered [41]. Typically, IC_{50} s, EC_{50} s, AC_{50} s, and other concentration-related parameters that are often log-normally distributed should be aggregated using a logarithmic scale.

For many machine learning algorithms, the data distribution of the experimental values has a significant impact on model performance [43]. Many models may over- or underpredict when the data are not evenly distributed across the experimental range. Skewed or complex data distributions may complicate uncertainty estimates and inferences around the mean, median, or mode of the data. Therefore, when modeling continuous values, scaling or transformation of the data is often advised. Transformations aim to create a more balanced data distribution. These commonly include logarithms or roots, which can make datasets more amenable for modeling. Perhaps the most common transformation example is the transformation of IC_{50} s to pIC_{50} s (Eq. 21.1). Unlike IC_{50} s that could be log-normally or log-logistically distributed, pIC_{50} s tend to approach normal distributions [44, 45]. Thus, the transformation allows one to compute Gaussian confidence intervals and enables more robust inference around the central point estimate. Data from many toxicological endpoints is transformed in this way.

By way of example, at Genentech, *human ether-a-go-go-related gene* (hERG or *KCNH2*)-encoded K⁺ channel (hERG) inhibition, bile salt export pump (BSEP) inhibition, primary human hepatocyte (PHH) cytotoxicity, and human liver microtissue (hLiMT) assays are typically screened in dose–response formats. When building models for these endpoints, the IC_{50} s derived from the dose–response curve are transformed into pIC_{50} s (Eq. 21.1).

$$pIC_{50} = -\log_{10}(IC_{50}) \quad (21.1)$$

However, some experiments do not produce an IC_{50} in default analysis. This usually happens when the response in all tested concentrations of the dose–response curve does not exceed 50% inhibition. While some guidance suggests that IC_{50} s should not be estimated from these data [46], empirical evidence suggests that IC_{50} s can be extrapolated from the data when some response is observed [47]. Internal Genentech analysis suggested that the IC_{50} s for hERG inhibition can be reliably extrapolated when a response that exceeds at least 20% of positive control is observed in at least one tested concentration [48]. These data extrapolation practices should only be used when the raw concentration–response data are available for manual examination [40, 49]. These data extrapolation practices are particularly impactful when developing models for toxicity endpoints. As discussed in more detail later in the chapter, toxicity endpoints often have limited data for model development. Consequently, computational toxicologists at Genentech attempt

to make the most of the existing data. Using extrapolated pIC_{50} s to supplement standard pIC_{50} s estimated from full dose–response curves increased the modelable data set for hERG inhibition fivefold and enabled the development of quantitative models for hERG inhibition. When extrapolating IC_{50} s, a simplified hill equation is used (Eq. 21.2)

$$\%Inhibition = \frac{100}{1 + \frac{IC_{50}}{[concentration]}} \quad (21.2)$$

Here, the upper asymptote of the hill equation was set to 100% inhibition, i.e. the inhibition observed in the positive control. The lower asymptote of the hill curve was fixed to 0% inhibition, i.e. normalized response observed in DMSO negative controls. The hill slope was set to 1, but may be adjusted based on the analysis of the empirical slope distribution observed in an assay. The resulting 1-parameter Hill curves were fitted using a likelihood-based optimization routine via gradient-free Nelder–Mead optimization method, as previously discussed [50].

Another example comes from the prediction of protein binding. In early drug screening, accurately predicting protein binding values is key for early IVIVE analysis in the absence of measured protein binding data [51]. However, the models directly predicting fraction unbound (f_u) not only lacked emphasis on accurately predicting and differentiating highly bound compounds [51–53], but also had tendencies to underpredict f_u due to the highly imbalanced distribution of compounds. Thus, the use of a log-scaled f_u ($\log f_u$) [54] or a pseudo-binding constant ($\ln K_a$ or $\log K_a$) [55] is more suitable for modeling.

21.2.1.2 Chemical Descriptors

When building a quantitative relationship between chemical structures and biological activities, the chemical structures need to be converted to a numeric format. In cheminformatics, this is often achieved through the calculation of chemical descriptors. Various types of descriptors have been developed. Common descriptor types [56–59] include fingerprint descriptors [60, 61], physicochemical descriptors [62], topological descriptors (connectivity and shape indices, adjacency, and distance matrix) [58, 63], pharmacophore features [64, 65], and quantum chemical descriptors (partial charge descriptors) [66]. Over the years, many software packages have emerged to calculate a variety of descriptors. Examples of commonly used commercial software for this purpose are Molecular Operating Environment (MOE) [67] and Dragon [68]. Several open-source packages or modules in different programming languages are capable of calculating descriptors with more flexibility [69] and can easily be integrated with automated QSAR modeling pipelines. Examples include RDKit (<http://www.rdkit.org>) and ChemoPy [70] in Python and ChemmineR [71] in R, among others.

The quality of descriptors greatly affects the efficiency and predictivity of QSAR models [56]. In cases where there are not enough representative descriptors, the chemical structures might not be properly represented and differentiated from one another, and the QSAR model may not reach its full capacity [56], even if the model is built using the optimal dataset and the most advanced machine learning

algorithm. On the other hand, where too many irrelevant or redundant descriptors are used to derive the relationship, it becomes exponentially harder for the model to find the optimal set of descriptors. In order to avoid such cases, the types and number of descriptors should be carefully selected using appropriate dimensionality reduction approaches (see below). Studies show that model predictivity can be significantly improved by optimizing the variety and number of descriptors [56], since introducing a large number of descriptors that may not be relevant to the endpoint being measured may lead to overfitting.

Dimensionality reduction to optimize the number of relevant descriptors enables the construction of efficient and meaningful QSAR models, especially when the number of descriptors used is higher than the number of data points [72]. Interpretability of the model also suffers as the number of descriptors grows, as well as generalizability to new chemical space/series. Thus, the goal of dimensionality reduction is to find a good balance between minimizing the dimensionality of descriptors and not having to lose significant, relevant, and useful information about the chemical structure. Common strategies employed to reduce the dimensionality include simply removing highly correlated descriptors [10], principal components analysis (PCA), and autoencoders (a neural network (NN) approach [73, 74]).

Feature scaling to normalize the descriptor values is often useful for some machine learning algorithms when the chemical descriptors have wide-ranging values (for example, when using both binary descriptors like fingerprints and continuous descriptors like molecular weight). If not normalized, descriptors of different ranges can be regarded as having different weights and importance by some machine learning algorithms like k-nearest neighbors (kNN), or when calculating the Euclidean distance between compounds. Commonly used methods include min–max normalization, which rescales all descriptors in the range [0, 1] or [−1, 1]; and variance standardization, which makes the values of each descriptor zero-mean and unit-variance [10]. The selection of appropriate methods for feature scaling depends on the number and distribution of descriptors as well as the method used for building the QSAR model.

21.2.1.3 Algorithms

Once the data and descriptors are curated and processed, QSAR models are ready to be built. Various algorithms can be applied for different purposes. A machine learning algorithm is a computational process that uses input data to achieve a desired task in a “soft coded” fashion, such that it automatically alters or adapts its parameters through repetition in order to become better at performing the desired task [75] (e.g. predicting ADMET properties). There are three types of machine learning: supervised learning, unsupervised learning, and semi-supervised learning [76, 77].

Supervised learning is when descriptors derived from the chemical structures are paired with an ADMET property of interest, such that model training is “supervised” by the ADMET property and the model learns which features are relevant to predicting the property. Supervised learning is the most commonly used machine

learning technique and has various subcategories such as linear regression, nonlinear regression, logistic regression; Naïve Bayes classifier [78], decision trees [79] like random forest (RF) [80, 81], extreme gradient boosting (XGB) [82], extremely randomized tree (XRT) [83], kNN [84, 85], support vector machines (SVM) [86, 87], and artificial NNs (ANN) [88]. There are several packages and modules available in different programming languages, such as scikit-learn [89] in Python, and RF [81] and e1071 in R [90].

A detailed description of the differences between these methods is largely beyond the scope of this chapter, and readers are encouraged to refer to the papers cited for such details. One of the main differences is the manner in which these algorithms handle descriptor selection. Methods like RF, XRT, and XGB can handle relatively large descriptor spaces while selecting the most important features. On the other hand, algorithms such as regression and SVM are likely to be negatively impacted by a large number of irrelevant features, and users are advised to perform feature selection before using them [91]. At the same time, it should be noted that the performance of algorithms like RF and XGB, which are relatively stable when using a large number of features, may start to deteriorate with an increasing number of irrelevant features [92–94]. This is likely due to the fact that such methods select a subset of features for different decision trees, and as the number of irrelevant features increases, the probability of selecting relevant features for a given tree can decrease.

Unsupervised learning describes a set of algorithms that learn patterns in the data without regard to the response labels (experimental outcome). The types of data learned can range from chemical structures or descriptors to biological knowledge graphs. Since these models learn patterns without an explicit guide, they are referred to as “unsupervised.” Some common unsupervised algorithms include principal component analysis (PCA) [95, 96], stochastic neighbor embedding (SNE) [97], and uniform manifold approximation and projection (UMAP) [98, 99]. Most of these algorithms are typically used for dimensionality reduction tasks. For example, unsupervised clustering methods are used to group observations into categories. In chemistry, this typically means grouping similar compounds based on structure or function. As a result, such groups can be visualized in chemical space to help make decisions about chemical synthesis, model applicability, or project progression. When it comes to QSAR models, chemical grouping and clustering are often used to assess model applicability and chemical space coverage. It should be noted that the use of unsupervised algorithms for applicability exploration is somewhat debatable [2, 100, 101]. It is proposed that clustering based on a given set of chemical descriptors/fingerprints simply to categorize them into subgroups might have limited relevance to individual ADMET endpoints since each endpoint is likely to have a unique set of relevant features.

Semi-supervised learning is a fusion of supervised and unsupervised learning, where only a portion of the data, typically small, is labeled and the rest of the data are unlabeled (unassigned experimental values) [76]. In ADMET applications, the unlabeled portion often describes the part of the data set where no ADMET properties have been measured and only information derived from chemical structure is available. A small portion of the data set may contain ADMET-related labels from

biological assays. Semi-supervised learning can improve the efficiency and predictivity of a model when the training dataset is relatively small, compared to conventional supervised and unsupervised learning [102]. Supervised learning models may not have enough data to learn basic patterns of the descriptors or other types of machine-engineered chemical representation features (see Section 21.2.3 for details) in these small data sets. In this case, an unlabeled dataset can be used to pre-train the model in an unsupervised fashion and transfer insights from the data patterns to the setting of model training with a labeled dataset [76, 102, 103].

One of the factors to consider while selecting the optimum algorithm and the corresponding chemical descriptors is the speed of generating predictions. This is especially important when using these models to screen a large number of virtual ideas (hundreds of thousands to millions). As the complexity of a model structure increases, allocated calculation resources need to be increased, and the infrastructure needs to be customized to compensate for the increased number of calculations, in order to deliver model predictions to the users in a reasonable time frame.

21.2.1.4 Model Validation and Evaluation Metrics

While models may be able to fit the data they were trained on, their performance can degrade significantly when predicting external datasets. Thus, in order to prevent overfitting and evaluate the model's true predictivity, leave set-out strategies should be applied during model building. This includes cross-validation methods such as leave one out and leave some out. In these strategies, the model is built on part of the dataset and evaluated against the portion that was left out during training. This procedure is repeated many times until all compounds have been in the test set at least once. It is worth noting that such cross-validation, while useful in exploring various descriptors and/or model-building methods, can also provide an over- or under-estimation of model performance in real drug discovery settings [34, 104] when coupled with random selection (too optimistic) or leave-class-out selection (too pessimistic). In this case, time-split selection should be used in addition to random selection as a standard for model evaluation [104].

In an industrial setting, typically ADMET QSAR models are built using all available data to maximize the coverage of existing chemical space, and cross-validation is used to estimate model predictivity at this stage. As project teams synthesize and test *new* compounds, those data constitute the prospective test set. The models are then continuously evaluated with the new data, so the scientists are aware of how well they perform in real-time, both with respect to their overall performance and the specific applicability for certain projects or chemical series of interest [24, 28, 29].

Various evaluation metrics are routinely used for model performance estimation. For models built on continuous endpoints, commonly used evaluation metrics include Pearson's multiple linear correlation coefficient (R^2) [105], mean absolute error (MAE) [106], root mean square error (RMSE) [106], mean fold error (MFE), and percentage within n -fold. For models built on categorical endpoints, the confusion matrix (as shown in Table 21.1) is usually used, where it reports the number of true positive (TP), true negative (TN), FP, and FN predictions. The predictivity of a categorical model is also evaluated using metrics such as sensitivity, specificity,

Table 21.1 Confusion matrix for binary classification.

Confusion matrix		Experimental value	
		Positive	Negative
Predicted value	Positive	True positive (TP)	False positive (FP)
	Negative	False negative (FN)	True negative (TN)

accuracy, and precision [24]. When evaluating model predictivity for an imbalanced dataset, that is, when the number of observations in the two classes varies greatly, balanced accuracy, G-means (GM), Matthews correlation coefficient (MCC), and Kappa index are the most appropriate statistics in order to avoid bias toward the majority class [107].

21.2.1.5 Utilizing Relevant Properties to Improve QSAR Model Performance

Research has shown that including relevant bioactivity information in the modeling process can improve the performance of traditional QSAR models [10, 33, 108]). This could be quite useful as some assays have higher throughput and are thus conducted in earlier tiers than others. Due to the larger dataset size and chemical space coverage as well as the mechanistic relevance, the utility of these assays to predict later-tier assays can largely improve the predictivity of the models that are related. For example, $\log D$ is an important physicochemical property that relates to several ADME properties. Similarly, the liver microsome stability assay, which is typically a first-tier ADME assay in industry, is correlated to the hepatocyte stability of the same species. Strategies that exploit these related bioactivities include multitasking machine learning algorithms and transfer learning strategies [33]. Alternatively, simpler model dependency strategies like the direct utility of bioactivities as additional features besides conventional chemical descriptors are also employed when building QSAR models [10, 108, 109].

21.2.1.6 Applicability Domain and Reliability of Individual Prediction

There are two major contributors to the reliability of predictions from a QSAR model: applicability domain and prediction boundary. With regard to the applicability domain, if a compound is outside of the chemical space of the model training dataset, called an outlier, it is regarded as out of the applicability domain [110]. The prediction boundary refers to the range of predicted values across which the model is expected to provide reliable predictions. This is mainly determined by the range of experimental values captured by the model training set. A compound is likely to be out of the applicability domain when the model's training set lacks the relevant region of the chemical space and hence is unable to extrapolate to it accurately. While the model may still make predictions for such compounds, such

predictions are likely to be less reliable, and decisions should be made accordingly. For example, a QSAR model trained on small molecules with MW < 1000 can be misleading when used for predicting the ADME properties of a macrocyclic peptide with MW approximately 1000–10 000 Da.

In practice, as the therapeutic program progresses, medicinal chemistry evolves, and as such, the team may require predictions for compounds in regions of chemical space that fall outside of the existing models' applicability domains. Consequently, continuous model evaluation, re-training, and expert judgment about prediction accuracy and model applicability to new chemical spaces are critical. Our previous study [11] showed that compounds synthesized within the 6-month period following a model release have a significant drop over time in the similarity to the model training set. As the chemical space of interest expands over time, an overdue model that does not include recent data will have significantly more out-of-applicability domain predictions and consequently reduced prediction accuracy over time [24].

As discussed above, model performance in new chemical space is an essential part of ADMET model evaluation and validation. Even predictions within an applicability domain may be affected by activity cliffs. An activity cliff is a pair of structurally similar compounds that vary substantially in activity against the same target [100]. When an activity cliff is present, a model may give inaccurate results for one of the compounds in the pair. This phenomenon usually occurs when the training dataset does not have the level of detail necessary to differentiate the activity cliffs. However, in some cases, the current state of knowledge has no theoretical explanation for the differences in activity, and only empirical evidence can distinguish molecules on the two sides of the cliff.

The study of the applicability domain allows estimating the uncertainty in the prediction for a particular molecule based on how “similar” it is to the training compounds [111]. There are various methods for evaluating the applicability domain of the model and assessing if a given compound is in or out of the domain. Typically, the applicability domain of a model is dependent on the training dataset and not on the machine learning algorithm. It is generally determined by calculating the chemical similarity or chemical distance, *in the descriptor space* used for the model. The more dissimilar a compound is from the training set, the more the model tends to extrapolate and the less reliable the predictions are. A common metric to evaluate the chemical similarity or distance (i.e. dissimilarity) is Euclidean distance, which calculates the geographical distance between two compounds in chemical space using normalized descriptors, typically, fingerprint descriptor-based chemical similarity using Tanimoto coefficient, dice coefficient, or the cosine coefficient. It is recommended that dimensionality reduction be applied before chemical similarity or distance calculations. PCA analysis can also be leveraged for the evaluation of the applicability of models by visualizing the distance of the compound from the model training set. Regardless of the method, it is critical that the similarity be based on the descriptor space used for the model, rather than a generic similarity based on standard fingerprints [100, 110, 112, 113].

There are various confidence estimation techniques that aim at quantifying the reliability of predictions. One example of the probabilistic predictions at Genentech is [11] the set of QSAR models predicting liver microsome stability, i.e. clearance

Cl_{hep} , where predictions from regression models are converted into a probability of a compound being stable. The reported probability provides the likelihood of a compound being stable. Another common practice is to use bagging and bootstrapping strategies to get multiple individual models that form an ensemble, and then to compute ensembled predictions alongside their variability across the panel of models [114–116]. Conformal prediction [113, 117] is another type of error estimation tool that can predict the reliability of individual predictions. Its recent popularity is due to the ease of interpretation of the computed prediction errors in both classification and regression tasks, as well as the feasibility of coupling it to any machine learning algorithm at little computational cost [117].

21.2.1.7 Model Interpretability vs. Predictivity

The earliest QSAR models were built with simplistic linear regression or decision trees using only a few physchem properties. These made the models interpretable to a large extent. However, with the increasing size and complexity of experimental data, especially in an industrial setting, more complex algorithms are typically required to effectively learn and predict from such data. As the field enters the era of an “AI paradise” with more sophisticated machine learning algorithms, the accuracy and breadth of predictions have increased at the expense of interpretability [100]. For a given data set, if an equally predictive model can be identified, which also affords interpretability to some extent, it is likely to gain better acceptance by end users.

Having said that, it should be important to note that the successful application of QSAR models requires a thorough understanding of the goals of a project team at any given stage, and interpretability may not always be strictly necessary. The complexity of biological systems, including those underlying various endpoints of ADME, cannot always be modeled by linear or “simple” models. In such cases, more sophisticated models may offer significant improvements in predictivity. Accordingly, when sufficient resolution is required to effectively triage or rank order a large set of virtual design ideas, a more complex but less interpretable model often proves to be effective in achieving such goals without limiting its acceptance by end users. Thus, there is room and need for both types of models.

21.2.1.8 Model Deployment and Accessibility

In order to ensure the maximum utilization and impact of a QSAR model in an industrial setting, it needs to be integrated seamlessly into the drug discovery pipeline. This is to ensure ease of access and interpretation of these models by end users, such as medicinal chemists. QSAR models are usually made available through various platforms and graphical user interfaces (GUIs), which are typically used by project teams for day-to-day data access and exploration. For more interactive design sessions, such models can also be deployed to generate real-time predictions as the user interactively modifies a design structure. In some cases, such models may need to be implemented in additional formats depending on their intended use. For example, for generating predictions for a large virtual library (thousands to millions of virtual structures) the models are typically run from their command line version on a high-efficiency server.

21.2.2 In Silico ADMET Models and Their Influence During Drug Discovery

21.2.2.1 In Silico ADME Models

At the lead generation stage, projects generate compounds that aim to increase knowledge about their respective chemical spaces, which can lead to informed decision-making [118]. At this stage, medicinal chemists build up knowledge about various chemical series of interest using a range of in vitro assay systems, assessing potencies, physchem properties, as well as ADMET properties [119]. When entering the hit-to-lead stage, the projects aim to refine each series to try to produce more potent and selective compounds, which possess PK properties adequate to examine their efficacy in suitable in vivo models [119]. Key compounds that meet the criteria of potency and ADME properties are then assessed for PK in preclinical species. During the lead optimization stage, the objective becomes to maintain favorable properties in lead compounds while improving on deficiencies in the lead structures [119]. Through the application of quantitative and statistical QSAR models during the learn and confirm cycles from lead generation to lead optimization, projects can benefit from a reduced number of iterative cycles in exploring the SAR in a chemical series as well as the ability to allocate the finite resources of in vitro and PK studies to more promising compounds/series.

A recent perspective from the in silico ADME IQ Consortium (i.e. International Consortium for Innovation through Quality in Pharmaceutical Development) [12] showed that (Figure 21.2) across member companies, a significant improvement of measured ADME properties is observed across all projects after the adoption of QSAR models, e.g. microsomal stability at Genentech, solubility at AstraZeneca, and CYP3A4 time-dependent inhibition at Eli Lilly.

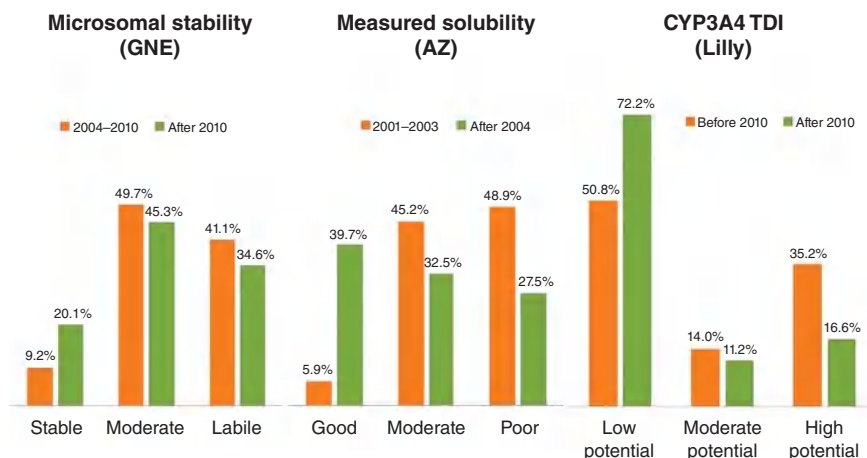


Figure 21.2 Changes observed across all projects in microsomal stability at Genentech (GNE), solubility at AstraZeneca (AZ), and CYP3A4 time-dependent inhibition at Eli Lilly (Lilly) after the adoption of in silico models. Source: Reproduced with permission from Lombardo et al. [12] American Chemical Society.

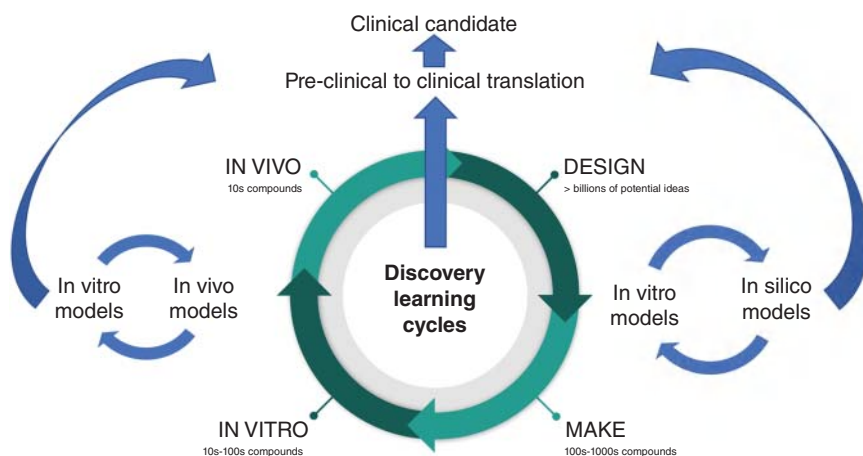


Figure 21.3 Integrated and iterative use of models in early-phase drug discovery, showing the recommended process to identify and integrate in silico, in vitro, and in vivo models.

While the “global” prospective validation is critical in establishing the trust and value of a QSAR model, it is equally important to assess its applicability to a given chemical series in question. To this end, after identifying a chemical series of interest for a given therapeutic program, a representative set of compounds spanning the range of predicted in silico values, including various physicochemical characteristics and capturing structural diversity, should be tested in the corresponding in vitro assays. As described by Danielson et al. (Eli Lilly) in their book chapter, it is equally important to explore the relationship between in vitro ADME models and the in vivo profile of compounds in order to select an appropriate suite of in vitro tools to prioritize the selection of compounds for in vivo assessment [23]. An iterative learning cycle, as shown in Figure 21.3, has been shown to be more effective than using a filtration approach where only the active compounds progress for in vitro and in vivo ADME measurements.

Desai et al. from Eli Lilly exemplified this strategy in their work on utilizing a P-gp efflux model for optimizing compound prioritization for synthesis and testing for three different therapeutic programs. They invoked custom modification of the strategy based on considerations of other physicochemical properties influencing P-gp efflux as well [24]. Another example of the application of ADME QSAR models for driving project decisions is our previous work on the use of a HLM stability QSAR model in the JAK1 project during lead optimization [11]. The chemical series in JAK1 suffered from metabolic stability issues. The decision to halt chemistry resources on a series of compounds was partially dependent on the consistently predicted poor metabolic stability across analogs in one chemical series. As the chemists focused on switching to another chemical series, predictions from the HLM QSAR model became an increasingly important filtering criterion prior to synthesis. By integrating the HLM QSAR model predictions into the process as a filtering criterion prior to synthesis, the experimental metabolic stability of the series was continuously improved, as the average measured HLM clearance of compounds

tested kept dropping over the months. Another Genentech example by Aliagas [11] corresponds to a different development stage, in which a good IVIVc of clearance had already been shown. In this case, the PI3K project used a similar strategy of prioritizing compounds for synthesis on the basis of predicted HLM clearance criteria, which resulted in only a small percentage of synthesized compounds having poor stability.

Many of the aforementioned applications of QSAR models focused on applying hard cutoffs on the predicted physchem or ADME properties to prioritize compounds to advance within the lead generation and optimization stages. At the same time, it is important to focus on multi-property optimization rather than a few properties in isolation. Different properties can come from the same underlying molecular characteristics, and simply optimizing compounds by improving one or a few ADME properties can result in another set of ADME-related liabilities [120]. The multiparameter optimization (MPO) strategy, in which we try to identify high-quality compounds with a balance of properties [121], can be leveraged to assess ADME properties as well as the potency of a compound in a more holistic way. Specifically, projects first aim to profile analogs using a broad range of assays to identify key parameters for optimization, then develop meaningful MPO scoring systems addressing prevalent issues for an advanced lead while maintaining the desired attributes for other properties [122]. In this sense, local MPO scores for a specific project or chemical series can be useful to optimize key parameters. An MPO scoring system can be used to rank order compounds before synthesis, using *all* in silico predictions. Of course, this would rely on having established a reasonable concordance between each of the in silico and in vitro endpoints constituting the MPO. As compounds progress through the assay cascade, with more experimental values available, uncertainties in the MPO scores are expected to decrease, thereby further strengthening the quality of decisions during subsequent cycles. The use of meaningful MPO scoring systems can reduce the attrition rate [121], reduce the number of design cycles, and speed up the identification of compounds with enhanced survival [123], enabling the allocation of limited resources wisely on more promising compounds [124].

21.2.2.2 In Silico Toxicity Models

QSAR models for toxicity endpoints typically revolve around in vitro assays included in the drug-candidate derisking pipeline. A few QSAR models have been developed for in vivo and clinical endpoints, such as drug-induced liver injury (DILI), nephrotoxicity, and cardiac effects [125–129]. However, given the scarcity and complexity of in vivo and clinical data, most toxicity models to date have focused on predicting defined molecular initiating events that are associated with an in vivo or clinical outcome and can be measured with an in vitro assay. Consequently, practically, most in silico QSAR models for toxicity attempt to predict in vitro assay endpoints. At Genentech, the current suite of in vitro endpoints available for modeling includes *hERG* or *KCNH2*-encoded K⁺ channel (*hERG*) inhibition [130], BSEP inhibition, PHH cytotoxicity [131], and *hLiMT* cytotoxicity [132], mitotoxicity [133], mutagenicity [134], and secondary pharmacology assays [47] (Table 21.2). The typical

Table 21.2 Selected list of current ADMET QSAR models at Genentech.

ADMET property	Modeled endpoint	Approximate training set size	Approximate R^2 range
Kinetic solubility	Log of kinetic solubility	169 K	0.5–0.6
$\log D$	$\log D$	143 K	0.8
Liver microsome (LM) stability	Log of human LM CLint	135 K	0.4–0.7
	Log of rat LM CLint	136 K	
	Log of mouse LM CLint	134 K	
	Log of dog LM CLint	130 K	
	Log of cyno LM CLint	124 K	
Hepatocyte stability	Log of human hepatocyte CLint	21 K	0.2–0.4
	Log of rat hepatocyte CLint	22 K	
	Log of mouse hepatocyte CLint	22 K	
	Log of dog hepatocyte CLint	18 K	
	Log of cyno hepatocyte CLint	18 K	
Permeability	Log of (g)MDCK Papp AB	22 K	0.5–0.8
	Log of (g)MDCK-MDRI Papp AB	10 K	
Efflux ratio (ER)	Log of (g)MDCK ER	22 K	0.3–0.6
	Log of (g)MDCK-MDRI ER	5 K	
	Log of (g)MDCK-hBCRP ER	2 K	
Binding	Human PPB log Ka	20 K	0.7–0.8
	Mouse PPB log Ka	20 K	
	Rat PPB log Ka	20 K	
	Dog PPB log Ka	1 K	
	Cyno PPB log Ka	1 K	
	Microsomal binding log Ka	4 K	0.9
Cytochrome P450 (CYP) inhibition	CYP3A4 inhibition pIC50	60 K	0.2–0.4
	CYP1A2 inhibition pIC50	60 K	
	CYP2C19 inhibition pIC50	60 K	
	CYP2C9 inhibition pIC50	58 K	
	CYP2D6 inhibition pIC50	60 K	
Cytochrome P450 time-dependent inhibition (TDI)	TDI 3A4 dAUC	8 K	0.2–0.5
	TDI 1A2 dAUC	8 K	
	TDI 2C9 dAUC	7 K	
Safety models	hERG	6 K	0.5
	Secondary pharmacology	3 K	0.3–0.8
	BSEP	1 K	0.7
	Cytotoxicity: PHH	1 K	0.6
	Cytotoxicity: HLiMT	1 K	NA
	Phospholipidosis	0.3 K	NA

CLint, intrinsic clearance; PPB, plasma protein binding; Papp AB, apparent permeability in the apical to basolateral direction; dAUC, area under the curve shift; PHH, primary human hepatocytes; BSEP, bile salt export pump; hERG, the human Ether-à-go-go-Related Gene. HLiMT, human liver microtissues. NA, no external validation metrics are currently available.

secondary pharmacology assay panel is outlined in the seminal paper by Bowes et al. [47] but may vary substantially based on project needs and pipeline experience.

The methodologies for in silico safety assessment and their applications vary vastly across industries, but commonly include a combination of QSAR models, structural alerts, chemical similarity, and clustering tools [135]. These methods can be used in isolation as an ensemble to reinforce prediction confidence or group individual predictions into the assessment of more complex endpoints. The common commercial and open-source tools used across the industry include Derek Nexus [136, 137], Toxtree [138], CASE Ultra Expert Rules [139, 140], and Leadscape Expert Alerts System [141].

One of the most widespread applications of these and other in silico tools across the pharmaceutical industry is screening for DNA reactive impurities. The manufacturing process can introduce impurities into the final product that are difficult to isolate or even identify. These impurities need to be evaluated for their potential to be DNA-reactive mutagens. The International Conference on Harmonization (ICH) M7 guideline (“Assessment and control of DNA reactive (mutagenic) impurities in pharmaceuticals to limit potential carcinogenic risk”) [142] specifies that two (Q)SAR methodologies may be used to assess impurities for DNA reactivity when no experimental data is available. The performance of commercial tools on proprietary data has been extensively evaluated by industry partners and is reviewed elsewhere [143]. Many partners found that model predictions can be improved by supplementing commercial tools with internal data that is closer to compounds of interest in chemical space [144]. In addition, many pharmaceutical companies increase their reliance on in silico methods for off-target profiling and kinase reactivity [135, 145]. These models include a combination of similarity-based tools, ligand- and target-based 3D approaches, and QSAR models. In many cases, the in silico predictions may be used in early drug development, during lead identification, and followed up with in vitro experiments during lead optimization [144, 145]. The relevance of in silico predictions to in vivo results is always of high interest in drug discovery. In validation studies for off-target effects predicted in silico, up to 50% of the off-target interactions could be validated with literature or experimental follow-up, demonstrating clear relevance for off-target potential predictions [145]. Other in silico models that may be used in lead identification include models for cytotoxicity, voltage-gated ion-channel, Kv11.1 (hERG) inhibition, CYP inhibition or induction, mitotoxicity, and phototoxicity [144]. Of these, models for hERG inhibition are often more widely applied due to the availability of larger data sets for model development in hERG inhibition. hERG blocking is associated with long QT syndrome, torsades de pointes, a potentially fatal condition [146]. Consequently, most drug candidates are screened for hERG inhibition in vitro, providing a large modeling data set. Industry scientists discussed that the implementation of in silico filters for hERG inhibition may lead to a reduction of highly potent hERG inhibitors that progress further into the pipeline [135, 147].

As a general rule, the in vitro toxicity assays are relatively expensive compared to many earlier-tier ADME endpoints, and limited in vitro data are available for model development. Historical datasets for toxicity endpoints may range from 100 s

to a few thousand compounds. Consequently, extracting the most information from historical data and ensuring data quality are paramount for robust model development. In safety assessment, computational toxicologists work to make the most of the historical data, support predictive model development, and help future projects learn from the historical data. As part of the efforts, Genentech scientists attempt to develop models for endpoints on the same scale as the outputs of the typical *in vitro* assay. For most assays, that means developing continuous models that predict pIC_{50} s (Eq. 21.1). At Genentech, the majority of toxicity assays are run in dose-response format with stable positive controls. Consequently, absolute IC_{50} s are typically derived from the dose-response data and reported to the project chemists and toxicologists. When no positive control data are readily available, such as in the case of secondary pharmacology binding assays, relative IC_{50} are calculated [47]. Continuous models for all relevant toxicity endpoints (Table 21.2) were developed based on the XGboost algorithm [148]. A combination of molecular fingerprints and physicochemical properties was typically used to describe chemical structures. The selection of descriptors is linked to their relevance and calculator speed. Consequently, the traditional Morgan fingerprints are at the core of most predictive models but can be augmented by partitioning, charge distribution, and molecular substructure-related features. Many of these descriptors can be calculated rapidly and enable fast model performance even with limited computational resources. A few of the mentioned descriptors, such as ionization constants, involve more complex calculations and can be resource-intensive. However, these are included in many models due to their direct relevance to toxicity endpoints. For example, basic centers are often implicated in hERG channel inhibition [149], and basic compounds are associated with higher promiscuity [150]. At present, the uncertainty around predictions is assessed by providing prediction errors calculated via boosting approaches. The continuous predictions on the pIC_{50} scale with uncertainty metrics allow project teams to get an idea for model relevance and rank order compounds for synthesis or further screening.

Late in drug discovery, project teams may screen large libraries to prioritize compounds for synthesis and downstream testing. The typical goal in early research is to identify regions of chemical space with high therapeutic potential and a lower probability of safety hazards. Computational models can help when no data is available. In these early stages, it is often desirable to provide models with a high positive predictive value (PPV) to ensure that the compounds identified as potentially hazardous have a high probability of being truly hazardous. This approach ensures that medicinal chemists are not eliminating chemical matter based on false-positive predictions. Since the compounds will be tested *in vitro* down the road, false negative predictions may be less concerning, but they can erode model confidence. Thus, models with high PPV have been more likely to get acceptance from chemists in early drug discovery. Later in safety assessment, it is often critical to identify and eliminate hazardous compounds. Consequently, high model sensitivity is paramount. In practice, predictions for compounds with high interest later in the pipeline are usually followed up with experimental assays. However, predictive safety models can help prioritize or rank order compounds for follow-up studies.

21.2.2.3 Typical QSAR Models for ADMET in the Industry

Typically, pharmaceutical companies with vast amounts of ADMET data build and maintain QSAR models based on the “global” dataset, with the intent of maximizing chemical space coverage to ensure the broadest possible applicability domain. For example, authors from Merck have shared an extensive list of ADMET models built using their internal and consistent set of data [151, 152]. This included both in vitro and in vivo endpoints for ADME and in vitro endpoints for safety. Similarly, other companies such as AstraZeneca [41], Bayer [28], and Eli Lilly [23] have also described the typical ADMET models routinely used during the drug discovery process to prioritize compound design, synthesis, and testing. In Table 21.2, we present the list of current ADMET QSAR models available at Genentech, along with the modeled endpoint, in order to show what endpoint transformation/scaling and handling were suitable for the endpoints internally, the training set size, and the model performance (R^2) of prospective predictions.

21.2.3 Emerging QSAR Technologies and Algorithms

Over the past decade, NN models have featured heavily in the ADME QSAR literature [7, 12, 20, 21, 23, 33, 153–161]. The accessibility of cloud computing infrastructure, the maturation of open-source deep learning (DL) libraries [162–164], and significant advancements in GPU computing power have all contributed to the emergence and development of DL approaches in drug discovery. This section addresses the key features of DL methods and outlines the challenges associated with them. Moreover, in addition to referring to studies from other companies, we share an internal Genentech case study examining the performance of deep multitask NNs in the ADME prediction setting. The promise of NN models generally is to mitigate human bias in the learning process, primarily by obviating the need to specify or engineer custom molecular descriptors (e.g. molecular fingerprints, physicochemical properties, shape descriptors, pK_a estimates, among many other possibilities). Instead, NNs can extract relevant features *directly* from the dataset using an end-to-end differentiable learning algorithm. In other words, molecular representations can be *machine-engineered* according to the task at hand, as opposed to being predetermined. Of course, as NN models are highly customizable, it is also possible to provide predefined molecular descriptors to the NN. Given the high dimensionality of synthetically accessible chemical space [165], though, the rationale behind NN models is that they may be more adept than humans at distilling structure–property relationships from large molecular datasets.

The general procedure for training a NN is analogous to that for training a classical ML algorithm (e.g. support vector machines [166, 167], XRT [83], logistic regression [167]). The first and the most important step of the process is the collection and curation of high-quality datasets, which encode the relevant relationships to be learned. For example, this dataset could consist of a mapping from molecular structures (often encoded as SMILES or SELFIES strings [168, 169]) to experimentally measured physicochemical properties like lipophilicity. Depending on the endpoint being modeled, this presents a challenge in itself owing to the data requirements of

training a NN. As NN models often consist of many stacked layers of “neurons,” with each layer containing multiple neurons, they are often *highly* parameterized and therefore susceptible to overfitting to the training set (cf. the curse of dimensionality [170, 171]) when the volume of training data is relatively small. This is often the case for later cascade assays (e.g. measuring metabolic stability in hepatocytes).

After data collection, filtering, and curation, the objective is to train the NN to recapitulate the structure–property relationship encoded by the dataset. This is achieved by: (i) specifying a loss function that expresses the deviation of model predictions from ground truth labels (e.g. in the regression context, a common choice is mean squared error); (ii) obtaining the gradient of the loss function with respect to the model parameters; and (iii) updating the parameters in the direction of the gradient (i.e. gradient descent). Steps (ii) and (iii) repeat until a local optimum of the loss function is found, as specified by a convergence criterion [172].

Let us consider the mechanics of the above learning algorithm in more detail. First, a fixed-dimensional matrix representation of the molecular structure data is prepared as input to the model. In the case of a graph NN (GNN), for example, these matrices encapsulate information about molecular nodes (e.g. atom types), graph edges (e.g. chemical bond types), and overall node connectivity [173]. In the *forward pass*, these molecular structure data are passed through the NN to obtain predictions that correspond to the current model parameters. Throughout the forward pass, the initial molecular representations undergo a series of nonlinear transformations consisting of matrix multiplication operations between the model parameters and the features, followed by the application of *nonlinear activation functions* (e.g. sigmoid, relu, tanh). While the precise details of these transformations depend on the chosen modeling paradigm (e.g. natural language processing, convolutional NNs, graph convolutional networks), the core idea is consistent: the learned representations are constructed iteratively, with each layer of the NN composing increasingly complex representations from the simpler preceding representations. The result of the forward pass is a vector of predictions, which are then used to compute the overall loss with respect to the ground truth labels.

The gradient of the loss function is obtained via backpropagation, which facilitates the efficient computation of partial derivatives of the loss with respect to the model’s parameters. This is obtainable because the exact sequence of mathematical operations comprising the forward pass is known, and therefore the gradient of the loss may be propagated backwards by invoking the chain rule of differential calculus [174].

Now that the contours of the learning algorithm for NNs have been described, let us consider an example of how such models have been explored within predictive ADME programs in industry, including at Genentech. One of the first systematic explorations of NN models for industrial ADME data was shared by Ma et al. [151], where they conducted a systematic comparison of RF and DNN models for a series of end points, including several from ADME. They found that while the DNN models appeared to outperform the RF models, the magnitude of the change in R^2 relative to RF appeared to be small for most datasets. They also provided insights into the effects of some key parameters on DNN’s predictive capability and

suggested a set of values for all DNN algorithmic parameters for large QSAR data sets in an industrial drug discovery environment. For example, they found that most single-task problems could be run with two hidden layers with fewer neurons (1000 and 500) and fewer epochs (75). Authors from Eli Lilly [157] conducted an exhaustive search of the optimum hyperparameters for DNN models for ADME data sets. They evaluated their in-house implementation of SVM models (benchmark models) with DNN models for 24 ADME end points, including both numerical and categorical models. In their findings, after applying the optimized parameters for each DNN model, the performance of DNN vs. SVM models was mostly equivalent based on two chronological test sets. Small but statistically significant improvement in performance was observed for the DNN models for end points with relatively larger training sets (>80 K), like the microsomal stability data, while the opposite behavior was noticed for relatively smaller training sets (<25 K), like the MDCK permeability.

In silico modeling has a long history at Genentech [11, 12, 16, 31–33, 175], and the use and development of modern ML algorithms have accelerated in recent years. For example, recent work [33] sought to benchmark the performance of GNNs against “traditional” QSAR approaches (e.g. XRT) by making use of Genentech’s large-scale historical ADME data as well as an external test set comprising Roche chemical space. As it is widely appreciated that establishing an ADME profile for a new molecule involves assays that vary in complexity, time, cost, and throughput and that the volume of data across ADME endpoints becomes progressively smaller as assays increase in complexity and cost, this presented a natural opportunity for the authors to evaluate both multitask (MT) learning and transfer learning. The remainder of this section will be concerned with the mechanics of MT network training, the core findings of Broccatelli et al., and some comments on future directions at Genentech.

In MT learning, multiple endpoints are modeled simultaneously in order to learn a shared molecular representation that is jointly predictive [153, 176, 177]. MT learning may be interpreted as a regularization strategy that exploits information transfer between related endpoints in order to learn more robust and generalizable representations. This strategy is especially appealing when modeling low-volume endpoints or data from the same endpoint that is obtained under multiple experimental protocols.

The mechanics of training a MT NN are much the same as training a single task (ST) NN. While the premise of simultaneously learning multiple QSAR relationships may at first seem more complicated, in practice this can be achieved simply by augmenting the loss function with terms associated with the additional endpoints. During backpropagation, the gradient of the total loss is used to update the model’s parameters, so loss terms corresponding to *all* of the endpoints influence the optimization trajectory.

In the work by Broccatelli et al. [33], GNN ST and MT architectures were benchmarked against classical ML approaches based solely on molecular fingerprints, as well as internal modeling workflows that leverage more custom descriptors and assay interdependencies *in addition* to fingerprints. For certain endpoints, they also had access to an external dataset of Roche data (enabled by the close ties

between Genentech and Roche), which afforded the opportunity to evaluate model generalizability to a more distant chemical space. Ultimately, XRT models trained on molecular fingerprints alone performed substantially worse than all other architectures explored, while the gap between the NN architectures and the XRT models built using more heavily curated descriptors was significantly narrower. This demonstrated the importance of benchmarking new methods appropriately and that publications relying solely on the use of molecular fingerprints in their benchmarking may overstate the outperformance of DL approaches over classical ones. Graph attention (GAT) networks outperformed other architectures in both the ST and MT settings, with the GAT ST model generalizing best to the Roche chemical space. However, when considering ST vs. MT GAT networks, the MT architecture only marginally outperformed the ST models, implying that transfer learning between the tasks was not as significant as anticipated. Ultimately, the DL algorithms exhibited strong performance and an expanded applicability domain, but additional research is required to fully explore the extent of their impact on the in silico ADME space.

In summary, NN algorithms can learn data-driven chemical features that are unbiased by human knowledge, as well as leverage data that are related semantically but cannot be directly pooled together (e.g. assay data from two separate companies corresponding to the same endpoint). These attributes result in both improved model predictivity [33] and an expanded applicability domain when compared to traditional QSAR models (e.g. RF).

21.3 Extended Scope of In Silico ADMET

21.3.1 Exploring New Chemical Space With Generative Models Considering ADMET

As previously discussed, in silico models assist in the lead generation to lead optimization of drug discovery projects by predicting ADMET properties of compounds; however, most QSAR models can only assess the properties of a compound proposed by a human and do not generate new chemical matter ideas. Alternatively, de novo molecular design tools can leverage the information from large databases to generate or suggest promising new compounds [178].

Matched molecular pair (MMP) analysis has emerged as an attractive computational approach to guide compound design and optimization due to its simplicity and interpretability. An MMP is a pair of compounds that only differ by a single localized structural change (i.e. transformation) [179]. An MMP analysis can extract all data from a given database, pair compounds, summarize, analyze, and ensemble all pairs of compounds consisting of certain transformations and their associated statistics (average property difference, StdDev, etc.) [180]. Building up an MMP database that summarizes structural transformations and their likely effect upon a broad range of properties (e.g. Log *D* [181–183], human liver microsome stability [183–185], aqueous solubility [182, 183, 186], plasma protein binding [183, 186], oral exposure

[186], hERG, and P450 metabolism [187]) has been shown to be useful during the lead optimization process when relatively conservative structural perturbations may be considered around a particular molecular core [4]. It can serve as a generative tool to find and suggest structural modifications that fit the need of improving ADMET properties. For example, MMP applications can suggest structural modifications that may improve metabolic stability, while maintaining permeability within a given range, without changing the potency core structure. The approach, which utilized all the data and knowledge across projects, often made more suggestions than could be processed manually. MMP applications can enable medicinal chemists to get a more objective sense and comprehensive view of how particular structural modifications on a compound may affect its key physchem or ADMET properties. Combining the QSAR model's predicted properties with the suggested new compounds as a comprehensive MPO scoring system, it provides a streamlined process to suggest, assess, and rank order compounds to be synthesized. Yet, it is worth noting that there are various aspects that MMP does not consider explicitly, e.g. synthesizability, and thus it will not be successful without medicinal chemists' manual input. In addition to MMP analysis, there are more sophisticated generative AI methods for automatically proposing novel chemical structures that optimally satisfy a desired molecular profile [188]. There are atom-, fragment-, and reaction-based approaches for generating novel structures, with various assessment methods to benchmark performance [188–191], using different molecular representations (text-based like SMILES string or graph-based, etc.), trained with a variety of molecular optimization algorithms either gradient-based or gradient-free [188]. The *de novo* molecular design and generative methods are still in the method development stage, and their applicability in drug discovery remains both theoretical and somewhat controversial [188, 192–194]. Several trial applications of such generative models have been evaluated in the field of drug discovery [195]. Merck and colleagues developed a generative model that designed compounds that are retinoid X and peroxisome proliferator-activated receptor agonists [196]. Researchers at AstraZeneca expanded the chemical space by tuning a sequence-based generative model to design compounds with almost optimal values for solubility, PK properties, bioactivity, and other parameters [197].

It should be noted that benchmark suites for *de novo* design are the most important yet most challenging task for creating a useful generative model for drug discovery. It is multifaceted, and current *de novo* design efforts are limited by a narrow view of the overall process [188]. When applied in drug discovery, it should be flexible and customizable at the project level as each project has its own challenges and key sets of properties to optimize. Most importantly, the key to success is to build trust and partnership between computers and humans – the computationally proposed compounds should be manually reviewed and evaluated by scientists in the project for the many facets that have not been undertaken by the models.

21.3.2 Mechanism-Based Models

It is the goal of drug discovery efforts to provide the safest and most efficacious drugs to patients. Consequently, the ultimate test of every *in silico*, *in vitro*, or *in vivo*

model is its relevance to humans. At each stage of drug discovery, with either the *in silico*, *in vitro*, or pre-clinical *in vivo* data at hand, the ultimate target is human prediction. In early drug discovery, various empirical MPO systems are developed combining physchem and ADME properties in order to explore, optimize and prioritize compounds [198]. Most of these MPO systems either set hard cutoffs for key properties [199] or build statistical functions to score the probability of compounds being successful [200]. Yet, the biological system is more complex than a few parameters, and some properties are interrelated, which makes it challenging to identify an optimum MPO scoring system. Similarly, such MPOs become less useful when projects are advanced to a later stage when the goal is to address prevailing issues without negatively impacting the existing favorable properties of the lead compounds [122]. The use of fully “bottom-up” physiologically-based pharmacokinetic (PBPK) modeling strategies early on, combined with QSAR models for the underlying ADME end points, has been proposed to be promising to enable human PK predictions that serve as MPO, from researches at GSK and Roche [18, 19, 201, 202]. This provides rank ordering of compounds holistically and mechanistically based on underlying properties. When combined with QSAR models for key properties, this strategy can be applied to virtual compounds for prioritization for synthesis. After compounds are synthesized and tested experimentally, the measured properties can replace the predicted ones to help reduce uncertainty. It has been proposed that the development and integration of such methods can potentially reduce discovery cycle times and animal experimentation. Admittedly, during early discovery stages, relative uncertainty is higher in predicting human dose (for example, *in silico* to *in vitro* disconnects from prediction errors, *in vitro* to *in vivo* disconnects from mechanisms that are not captured in the generic PBPK models). However, with the data, information, and knowledge that is at hand at this stage, this approach is still expected to be better than using the “traditional” MPOs, which lack the ability to integrate the net balance of the properties required to achieve the desired PK in the clinic. Plus, it provides mechanistic insights and enables ADME scientists to influence compound design.

21.3.3 Predictive MetID (Metabolites Identification From Chemical Structures)

Drug metabolism is one of the most important processes involving a drug after administration [203], as it influences the PKs and pharmacodynamics of compounds while altering their pharmacological activity (a.k.a. biochemical reactivity/potency) and toxicities [22, 204, 205]. The prediction of metabolic sites can be beneficial to help narrow down potential sites of metabolism when the experimental identification is ambiguous with partially solved metabolic structures. An accurate prediction of the metabolic site either experimentally or computationally can allow chemists to understand, avoid, or modify the fragments with potential metabolic liabilities.

There are three types of computational methods to predict sites of metabolism, i.e. rule-based, physics-based, and machine-learning-based [204]. The rule-based methods use empirical rules to predict the sites of metabolisms of compounds based

on historical data. The physics-based methods are usually performed with docking or quantum mechanical (QM) simulations leveraging the 3D structure of the compounds, which are assessed on the best fitting pose against the reaction sites of various metabolic enzymes, majorly for CYP450 [204, 206–210], as well as other non-CYP enzymes [23, 211–214]. The machine-learning-based methods use computational algorithms to allow automatic learning from the previous experimental MetID datasets. These ML models typically utilize custom descriptors to capture compound's likelihood of binding in the active site of key metabolizing enzymes and those related to quantum chemical estimations to capture the lability of the atoms [215–217]. Each of these types of predictive MetID methodologies has its pros and cons while most of the available MetID software packages use more than one of these strategies to comprehend the metabolite prediction. Readers should refer to other articles for a more comprehensive overview [22, 204].

Nevertheless, although the predictive MetID software has been shown to be quite useful tool, caution should be used when applying such predictive tools. It is best practice to properly assess the predictions case by case, combine these with experimental MetID reports if available, in consultation with a MetID scientist to minimize overinterpretation. It should also be noted that removing or modifying the dominant metabolic site might not necessarily improve the overall metabolic stability significantly, as other metabolic sites can become dominant (i.e. metabolic switching). Plus, structural changes for mediating metabolic liability may perturb many other key physchem and ADME properties that might be undesired. This is also aligned with what our Lilly colleagues tried to appeal [23].

21.4 Conclusion

In silico tools are becoming increasingly popular in drug discovery. They provide fast, cost-effective information, often when no other information is available. Consequently, predictive methods aid in enabling scientists to deliver efficacious, safer drugs to patients faster and at lower cost. However, like any other tool or model, the successful impact of these models requires a thorough understanding of the underlying data and key factors, such as the need for estimation of prospective predictivity and their applicability domains. The ever-changing chemical space explored by medicinal chemists in search of new drugs poses a challenge for the models and highlights the importance of regular updates to capture the emerging chemical space to maintain their applicability, as well as the need to include more quality data to populate the chemical space coverage and suitable ML methods to properly leverage the extra data. Computational tools should be reliable, transparent, and applicable to the question at hand to increase their impact on drug discovery. In many cases, in silico tools work best when integrated into the iterative learning cycles of discovery projects, wherein specific tools may need to be developed across various stages of a project. While the acceptability of silico ADME models for driving decisions varies across companies and project teams, these methods have already demonstrated a significant impact on drug discovery for over a decade. As shown in Figure 21.4, the

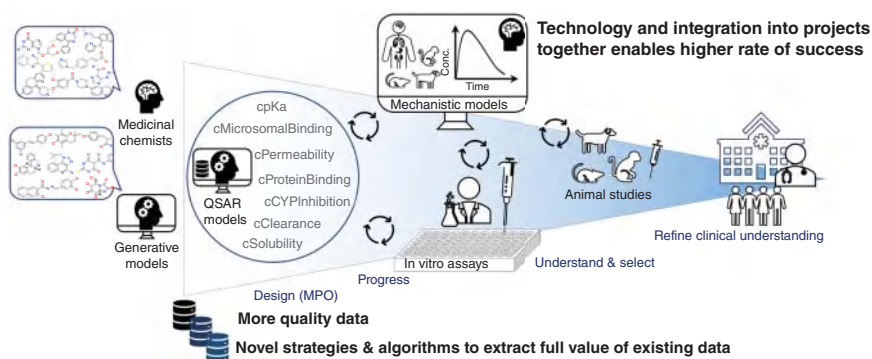


Figure 21.4 Vision and future directions for in silico predictive ADMET.

computational tools for ADMET have been used to assess, prioritize, progress compounds, assist in understanding ADMET mechanisms, and even suggest new candidates. These tools allow the exploration of a large set of (virtual) compounds toward finding novel chemical space and enable holistic and mechanistic optimization for multiple parameters in parallel (MPO), at a much faster rate than the traditional manual process. They hold a great potential to further impact drug discovery, yet it is critical to note that no tool has been able to replace the contributions of human investigators. Most methods aim to complement, augment, and simplify drug discovery research. They enable scientists to extract knowledge from a complex array of historical data and apply the learning to inform future drug discovery programs. Thus, the ultimate success of the in silico ADMET-aided drug discovery relies greatly on the on-going collaborations and trust between laboratories, computational scientists, and end users like medicinal chemists and Drug Metabolism and Pharmacokinetics (DMPK) scientists, making the decision to synthesize and test compounds.

References

- 1 Gao, W. and Coley, C.W. (2020). The synthesizability of molecules proposed by generative models. *J Chem Inf Model* 60: 5714–5723.
- 2 Coley, C.W. (2021). Defining and exploring chemical spaces. *Trends Chem* 3: 133–145.
- 3 Jensen, J.H. (2019). A graph-based genetic algorithm and generative model/Monte Carlo tree search for the exploration of chemical space. *Chem Sci* 10: 3567–3572.
- 4 Griffen, E., Leach, A.G., Robb, G.R., and Warner, D.J. (2011). Matched molecular pairs as a medicinal chemistry tool. *J Med Chem* 54: 7739–7750.
- 5 Awale, M., Riniker, S., and Kramer, C. (2020). Matched molecular series analysis for ADME property prediction. *J Chem Inf Model* 60: 2903–2914.
- 6 Boitreau, J., Mallet, V., Oliver, C., and Waldispühl, J. (2020). OptiMol: optimization of binding affinities in chemical space for drug discovery. *J Chem Inf Model* 60: 5658–5666.

- 7 Wang, Y., Zhan, Y., Liub, C., and Zhan, W. (2022). Application of machine learning technology in the prediction of ADME related pharmacokinetic parameters. *Curr Med Chem* 30 (17): 1945–1962.
- 8 Kearnes, S., Goldman, B., and Pande, V. (2016). Modeling industrial ADMET data with multitask networks. *Arxiv*. <https://doi.org/10.48550/arxiv.1606.08793>.
- 9 Xu, T. et al. (2020). Predictive models for human organ toxicity based on in vitro bioactivity data and chemical structure. *Chem Res Toxicol* 33: 731–741.
- 10 Wang, W., Kim, M.T., Sedykh, A., and Zhu, H. (2015). Developing enhanced blood–brain barrier permeability models: integrating external bio-assay data in QSAR modeling. *Pharm Res* 32: 3055–3065.
- 11 Aliagas, I. et al. (2015). A probabilistic method to report predictions from a human liver microsomes stability QSAR model: a practical tool for drug discovery. *J Comput Aid Mol Des* 29: 327–338.
- 12 Lombardo, F. et al. (2017). In silico absorption, distribution, metabolism, excretion, and pharmacokinetics (ADME-PK): utility and best practices. An industry perspective from the international consortium for innovation through quality in pharmaceutical development. *J Med Chem* 60: 9097–9113.
- 13 Emami, J. (2006). In vitro–in vivo correlation: from theory to applications. *J Pharm Pharm Sci Publ Can Soc Pharm Sci Soc Can Des Sci Pharm* 9: 169–189.
- 14 Caldwell, G.W. (2000). Compound optimization in early- and late-phase drug discovery: acceptable pharmacokinetic properties utilizing combined physico-chemical, in vitro and in vivo screens. *Curr Opin Drug Discov* 3: 30–41.
- 15 Jones, H.M., Gardner, I.B., and Watson, K.J. (2009). Modelling and PBPK simulation in drug discovery. *AAPS J* 11: 155–166.
- 16 Kenny, J.R. (2013). Predictive DMPK: in silico ADME predictions in drug discovery. *Mol Pharm* 10: 1151–1152.
- 17 Parrott, N. and Lave, T. (2008). Applications of physiologically based absorption models in drug discovery and development. *Mol Pharm* 5: 760–775.
- 18 Parrott, N., Manevski, N., and Olivares-Morales, A. (2022). Can we predict clinical pharmacokinetics of highly lipophilic compounds by integration of machine learning or in vitro data into physiologically based models? A feasibility study based on 12 development compounds. *Mol Pharm* 19 (11): 3858–3868. <https://doi.org/10.1021/acs.molpharmaceut.2c00350>.
- 19 Naga, D., Parrott, N., Ecker, G.F., and Olivares-Morales, A. (2022). Evaluation of the success of high-throughput physiologically based pharmacokinetic (HT-PBPK) modeling predictions to inform early drug discovery. *Mol Pharm* 19: 2203–2216.
- 20 Obrezanova, O. et al. (2022). Prediction of in vivo pharmacokinetic parameters and time–exposure curves in rats using machine learning from the chemical structure. *Mol Pharm* 19: 1488–1504.
- 21 Kosugi, Y. and Hosea, N. (2021). Prediction of oral pharmacokinetics using a combination of in silico descriptors and in vitro ADME properties. *Mol Pharm* 18: 1071–1079.
- 22 Kirchmair, J. et al. (2015). Predicting drug metabolism: experiment and/or computation? *Nat Rev Drug Discov* 14: 387–404.

- 23 Bhattachar, S.N., Tan, J.S., and Bender, D.M. (2017). Translating molecules into medicines, cross-functional integration at the drug discovery-development interface. *AAPS Adv Pharm Sci Ser* 25: 231–266. https://doi.org/10.1007/978-3-319-50042-3_7.
- 24 Desai, P.V., Sawada, G.A., Watson, I.A., and Raub, T.J. (2013). Integration of in silico and in vitro tools for scaffold optimization during drug discovery: predicting p-glycoprotein efflux. *Mol Pharm* 10: 1249–1261.
- 25 Dolgikh, E. et al. (2016). QSAR model of unbound brain-to-plasma partition coefficient, $K_{p,uu,brain}$: incorporating p-glycoprotein efflux as a variable. *J Chem Inf Model* 56: 2225–2233.
- 26 Danielson, M.L., Sawada, G.A., Raub, T.J., and Desai, P.V. (2018). In silico and in vitro assessment of OATP1B1 inhibition in drug discovery. *Mol Pharm* 15: 3060–3068.
- 27 Hu, B., Zhou, X., Mohutsky, M.A., and Desai, P.V. (2020). Structure–property relationships and machine learning models for addressing CYP3A4-mediated victim drug–drug interaction risk in drug discovery. *Mol Pharm* 17: 3600–3608.
- 28 Göller, A.H. et al. (2020). Bayer’s in silico ADMET platform: a journey of machine learning over the past two decades. *Drug Discov Today* 25: 1702–1709.
- 29 Sheridan, R.P., Culberson, J.C., Joshi, E. et al. (2022). Prediction accuracy of production ADMET models as a function of version: activity cliffs rule. *J Chem Inf Model* 62: 3275–3280.
- 30 Keefer, C.E., Kauffman, G.W., and Gupta, R.R. (2013). Interpretable, probability-based confidence metric for continuous quantitative structure–activity relationship models. *J Chem Inf Model* 53: 368–383.
- 31 Tsui, V., Ortwine, D.F., and Blaney, J.M. (2017). Enabling drug discovery project decisions with integrated computational chemistry and informatics. *J Comput Aid Mol Des* 31: 287–291.
- 32 Ortwine, D.F. and Aliagas, I. (2013). Physicochemical and DMPK in silico models: facilitating their use by medicinal chemists. *Mol Pharm* 10: 1153–1161.
- 33 Broccatelli, F., Trager, R., Reutlinger, M. et al. (2022). Benchmarking accuracy and generalizability of four graph neural networks using large in vitro ADME datasets from different chemical spaces. *Mol Inform* 41: 2100321.
- 34 Tropsha, A. (2010). Best practices for QSAR model development, validation, and exploitation. *Mol Inform* 29: 476–488.
- 35 Marcou, G. and Varnek, A. (2017). Data curation. In: *Tutorials in chemoinformatics*, 1–36. <https://doi.org/10.1002/9781119161110.ch1>.
- 36 Winiwarter, S. et al. (2015). Time dependent analysis of assay comparability: a novel approach to understand intra- and inter-site variability over time. *J Comput Aid Mol Des* 29: 795–807.
- 37 Stresser, D.M., Mao, J., Kenny, J.R. et al. (2014). Exploring concepts of in vitro time-dependent CYP inhibition assays. *Expert Opin Drug Met* 10: 157–174.
- 38 Mendes, M.D.S. et al. (2020). A laboratory specific scaling factor to predict the in vivo human clearance of aldehyde oxidase substrates. *Drug Metab Dispos* 48, DMD-AR-2020-000082.

- 39 Khojasteh, S.C., Wong, H., Zhang, D., and Hop, C.E.C.A. (2022). Discovery DMPK quick guide. In: *Guide to data interpretation and integration*, 175–215. https://doi.org/10.1007/978-3-031-10691-0_6.
- 40 Johnson, C. et al. (2022). Evaluating confidence in toxicity assessments based on experimental data and in silico predictions. *Comput Toxicol* 21.
- 41 Wenlock, M.C. and Carlsson, L.A. (2015). How experimental errors influence drug metabolism and pharmacokinetic QSAR/QSPR models. *J Chem Inf Model* 55: 125–134.
- 42 Chen, E.C. et al. (2018). Evaluating the utility of canine Mdr1 knockout Madin-Darby canine kidney I cells in permeability screening and efflux substrate determination. *Mol Pharm* 15: 5103–5113.
- 43 Zakharov, A.V., Peach, M.L., Sitzmann, M., and Nicklaus, M.C. (2014). QSAR modeling of imbalanced high-throughput screening data in pubchem. *J Chem Inf Model* 54: 705–712.
- 44 Elkins, R.C. et al. (2013). Variability in high-throughput ion-channel screening data and consequences for cardiac safety assessment. *J Pharmacol Toxicol* 68: 112–122.
- 45 Kalliokoski, T., Kramer, C., Vulpetti, A., and Gedeck, P. (2013). Comparability of mixed IC50 data – a statistical analysis. *PloS One* 8: e61007.
- 46 Sebaugh, J.L. (2011). Guidelines for accurate EC50/IC50 estimation. *Pharm Stat* 10: 128–134.
- 47 Bowes, J. et al. (2012). Reducing safety-related drug attrition: the use of in vitro pharmacological profiling. *Nat Rev Drug Discov* 11: 909–922.
- 48 Melnikov, F., Anger, L.T., and Hasselgren, C. (2022). Toward quantitative models in safety assessment: a case study to show impact of dose–response inference on hERG inhibition models. *Int J Mol Sci* 24: 635.
- 49 López-Massaguer, O. et al. (2017). Generating modeling data from repeat-dose toxicity reports. *Toxicol Sci* 162: 287–300.
- 50 Melnikov, F., Hsieh, J.-H., Sipes, N.S., and Anastas, P.T. (2018). Channel interactions and robust inference for ratiometric β -lactamase assay data: a Tox21 library analysis. *ACS Sustain Chem Eng* 6: 3233–3241.
- 51 Zhang, F., Xue, J., Shao, J., and Jia, L. (2012). Compilation of 222 drugs' plasma protein binding data and guidance for study designs. *Drug Discov Today* 17: 475–485.
- 52 Pellegatti, M., Pagliarusco, S., Solazzo, L., and Colato, D. (2011). Plasma protein binding and blood-free concentrations: which studies are needed to develop a drug? *Expert Opin Drug Met* 7: 1009–1020.
- 53 Hall, L., Hall, L., and Kier, L. (2009). Methods for predicting the affinity of drugs and drug-like compounds for human plasma proteins: a review. *Curr Comput Aid Drug Des* 5: 90–105.
- 54 Toma, C. et al. (2018). QSAR development for plasma protein binding: influence of the ionization state. *Pharm Res* 36: 28.
- 55 Zhu, X.-W., Sedykh, A., Zhu, H. et al. (2013). The use of pseudo-equilibrium constant affords improved QSAR models of human plasma protein binding. *Pharm Res* 30: 1790–1798.

- 56 Danishuddin and Khan, A.U. (2016). Descriptors and their selection methods in QSAR analysis: paradigm for drug design. *Drug Discov Today* 21: 1291–1302.
- 57 Gedeck, P., Rohde, B., and Bartels, C. (2006). QSAR – how good is it in practice? Comparison of descriptor sets on an unbiased cross section of corporate data sets. *J Chem Inf Model* 46: 1924–1936.
- 58 Katritzky, A.R. and Gordeeva, E.V. (1993). Traditional topological indexes vs electronic, geometrical, and combined molecular descriptors in QSAR/QSPR research. *J Chem Inf Comput Sci* 33: 835–857.
- 59 Dudek, A., Arodz, T., and Galvez, J. (2006). Computational methods in developing quantitative structure-activity relationships (QSAR): a review. *Comb Chem High T Scr* 9: 213–228.
- 60 Lo, Y.-C., Rensi, S.E., Torng, W., and Altman, R.B. (2018). Machine learning in chemoinformatics and drug discovery. *Drug Discov Today* 23: 1538–1546.
- 61 Willett, P. (2010). Chemoinformatics and computational chemical biology. *Methods Mol Biol* 672: 133–158.
- 62 Raevsky, O. (2004). Physicochemical descriptors in property-based drug design. *Mini Rev Med Chem* 4: 1041–1052.
- 63 Gozalbes, R., Doucet, J., and Derouin, F. (2002). Application of topological descriptors in QSAR and drug design: history and new trends. *Curr Drug Targets Infect Disord* 2: 93–102.
- 64 Akamatsu, M. (2002). Current state and perspectives of 3D-QSAR. *Curr Top Med Chem* 2: 1381–1394.
- 65 Tropsha, A. and Weifan, Z. (2001). Identification of the descriptor pharmacophores using variable selection QSAR applications to database mining. *Curr Pharm Design* 7: 599–612.
- 66 Karelson, M., Lobanov, V.S., and Katritzky, A.R. (1996). Quantum-chemical descriptors in QSAR/QSPR studies. *Chem Rev* 96: 1027–1044.
- 67 Chemical Computing Group (CCG) | Research. https://www.chemcomp.com/Research-Citing_MOE.htm.
- 68 Tetko, I.V. et al. (2005). Virtual computational chemistry laboratory – design and description. *J Comput Aid Mol Des* 19: 453–463.
- 69 Wang, W. *From QSAR to QNAR, developing enhanced models for drug discovery*. (2020) <https://doi.org/10.7282/t3bz69nc>.
- 70 Cao, D.-S., Xu, Q.-S., Hu, Q.-N., and Liang, Y.-Z. (2013). ChemoPy: freely available python package for computational biology and chemoinformatics. *Bioinformatics* 29: 1092–1094.
- 71 Cao, Y., Charisi, A., Cheng, L.-C. et al. (2008). ChemmineR: a compound mining framework for R. *Bioinformatics* 24: 1733–1734.
- 72 Khan, P.M. and Roy, K. (2018). Current approaches for choosing feature selection and learning algorithms in quantitative structure-activity relationships (QSAR). *Expert Opin Drug Discov* 13: 1075–1089.
- 73 Wang, Y., Yao, H., and Zhao, S. (2016). Auto-encoder based dimensionality reduction. *Neurocomputing* 184: 232–242.
- 74 Hinton, G.E. and Salakhutdinov, R.R. (2006). Reducing the dimensionality of data with neural networks. *Science* 313: 504–507.

- 75 Naqa, I. E. and Murphy, M. J. (2015). Machine learning in radiation oncology, theory and applications. 3–11 https://doi.org/10.1007/978-3-319-18305-3_1.
- 76 Love, B.C. (2002). Comparing supervised and unsupervised category learning. *Psychon B Rev* 9: 829–835.
- 77 Jamal, S., Goyal, S., Grover, A. and Shanker, A. (2018). Bioinformatics: sequences, structures, phylogeny. 359–374 https://doi.org/10.1007/978-981-13-1562-6_16.
- 78 Friedman, N., Geiger, D., and Goldszmidt, M. (1997). Bayesian network classifiers. *Mach Learn* 29: 131–163.
- 79 Kingsford, C. and Salzberg, S.L. (2008). What are decision trees? *Nat Biotechnol* 26: 1011–1013.
- 80 Biau, G. and Scornet, E. (2016). A random forest guided tour. *Test* 25: 197–227.
- 81 Breiman, L. (2001). Random forests. *Mach Learn* 45: 5–32.
- 82 Sheridan, R.P., Wang, W.M., Liaw, A. et al. (2016). Extreme gradient boosting as a method for quantitative structure–activity relationships. *J Chem Inf Model* 56: 2353–2360.
- 83 Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Mach Learn* 63: 3–42.
- 84 Zheng, W. and Tropsha, A. (2000). Novel variable selection quantitative structure–property relationship approach based on the k-nearest-neighbor principle. *J Chem Inf Comput Sci* 40: 185–194.
- 85 Altman, N.S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *Am Stat* 46: 175–185.
- 86 Suthaharan, S. (2016). Machine learning models and algorithms for big data classification, thinking with examples for effective learning. 79–97 https://doi.org/10.1007/978-1-4899-7641-3_4.
- 87 Noble, W.S. (2006). What is a support vector machine? *Nat Biotechnol* 24: 1565–1567.
- 88 Jain, A.K., Mao, J., and Mohiuddin, K.M. (1996). Artificial neural networks: a tutorial. *Computer* 29: 31–44.
- 89 Pedregosa, F. et al. (2012). Scikit-learn: machine learning in python. *Arxiv* <https://doi.org/10.48550/arxiv.1201.0490>.
- 90 CRAN - Package e1071. <https://cran.r-project.org/web/packages/e1071/index.html>.
- 91 Fung, G.M. and Mangasarian, O.L. (2004). A feature selection newton method for support vector machine classification. *Comput Optim Appl* 28: 185–202.
- 92 Khan, N.M., Nalina Madhav, C., Negi, A., and Thaseen, I.S. (2019). Intelligent systems design and applications, 18th international conference on intelligent systems design and applications (ISDA 2018) held in Vellore, India, December 6–8, 2018, volume 2. *Adv Intell Syst.* 69–77. https://doi.org/10.1007/978-3-030-16660-1_7.
- 93 Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinform* 8: 25.

- 94 Pes, B. (2020). Ensemble feature selection for high-dimensional data: a stability analysis across multiple domains. *Neural Comput Appl* 32: 5951–5973.
- 95 Bro, R. and Smilde, A.K. (2014). Principal component analysis. *Anal Methods UK* 6: 2812–2831.
- 96 Abdi, H. and Williams, L.J. (2010). Principal component analysis. *Wiley Interdiscip Rev Comput Stat* 2: 433–459.
- 97 Peluffo-Ordóñez, D.H., Lee, J.A., and Verleysen, M. Advances in self-organizing maps and learning vector quantization, proceedings of the 10th international workshop, WSOM 2014, Mittweida, Germany, July 2–4, 2014. *Adv Intell Syst* (2014): 65–74. https://doi.org/10.1007/978-3-319-07695-9_6.
- 98 Lovrić, M. et al. (2021). Should we embed in chemistry? A comparison of unsupervised transfer learning with PCA, UMAP, and VAE on molecular fingerprints. *Pharmaceuticals (Basel)* 14: 758.
- 99 Smets, T. et al. (2019). Evaluation of distance metrics and spatial autocorrelation in uniform manifold approximation and projection applied to mass spectrometry imaging data. *Anal Chem* 91: 5706–5714.
- 100 Muratov, E.N. et al. (2020). QSAR without borders. *Chem Soc Rev* 49: 3525–3564.
- 101 Tropsha, A. and Golbraikh, A. (2007). Predictive QSAR modeling workflow, model applicability domains, and virtual screening. *Curr Pharm Design* 13: 3494–3504.
- 102 Levatić, J., Ceci, M., Stepišnik, T. et al. (2020). Semi-supervised regression trees with application to QSAR modelling. *Expert Syst Appl* 158: 113569.
- 103 Ouali, Y., Hudelot, C., and Tami, M. (2020). An overview of deep semi-supervised learning. *Arxiv*. <https://doi.org/10.48550/arxiv.2006.05278>.
- 104 Sheridan, R.P. (2013). Time-split cross-validation as a method for estimating the goodness of prospective prediction. *J Chem Inf Model* 53: 783–790.
- 105 Schober, P., Boer, C., and Schwarte, L.A. (2018). Correlation coefficients. *Anesth Analg* 126: 1763–1768.
- 106 Chai, T. and Draxler, R.R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geosci Model Dev* 7: 1247–1250.
- 107 Casanova-Alvarez, O., Morales-Helguera, A., Cabrera-Pérez, M.A. et al. (2021). A novel automated framework for QSAR modeling of highly imbalanced leishmania high-throughput screening data. *J Chem Inf Model* 61: 3213–3231.
- 108 Sedykh, A. et al. (2011). Use of in vitro HTS-derived concentration–response data as biological descriptors improves the accuracy of QSAR models of in vivo toxicity. *Environ Health Perspect* 119: 364–370.
- 109 Zhu, H., Rusyn, I., Richard, A., and Tropsha, A. (2008). Use of cell viability assay data improves the prediction accuracy of conventional quantitative structure–activity relationship models of animal carcinogenicity. *Environ Health Perspect* 116: 506–513.
- 110 Sheridan, R.P. (2012). Three useful dimensions for domain applicability in QSAR models using random forest. *J Chem Inf Model* 52: 814–823.

- 111 Kar, S., Roy, K., and Leszczynski, J. (2018). Computational toxicology, methods and protocols. *Methods Mol Biol* 1800: 141–169.
- 112 Sheridan, R.P. (2013). Using random forest to model the domain applicability of another random forest model. *J Chem Inf Model* 53: 2837–2850.
- 113 Norinder, U., Carlsson, L., Boyer, S., and Eklund, M. (2014). Introducing conformal prediction in predictive modeling. A transparent and flexible alternative to applicability domain determination. *J Chem Inf Model* 54: 1596–1603.
- 114 Breiman, L. (1996). Stacked regressions. *Mach Learn* 24: 49–64.
- 115 Freund, Y. and Schapire, R.E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *J Comput Syst Sci* 55: 119–139.
- 116 Kwon, S., Bae, H., Jo, J., and Yoon, S. (2019). Comprehensive ensemble in QSAR prediction for drug discovery. *BMC Bioinform* 20: 521.
- 117 Cortés-Ciriano, I. and Bender, A. (2019). Concepts and applications of conformal prediction in computational drug discovery. *Arxiv*. <https://doi.org/10.48550/arxiv.1908.03569>.
- 118 Alanine, A., Nettekoven, M., Roberts, E., and Thomas, A. (2003). Lead generation-enhancing the success of drug discovery by investing in the hit to lead process. *Comb Chem High T Scr* 6: 51–66.
- 119 Hughes, J., Rees, S., Kalindjian, S., and Philpott, K. (2011). Principles of early drug discovery. *Br J Pharmacol* 162: 1239–1249.
- 120 Broccatelli, F., Aliagas, I., and Zheng, H. (2018). Why decreasing lipophilicity alone is often not a reliable strategy for extending IV half-life. *ACS Med Chem Lett* 9: 522–527.
- 121 Segall, M.D. (2012). Multi-parameter optimization: identifying high quality compounds with a balance of properties. *Curr Pharm Design* 18: 1292–1310.
- 122 Pennington, L.D. and Muegge, I. (2021). Holistic drug design for multiparameter optimization in modern small molecule drug discovery. *Bioorg Med Chem Lett* 41: 128003.
- 123 Wager, T.T., Hou, X., Verhoest, P.R., and Villalobos, A. (2010). Moving beyond rules: the development of a central nervous system multiparameter optimization (CNS MPO) approach to enable alignment of druglike properties. *ACS Chem Neurosci* 1: 435–449.
- 124 Ferreira, L.L.G., de Moraes, J., and Andricopulo, A.D. (2022). Approaches to advance drug discovery for neglected tropical diseases. *Drug Discov Today* 27: 2278–2287.
- 125 Przybylak, K.R. and Cronin, M.T.D. (2012). In silico models for drug-induced liver injury – current status. *Expert Opin Drug Metab Toxicol* 8: 201–217.
- 126 Chen, M. et al. (2014). Toward predictive models for drug-induced liver injury in humans: are we there yet? *Biomark Med* 8: 201–213.
- 127 Bassan, A. et al. (2021). In silico approaches in organ toxicity hazard assessment: current status and future needs for predicting heart, kidney and lung toxicities. *Comput Toxicol* 20.
- 128 Siramshetty, V.B. et al. (2020). Critical assessment of artificial intelligence methods for prediction of hERG channel inhibition in the “big data” era. *J Chem Inf Model* 60: 6007–6019.

- 129 Martin, M.T. et al. (2022). Early drug-induced liver injury risk screening: “free,” as good as it gets. *Toxicol Sci* 188: 208–218.
- 130 Garrido, A., Lepailleur, A., Mignani, S.M. et al. (2020). hERG toxicity assessment: useful guidelines for drug design. *Eur J Med Chem* 195: 112290.
- 131 Moeller, T.A., Shukla, S.J., and Xia, M. (2012). Assessment of compound hepatotoxicity using human plateable cryopreserved hepatocytes in a 1536-well-plate format. *Assay Drug Dev Technol* 10: 78–87.
- 132 Proctor, W.R. et al. (2017). Utility of spherical human liver microtissues for prediction of clinical drug-induced liver injury. *Arch Toxicol* 91: 2849–2863.
- 133 Espinosa, J.A., Pohan, G., Arkin, M.R., and Markossian, S. (2021). Real-time assessment of mitochondrial toxicity in HepG2 cells using the seahorse extracellular flux analyzer. *Curr Protoc* 1: e75.
- 134 Miller, B. et al. (1998). Evaluation of the in vitro micronucleus test as an alternative to the in vitro chromosomal aberration assay: position of the GUM working group on the in vitro micronucleus test. *Mutat Res Rev Mutat Res* 410: 81–116.
- 135 Hasselgren, C. and Myatt, G.J. (2018). Computational toxicology, methods and protocols. *Methods Mol Biol* 1800: 233–244.
- 136 Judson, P. (2010). Using computer reasoning about qualitative and quantitative information to predict metabolism and toxicity. In: *Pharmacokinetic profiling in drug research: biological, physicochemical, and computational strategies*, 417–429. <https://doi.org/10.1002/9783906390468.ch24>.
- 137 Greene, N., Judson, P.N., Langowski, J.J., and Marchant, C.A. (1999). Knowledge-based expert systems for toxicity and metabolism prediction: DEREK, StAR and METEOR. *SAR QSAR Environ Res* 10: 299–314.
- 138 ToxTree version 2.6.6. (2015).
- 139 Chakravarti, S.K., Saiakhov, R.D., and Klopman, G. (2012). Optimizing predictive performance of CASE ultra expert system models using the applicability domains of individual toxicity alerts. *J Chem Inf Model* 52: 2609–2618.
- 140 Saiakhov, R., Chakravarti, S., and Klopman, G. (2013). Effectiveness of CASE ultra expert system in evaluating adverse effects of drugs. *Mol Inform* 32: 87–97.
- 141 Leadscape Expert Alerts version 3.2.4-1. http://www.leadscope.com/expert_alerts (2015).
- 142 EMA (2015). ICH guideline M7 (R1) on assessment and control of DNA reactive (mutagenic) impurities in pharmaceuticals to limit potential carcinogenic risk. https://www.ema.europa.eu/en/documents/scientific-guideline/ich-guideline-m7r1-assessment-control-dna-reactive-mutagenic-impurities-pharmaceuticals-limit_en.pdf.
- 143 Sutter, A. et al. (2013). Use of in silico systems and expert knowledge for structure-based assessment of potentially mutagenic impurities. *Regul Toxicol Pharmacol* 67: 39–52.
- 144 Brigo, A. and Muster, W. (2016). In silico methods for predicting drug toxicity. *Methods Mol Biol* 1425: 475–510.
- 145 Schmidt, F., Matter, H., Hessler, G., and Czich, A. (2014). Predictive in silico off-target profiling in drug discovery. *Future Med Chem* 6: 295–317.

- 146 Brown, A.M. (2004). Drugs, hERG and sudden death. *Cell Calcium* 35: 543–547.
- 147 Hasselgren, C. et al. (2013). Chemoinformatics and beyond. In: *Chemoinformatics for drug discovery*, 267–290. <https://doi.org/10.1002/9781118742785.ch12>.
- 148 Krishnapuram, B. et al. (2016). XGBoost. *Proc 22nd Acm SIGKDD Int Conf Knowl Discov Data Min* 785–794. <https://doi.org/10.1145/2939672.2939785>.
- 149 Aronov, A.M. (2006). Common pharmacophores for uncharged human ether-a-go-go-related gene (hERG) blockers. *J Med Chem* 49: 6917–6921.
- 150 Sameshima, T. et al. (2020). Small-scale panel comprising diverse gene family targets to evaluate compound promiscuity. *Chem Res Toxicol* 33: 154–161.
- 151 Ma, J., Sheridan, R.P., Liaw, A. et al. (2015). Deep neural nets as a method for quantitative structure–activity relationships. *J Chem Inf Model* 55: 263–274.
- 152 Chen, B., Sheridan, R.P., Hornak, V., and Voigt, J.H. (2012). Comparison of random forest and pipeline pilot naïve bayes in prospective QSAR predictions. *J Chem Inf Model* 52: 792–803.
- 153 Feinberg, E.N., Joshi, E., Pande, V.S., and Cheng, A.C. (2020). Improvement in ADMET prediction with multitask deep featurization. *J Med Chem* 63: 8835–8848.
- 154 Cáceres, E.L., Tudor, M., and Cheng, A.C. (2020). Deep learning approaches in predicting ADMET properties. *Future Med Chem* 12: 1995–1999.
- 155 Venkatraman, V. (2021). FP-ADMET: a compendium of fingerprint-based ADMET prediction models. *J Chem* 13: 75.
- 156 Montanari, F., Kuhnke, L., Laak, A.T., and Clevert, D.-A. (2019). Modeling physico-chemical ADMET endpoints with multitask graph convolutional networks. *Molecules* 25: 44.
- 157 Zhou, Y. et al. (2019). Exploring tunable hyperparameters for deep neural networks with industrial ADME data sets. *J Chem Inf Model* 59: 1005–1016.
- 158 Alexander Heifetz (2022). *Artificial intelligence in drug design*. Methods in molecular biology. <https://doi.org/10.1007/978-1-0716-1787-8>.
- 159 Klambauer, G., Hochreiter, S., and Rarey, M. (2019). Machine learning in drug discovery. *J Chem Inf Model* 59: 945–946.
- 160 Bhatarai, B., Walters, W.P., Hop, C.E.C.A. et al. (2019). Opportunities and challenges using artificial intelligence in ADME/tox. *Nat Mater* 18: 418–422.
- 161 Wenzel, J., Matter, H., and Schmidt, F. (2019). Predictive multitask deep neural network models for ADME-tox properties: learning from large data sets. *J Chem Inf Model* 59: 1253–1268.
- 162 Abadi, M. et al. (2015). *TensorFlow: large-scale machine learning on heterogeneous distributed systems*. <https://doi.org/10.5281/zenodo.4724125>.
- 163 Wang, M. et al. (2019). Deep graph library: a graph-centric, highly-performant package for graph neural networks. *Arxiv* <https://doi.org/10.48550/arxiv.1909.01315>.
- 164 Paszke, A. et al. (2019). PyTorch: an imperative style, high-performance deep learning library. *Arxiv*. <https://doi.org/10.48550/arxiv.1912.01703>.
- 165 Polishchuk, P.G., Madzhidov, T.I., and Varnek, A. (2013). Estimation of the size of drug-like chemical space based on GDB-17 data. *J Comput Aid Mol Des* 27: 675–679.

- 166 Hearst, M.A., Dumais, S.T., Osuna, E. et al. (1998). Support vector machines. *IEEE Intell Syst Appl* 13: 18–28.
- 167 Hastie, T., Tibshirani, R., and Friedman, J. (2009). The elements of statistical learning, data mining, inference, and prediction. *Springer Ser Stat* <https://doi.org/10.1007/978-0-387-84858-7>.
- 168 Krenn, M., Hse, F., Nigam, A. et al. (2020). Self-referencing embedded strings (SELFIES): a 100% robust molecular string representation. *Mach Learn Sci Technol* 1: 045024.
- 169 Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Model* 28: 31–36.
- 170 Keogh, E. and Mueen, A. (2017). *Encyclopedia of machine learning and data mining*, 314–315. https://doi.org/10.1007/978-1-4899-7687-1_192.
- 171 Bellman, R.E. (2010). *Dynamic programming*. Princeton University Press.
- 172 Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT Press.
- 173 Kearnes, S., McCloskey, K., Berndl, M. et al. (2016). Molecular graph convolutions: moving beyond fingerprints. *J Comput Aid Mol Des* 30: 595–608.
- 174 Baydin, A.G., Pearlmutter, B.A., Radul, A.A., and Siskind, J.M. (2018). Automatic differentiation in machine learning: a survey. *J Mach Learn Res* 1–43. <https://doi.org/10.48550/arxiv.1502.05767>.
- 175 Broccatelli, F. et al. (2016). Predicting passive permeability of drug-like molecules from chemical structure: where are we? *Mol Pharm* 13: 4199–4208.
- 176 Caruana, R. (1997). Multitask learning. *Mach Learn* 28: 41–75.
- 177 Sosnin, S. et al. (2019). A survey of multi-task learning methods in chemoinformatics. *Mol Inform* 38: 1800108.
- 178 Rohall, S.L. et al. (2020). An artificial intelligence approach to proactively inspire drug discovery with recommendations. *J Med Chem* 63: 8824–8834.
- 179 Hussain, J. and Rea, C. (2010). Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. *J Chem Inf Model* 50: 339–348.
- 180 Dalke, A., Hert, J., and Kramer, C. (2018). Mmpdb: an open-source matched molecular pair platform for large multiproperty data sets. *J Chem Inf Model* 58: 902–910.
- 181 Landry, M.L. and Crawford, J.J. (2020). LogD contributions of substituents commonly used in medicinal chemistry. *ACS Med Chem Lett* 11: 72–76.
- 182 Ritchie, T.J., Macdonald, S.J.F., and Pickett, S.D. (2015). Insights into the impact of N- and O-methylation on aqueous solubility and lipophilicity using matched molecular pair analysis. *MedChemComm* 6: 1787–1797.
- 183 Kramer, C. et al. (2018). Learning medicinal chemistry absorption, distribution, metabolism, excretion, and toxicity (ADMET) rules from cross-company matched molecular pairs analysis (MMPA). *J Med Chem* 61: 3277–3292.
- 184 Landry, M.L., Trager, R., Broccatelli, F., and Crawford, J.J. (2022). When cofactors aren't X factors: functional groups that are labile in human liver microsomes in the absence of NADPH. *ACS Med Chem Lett* 13: 727–733.
- 185 Stepan, A.F., Kauffman, G.W., Keefer, C.E. et al. (2013). Evaluating the differences in cycloalkyl ether metabolism using the design parameter “lipophilic

- metabolism efficiency” (LipMetE) and a matched molecular pairs analysis. *J Med Chem* 56: 6985–6990.
- 186 Leach, A.G. et al. (2006). Matched molecular pairs as a guide in the optimization of pharmaceutical properties; a study of aqueous solubility, plasma protein binding and oral exposure. *J Med Chem* 49: 6672–6682.
 - 187 Gleeson, P., Bravi, G., Modi, S., and Lowe, D. (2009). ADMET rules of thumb II: a comparison of the effects of common substituents on a range of ADMET parameters. *Bioorg Med Chem* 17: 5906–5919.
 - 188 Meyers, J., Fabian, B., and Brown, N. (2021). De novo molecular design and generative models. *Drug Discov Today* 26: 2707–2715.
 - 189 Bickerton, G.R., Paolini, G.V., Besnard, J. et al. (2012). Quantifying the chemical beauty of drugs. *Nat Chem* 4: 90–98.
 - 190 Firth, N.C., Atrash, B., Brown, N., and Blagg, J. (2015). MOARF, an integrated workflow for multiobjective optimization: implementation, synthesis, and biological evaluation. *J Chem Inf Model* 55: 1169–1180.
 - 191 Polykovskiy, D. et al. (2020). Molecular sets (MOSES): a benchmarking platform for molecular generation models. *Front Pharmacol* 11: 565644.
 - 192 Chen, H. and Engkvist, O. (2019). Has drug design augmented by artificial intelligence become a reality? *Trends Pharmacol Sci* 40: 806–809.
 - 193 Zhavoronkov, A. and Aspuru-Guzik, A. (2020). Reply to ‘Assessing the impact of generative AI on medicinal chemistry’. *Nat Biotechnol* 38: 146–146.
 - 194 Walters, W.P. and Murcko, M. (2020). Assessing the impact of generative AI on medicinal chemistry. *Nat Biotechnol* 38: 143–145.
 - 195 Vamathevan, J. et al. (2019). Applications of machine learning in drug discovery and development. *Nat Rev Drug Discov* 18: 463–477.
 - 196 Merk, D., Friedrich, L., Grisoni, F., and Schneider, G. (2018). De novo design of bioactive small molecules by artificial intelligence. *Mol Inform* 37: 1700153.
 - 197 Olivecrona, M., Blaschke, T., Engkvist, O., and Chen, H. (2017). Molecular de-novo design through deep reinforcement learning. *J Chem* 9: 48.
 - 198 Muegge, I. (2003). Selection criteria for drug-like compounds. *Med Res Rev* 23: 302–321.
 - 199 Lipinski, C.A., Lombardo, F., Dominy, B.W., and Feeney, P.J. (2001). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev* 39: 237–257. The article was originally published in *Advanced Drug Delivery Reviews* 23 (1997) 3–25. *Adv Drug Deliv Rev* 46: 3–26.
 - 200 Segall, M.D., Beresford, A.P., Gola, J.M. et al. (2006). Focus on success: using a probabilistic approach to achieve an optimal balance of compound properties in drug discovery. *Expert Opin Drug Met* 2: 325–337.
 - 201 Chen, E.P., Bondi, R.W., and Michalski, P.J. (2021). Model-based target pharmacology assessment (mTPA): an approach using PBPK/PD modeling and machine learning to design medicinal chemistry and DMPK strategies in early drug discovery. *J Med Chem* 64: 3185–3196.
 - 202 Chen, E.P. et al. (2022). Applications of model-based target pharmacology assessment in defining drug design and DMPK strategies: GSK experiences. *J Med Chem* 65: 6926–6939.

- 203 He, C. and Wan, H. (2018). Drug metabolism and metabolite safety assessment in drug discovery and development. *Expert Opin Drug Met* 14: 1071–1085.
- 204 Smith, A.M.E., Lanevskij, K., Sazonovas, A., and Harris, J. (2022). Impact of established and emerging software tools on the metabolite identification landscape. *Front Toxicol* 4: 932445.
- 205 Manikandan, P. and Nagini, S. (2018). Cytochrome P450 structure, function and clinical significance: a review. *Curr Drug Targets* 19: 38–54.
- 206 Li, J., Schneebeli, S.T., Bylund, J. et al. (2011). IDSite: an accurate approach to predict P450-mediated drug metabolism. *J Chem Theory Comput* 7: 3829–3845.
- 207 Öeren, M. et al. (2022). Predicting regioselectivity of AO, CYP, FMO, and UGT metabolism using quantum mechanical simulations and machine learning. *J Med Chem* 65: 14066–14081.
- 208 Moors, S.L.C., Vos, A.M., Cummings, M.D. et al. (2011). Structure-based site of metabolism prediction for cytochrome P450 2D6. *J Med Chem* 54: 6098–6105.
- 209 Tarcsay, Á., Kiss, R., and Keserű, G.M. (2010). Site of metabolism prediction on cytochrome P450 2C9: a knowledge-based docking approach. *J Comput Aid Mol Des* 24: 399–408.
- 210 Vasanthanathan, P. et al. (2009). Virtual screening and prediction of site of metabolism for cytochrome P450 1A2 ligands. *J Chem Inf Model* 49: 43–52.
- 211 Hughes, T.B., Miller, G.P., and Swamidass, S.J. (2015). Site of reactivity models predict molecular reactivity of diverse chemicals with glutathione. *Chem Res Toxicol* 28: 797–809.
- 212 Kirchmair, J. et al. (2013). FASt METabolizer (FAME): a rapid and accurate predictor of sites of metabolism in multiple species by endogenous enzymes. *J Chem Inf Model* 53: 2896–2907.
- 213 Peng, J. et al. (2014). In silico site of metabolism prediction for human UGT-catalyzed reactions. *Bioinformatics* 30: 398–405.
- 214 Smith, P.A., Sorich, M.J., Low, L.S.C. et al. (2004). Towards integrated ADME prediction: past, present and future directions for modelling metabolism by UDP-glucuronosyltransferases. *J Mol Graph Model* 22: 507–517.
- 215 Tyzack, J.D., Hunt, P.A., and Segall, M.D. (2016). Predicting regioselectivity and lability of cytochrome P450 metabolism using quantum mechanical simulations. *J Chem Inf Model* 56: 2180–2193.
- 216 Zaretski, J. et al. (2012). RS-predictor models augmented with SMARTCyp reactivities: robust metabolic regioselectivity predictions for nine CYP isozymes. *J Chem Inf Model* 52: 1637–1659.
- 217 Cruciani, G. et al. (2005). MetaSite: understanding metabolism in human cytochromes from the perspective of the chemist. *J Med Chem* 48: 6970–6979.

Part VII

Computational Approaches for New Therapeutic Modalities

22

Modeling the Structures of Ternary Complexes Mediated by Molecular Glues

Michael L. Drummond

Chemical Computing Group, Montreal, Quebec H3A 2R7, Canada

22.1 Introduction

The term “molecular glue” (abbreviated hereafter as MG) was first coined in 1992 by Stuart Schreiber [1] to describe how the macrocyclic natural products cyclosporin A, rapamycin, and FK506 induce the formation of ternary complexes (i.e. complexes made from three components). Specifically, these MGs exhibit an immunosuppressant effect by first binding to their endogenous receptors (the so-called immunophilins, such as cyclophilin and FKBP); these binary complexes then engage with a second target protein, such as calcineurin or FRAP. Importantly, it was noted that calcineurin does not bind with appreciable affinity to either the free small molecules or the immunophilins, and thus these natural products “glue” the two proteins together – a finding later elucidated via crystallography [2]. The calcineurin-cyclosporin A-cyclophilin ternary complex was later invoked as precedent in 2014 [3] to characterize the structural interaction between the infamous small molecule thalidomide (and derivatives), its receptor protein cereblon (CRBN), and the (as of 2014) undetermined target protein. Today, thalidomide and its analogs and derivatives, collectively known as IMiDs (immunomodulatory drugs), are by far the most important class of MGs [4], particularly among compounds that have advanced to or emerged from the clinic (Figure 22.1) [5]. Targeted protein degradation using E3 ligases besides CRBN, such as DCAF15 [6], has also been effected, utilizing different scaffolds.

Soon after the publication of the first crystal structures of these IMiD-CRBN complexes [3], two independent efforts incorporated thalidomide [7] or pomalidomide [8] as the CRBN-recruiting moieties using the PROTAC approach. First constructed as polypeptidic moieties [9], but later (and almost exclusively) as small molecules [10], PROteolysis-TArgeting Chimeras are a class of protein–protein proximity inducers similar to, but conceptually distinct from, MGs [11] – with an even greater presence in the clinic [12]. PROTACs are, by definition, bifunctional molecules, where the two “ends” of the molecules, connected by a linker, are each responsible for binding to a distinct protein. Although in reality, this demarcation is

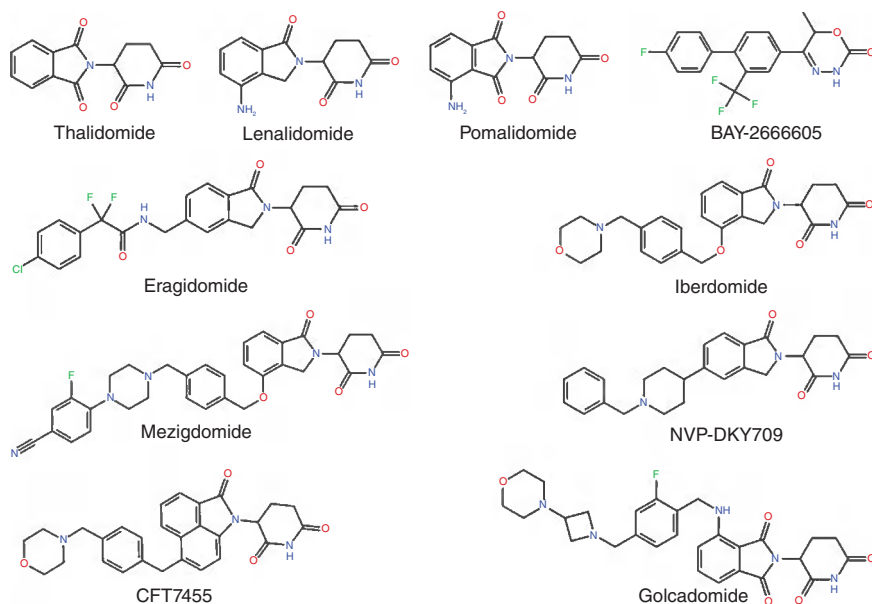


Figure 22.1 Some of the molecular glues that are either in clinical trials or already on the market.

not quite so strict – interactions between both proteins and both ends of a PROTAC can be observed in crystal structures [13] – this conceptual framework nonetheless lends itself to a modular design approach for PROTACs, so that dozens of proteins can be degraded via a PROTAC approach by adjusting one binding end, along with concomitant optimization of the linker (often dubbed “linkerology”). [14, 15] By contrast, MGs are typically considered as monovalent molecules, where both proteins simultaneously interact with a single moiety – although larger MGs such as mezigdomide (Figure 22.1) should perhaps be viewed as only nominally monovalent. Regardless, as a class MGs are generally smaller and more “drug-like” than PROTACs, which is largely responsible for intense interest in their development as potential therapeutics. However, the colocation of two separate protein-binding moieties into a single molecular entity greatly complicates the rational design of MGs. Indeed, to date, the initial discovery of most MGs has been driven by “serendipity,” although the rational design has been applied to refine these initial, serendipitous molecules [16].

We have previously [17, 18] utilized the inherent modularity of PROTACs to construct computational methods for modeling the structures of PROTAC-mediated ternary complexes. In our most successful modeling approach, Method 4B, three inputs are required: the PROTACs themselves, as well as two binary protein–ligand complexes, where the respective protein pockets contain warheads that (largely) match the binding ends of the PROTACs. Here, we describe the extension of these PROTAC modeling techniques to predict the structure of ternary complexes mediated by MGs.

Two distinct approaches will be detailed. The first approach treats MGs as they are commonly conceptualized – as whole, indivisible molecules, placed via small molecule docking at protein–protein interfaces (PPIs), which are themselves predicted by protein–protein docking. The second approach instead treats MGs as “linkerless PROTACs,” i.e. as molecules that, despite their nominal monovalency, can be partitioned into two binding parts (*cf.* three parts for a PROTAC: binder-linker-binder), each of which can be viewed as primarily interacting with just one of the proteins in the ternary complex. Thus, after MG partitioning, the computational protocol is analogous to Method 4B [18]. As will be shown, although this second approach requires additional information about the system when compared to the first approach, the accuracy of the predicted ternary complex models is improved when MGs are treated via this PROTAC-like approach. A unified interface has been developed for both approaches, where the user can decide whether or not to partition their MGs (and thus whether the first or second approach should be utilized). Moreover, although only the MGs and the structures of the proteins are required as minimal input, additional information describing the nature of the protein–protein and/or protein-MG interfaces can optionally be specified to guide the simulation. To the best of our knowledge, the tools described herein are the first computational methods specifically designed to model MG-mediated ternary complexes.

22.2 Methodology

All results described herein were produced using the MOE software package [19]. The computational protocols described below were implemented in SVL (Scientific Vector Language), MOE’s integrated programming language, and are freely available upon request to anyone with access to MOE. In order to judge the accuracy of the MG-mediated ternary complex structures predicted with the two computational approaches described in this work, a set of 32 known MG-containing ternary complex crystal structures was assembled (Table 22.1). This validation set contains many of the crystal structures collated in a recent MG review [11], augmented with newer crystal structures featuring the E3 ligases CRBN [20–25] and DCAF15 [6, 26, 27], as well as a set of rationally designed MGs targeting the protein 14-3-3 [28]. In constructing this dataset, an expansive definition of “molecular glue” has been adopted – to wit, in some of these complexes, the two proteins may have appreciable interactions even in the absence of the accompanying MG. Regardless, a small molecule of MG can be found “sandwiched” between two proteins in all complexes in Table 22.1.

The accuracy of the predicted MG-mediated ternary complex structures is evaluated in this work using the same metric previously used to judge the accuracy of predicted PROTAC-mediated ternary complexes [17, 18]. That is, a successful prediction must have $<10 \text{ \AA}$ RMSD between the alpha carbons of the protein chain that move during protein–protein docking and their crystallographic positions, after

Table 22.1 The 32 molecular glue-containing ternary complex crystal structures used in this validation study.

PDB	Protein 1	Protein 2	Molecular glue	PDB	Protein 1	Protein 2	Molecular glue
1FAP	FKBP12	FRAP	Rapamycin	6Q0W	DCAF15	RBM39	Indisulam
2P1N	TIR1	IAA7	Auxin-like	6RHC	14-3-3	TAZpS89	AZ-003
2P1O	TIR1	IAA7	Auxin-like	6RJL	14-3-3	TAZpS89	AZ-018
2P1Q	TIR1	IAA7	Auxin	6RJQ	14-3-3	TAZpS89	AZ-006
3M50	14-3-3	PMA2	Epibestatin	6RKK	14-3-3	p53pT387	AZ-021
3M51	14-3-3	PMA2	Pyrrolidone 1	6RP6	14-3-3	TAZpS89	AZ-019
3OGL	COI1	JAZ	JA-Ile	6RX2	14-3-3	p53pT387	AZ-005
3OGM	COI1	JAZ	Coronatine	6SJ7	DCAF15	RBM39	Indisulam
4IHL	14-3-3	CRAFpep	Cotylenin A	6SLW	14-3-3	TAZpS89	AZ-004
5FQD	CRBN	CK1alpha	Lenalidomide	6SLX	14-3-3	TAZpS89	AZ-010
5HXB	CRBN	GSPT1	CC-885	6UD7	DCAF15	RBM39	Indisulam
6H0F	CRBN	ZF2	Pomalidomide	6UE5	DCAF15	RBM39	Sulfonamide
6H0G	CRBN	ZF4	Pomalidomide	6UML	CRBN	SALL4	Pomalidomide
6PAI	DCAF15	RBM39	E7820	6XK9	CRBN	GSPT1	CC-90009
6Q0R	DCAF15	RBM39	E7820	7BQU	CRBN	SALL4	Thalidomide
6Q0V	DCAF15	RBM39	Tasisulam	7BQV	CRBN	SALL4	5-OH-thalidomide

rigid body superposition of the accompanying stationary protein onto its crystallographic position. Additionally, an analogous metric has been used to judge the successful placement of the MGs: after the protein-based superposition just described, the heavy atoms of the predicted MG must be within 10 Å RMSD of its crystallographic coordinates to be deemed successful. Most analyses in this work describe performance against the entire 32-member validation set, but occasionally special attention will be paid to the eight CRBN-containing ternary complexes in Table 22.1, due to the obvious therapeutic interest in this protein.

Finally, it should be noted that ternary complex structures were predicted using the component protein chains as found in the PDB structures of Table 22.1. That is, the structure of 14-3-3 used to predict the ternary complex described by 3M50 was the A chain of 3M50 itself (and similarly, the P chain was used for the PMA2 protein in 3M50). Originally, separate (and ideally *apo*) crystal structures were sought for use as inputs, but the quality (and availability) of these separate crystal structures was deemed to be too variable. Although protein sidechains were always repacked during the protein–protein docking simulations of this work, backbone conformations do not deviate from their crystallographic geometries. Thus, the results in this work are something of a “best case” scenario, and performance should be expected to decrease if ternary complex formation is accompanied by substantial protein conformational deformation, as has recently been demonstrated to exist for CRBN [29].

22.3 Results and Discussion

22.3.1 Approach 1: Treating MGs as Whole, Indivisible Molecules

Procedure. The first approach to model MG-mediated ternary complexes, which is the default setting, is most appropriate when limited knowledge is available about the system; when additional information is available, the second approach (detailed below) generally produces more accurate results. Approach 1 only requires the structures of the two proteins that are to be glued together, in addition to the putative MGs. Typically, the protein structures come from structures solved using X-ray crystallography or cryo-EM, although a high-quality homology model, such as an AlphaFold structure [30], could also be used; the impact on accuracy of using a homology model in lieu of an experimentally determined structure has not yet been explicitly determined.

In Approach 1, the protein-MG-protein ternary complex is assembled first by predicting potential PPIs between the two apo proteins using MOE's protein-protein docking algorithm [19]. Up to 100 potential PPI poses are generated – occasionally slightly fewer may result, if the routine produces duplicate poses – and are sorted by the default protein-protein docking score, which is the forcefield interaction energy. Each of these protein-protein poses is then analyzed with MOE's Site Finder tool, to identify potential pockets and cavities at the predicted PPIs. The predicted pockets are ranked by the default score [31], based on their resemblance to small molecule binding sites found in extant crystal structures. The highest scoring pocket that is located at the PPI, i.e. within 4.5 Å of both proteins, is designated as the site the user-supplied MGs are docked into, using MOE's default, non-covalent Dock algorithm. The MG docked pose that scores best using the default GBVI/wSA dG score [32] is chosen as the final position of that MG with that particular PPI, and is written to the final output database. This procedure is repeated for each of the predicted PPIs, so that 100 potential ternary complexes are generated for each MG.

Even though Approach 1 is well-suited for situations where minimal knowledge is on-hand, it is possible to add constraints to guide the PPIs generated during the protein-protein docking procedure. Specifically, it is possible to add no constraints (the default setting), a site constraint to one protein, to the “other” protein, or to both proteins simultaneously. These various combinations of user-provided information will be referred to as Scenarios in this work and are summarized in Table 22.2, along with additional Scenarios only relevant for Approach 2 (see below).

Results. The results using Approach 1 on the 32 ternary complexes of Table 22.1 using Scenarios 1–4 of Table 22.2 are summarized in Table 22.3. With 100 PPIs predicted for each ternary complex – each of which contains one best-scoring docked pose of the corresponding MG – all told, 3200 ternary complexes were predicted using each Scenario. The “All Predictions” section of Table 22.3 provides a broad overview of the ability of Approach 1 to produce ternary complexes that resemble the experimental structures, in regards to reproducing both the protein and MG coordinates. As expected, due to the smaller size of the MGs compared to the proteins, it is easier to predict the placement of the MGs to within 10 Å of their crystallographic

Table 22.2 The nine Scenarios governing how much user-specified information is provided to the validation simulations in this work.

Scenario ^{a)}	Site1 ^{b), c)}	Site2 ^{b), c)}	Pose1 ^{c), d)}	Pose2 ^{c), d)}
1	Unknown	Unknown	Unknown	Unknown
2	Pocket	Unknown	Unknown	Unknown
3	Unknown	Pocket	Unknown	Unknown
4	Pocket	Pocket	Unknown	Unknown
5	Pocket	Unknown	Crystal	Unknown
6	Unknown	Pocket	Unknown	Crystal
7	Pocket	Pocket	Crystal	Unknown
8	Pocket	Pocket	Unknown	Crystal
9	Pocket	Pocket	Crystal	Crystal

- a) Scenarios 1–4 were explored in conjunction with Approach 1, and Scenarios 4–9 with Approach 2. Whenever two Scenarios have the same number of constraints added (e.g. Scenarios 2 and 3), the larger protein carries the constraint in the lower number Scenario, and the smaller protein in the higher number Scenario.
- b) “Pocket” indicates that the protein residues within 4.5 Å of the crystallographically positioned ligand were used to define the corresponding site constraint.
- c) “Unknown” indicates that no experimental information was provided to the simulation.
- d) “Crystal” indicates that the crystallographic coordinates of the molecular glue were provided to the simulation.

positions than it is to correctly predict the protein positions. Thus, in Scenario 4, only 330 out of the 3200 total predicted ternary complexes (10.3%) correctly reproduce the protein positions, whereas 828/3200 (25.9%) of these complexes have the MG near its crystallographic position. Additionally, across all predictions, there is a clear benefit to providing additional information to guide the protein–protein docking part of Approach 1. Adding a docking constraint to the larger protein (as in Scenarios 2 and 4) seems especially worthwhile and yields improved performance for correctly predicting both the protein and MG geometries; adding a constraint to just the smaller protein (Scenario 3) shows only a modest improvement for predicting the PPIs and has little – or perhaps even a slightly deleterious – effect on correctly placing the MGs.

Considering the predictions on a Per System basis (Table 22.3) reveals that, for nearly all of the 32 ternary complexes (Table 22.1), at least one pose out of the 100 predicted in total has the MG correctly placed, with only a slightly lower rate of success when judging the PPIs. This finding is true regardless of whether additional information was provided to guide the protein–protein docking procedure. However, unfortunately, the few ternary complexes that never have a prediction <10 Å are almost entirely complexes containing CRBN – perhaps a consequence of the conformational flexibility (“plasticity”) of CRBN. [29, 33]

In addition to considering whether *any* ternary complexes can be correctly predicted for each System, it is also useful to consider whether these correct structures

Table 22.3 Results for Approach 1 applied to the 32 ternary complexes of Table 22.1 across Scenarios 1–4.

	Scenario 1	Scenario 2	Scenario 3	Scenario 4
<i>All Predictions^{a)}</i>				
<10 Å (Protein)	167	223	185	330
<10 Å (MG)	350	754	344	828
<i>Per System^{b)}</i>				
<10 Å (Protein)	27	28	27	30
<10 Å (MG)	31	32	31	32
<i>Best protein score, per System^{c)}</i>				
<10 Å (Protein)	20	19	19	22
<10 Å (MG)	9	12	8	11
<10 Å (Both)	6	8	5	8
PDBs with <10 Å (Both)	3M50, 4IHL, 6Q0R, 6Q0V, 6RHC, 6SJ7	3M50, 4IHL, 6Q0R, 6Q0V, 6RHC, 6R JL, 6RKK, 6SJ7	3M50, 4IHL, 6Q0R, 6R JL, 6SJ7	3M50, 3M51, 4IHL, 6Q0R, 6RHC, 6R JL, 6SJ7, 6SLW
<i>Best MG score, per System^{d)}</i>				
<10 Å (Protein)	4	2	2	8
<10 Å (MG)	14	16	10	16
<10 Å (Both)	3	2	2	3
PDBs with <10 Å (Both)	1FAP, 6Q0R, 6Q0V	1FAP, 6PAI	1FAP, 3M50	1FAP, 4IHL, 6Q0R

a) The number of successful predictions, across 3200 total predictions per Scenario.

b) The number of Systems (ternary complexes in Table 22.1) with successful predictions, across 32 total Systems.

c) The number of successful predictions considering only those that scored best using the protein–protein docking score, across 32 total Systems.

d) The number of successful predictions considering only those that scored best using the MG-based GBVI/wSA dG score, across 32 total Systems.

can be identified *a priori* via scoring. Two separate scores are generated for each predicted ternary complex during Approach 1: the “protein score,” i.e. the forcefield interaction energy between the two apo proteins as the PPIs are generated by protein–protein docking, and the “MG score,” i.e. the docking score resulting from placing the MG into the interfacial pocket. The utility of these two scores to identify experimentally relevant complexes is also evaluated in Table 22.3. Considering only the single ternary complex with the best protein score for each System does show a substantial number of the ternary complexes of Table 22.1 correctly reproduced in terms of the protein geometry – roughly two-thirds (19–22 out of 32), regardless of the Scenario. Moreover, of these ternary complexes with the best protein score, roughly one-third of them contain the MG correctly placed (8–12 out of 32).

However, no more than 8/32 of these best-scoring complexes have the proteins *and* the MGs both placed to within 10 Å of their crystallographic positions. The Systems with the best protein scores that also have both the proteins and the MGs correctly placed are listed in Table 22.3; these are exclusively systems containing 14-3-3 or DCAF15, the latter of which is relevant in a targeted protein degradation context. The MG score, i.e. the GBVI/wSA dG MG docking score – proves to be even less correlated to the correct predictions: no more than half of the 32 systems have their MGs correctly located in these top-scoring complexes, and the proteins are correct in, at best, only 8/32 systems. The complexes that scored best by the MG score very rarely (<10%) have both the protein *and* MG components predicted correctly. These few entirely correct predictions with the best MG score also generally contain either 14-3-3 or DCAF15. The structure of 1FAP, which is the only ternary complex in this study with a macrocyclic MG (rapamycin), is also correctly reproduced across all four Scenarios using the MG score.

Discussion. Using Approach 1, the validation ternary complexes of Table 22.1 can be recapitulated with a modest level of success (25%) if (a) only the protein–protein docking score is used to identify the single prediction of interest out of all 100 generated ternary complexes (i.e., the score when the MG is docked to the interfacial pocket is ignored); and (b) the larger protein has a site constraint added to guide PPI prediction during protein–protein docking (Scenarios 2 and 4). From a targeted protein degradation perspective, which is the primary potential therapeutic application of MGs at present [4, 16], the ability to successfully reproduce DCAF15-containing crystal structures is encouraging, but the failure to correctly reproduce *any* CRBN-containing crystal structures with Approach 1 is certainly disappointing.

However, understanding the underlying causes of this failure, particularly in regards to the CRBN-containing systems, is instructive, and indeed spurred the development of Approach 2 (see below). One limitation of Approach 1 is its reliance on the protein–protein docking score to identify the predicted pose most likely to reflect the experimental result. This score in MOE is simply the forcefield interaction energy score and was not trained on or parameterized against known protein–protein crystal structures, as is often done [34, 35] – although the applicability of this interaction energy as a scoring function was validated against protein–protein crystal structures [36]. However, these protein–protein crystal structures generally represent pairs of proteins that have coevolved to effectively interact with each other – a vastly different situation from any potential application of MGs, where the proteins that are to be glued together are chosen based only on their therapeutic relevance and not based on any inherent complementarity. As a consequence, all PPIs generated for two non-related proteins should be expected to be unoptimized and nonspecific, which is clearly challenging for any *in silico* score – and might prove especially problematic for a docking score that was empirically trained to reproduce known structures of coevolved proteins.

Another limitation of Approach 1 is the order of events in which the ternary complex is formed: prediction of the apo-apo PPI first, followed by the introduction of the MG. As traditionally defined, an MG is *necessary* in order for the correct PPI to form

in the first place – although an MG that stabilizes a PPI that forms natively, but only weakly, would also be useful – and thus Approach 1 is fundamentally flawed in this respect. However, additional work showed that modifying the order in which the ternary complex is formed does not appreciably improve the quality of the results. Specifically, for the eight CRBN-containing ternary complexes in Table 22.1, the crystallographic positions of CRBN *and* their co-crystallized MGs were protein–protein docked, as a complex, against the second, apo protein, using Scenario 2 – a reflection of the actual means by which these ternary complexes form, where the well-known neomorphic interface of liganded CRBN interacts with neosubstrates [3]. However, only one single ternary complex crystal structure from Table 22.1 (6XK9) was successfully reproduced with this modified variant of Approach 1. It should also be noted that any potential application of this variant to novel systems would require prospective knowledge of which protein can productively bind to the MG in the absence of the second protein. As this knowledge will not necessarily be available in early stage discovery projects on novel systems, further development and refinement of this (protein+MG) + protein variant of Approach 1 was not pursued.

In addition, a comparison between the PPIs produced with protein–protein docking using either apo or liganded CRBN against their accompanying second proteins revealed limited overlap between the predicted PPI ensembles (data not shown). Fundamentally, this difference is expected, as liganded CRBN presents a protein surface at the MG binding site that differs quite substantially from the surface of apo CRBN. Figure 22.2 illustrates the different surfaces for (a) apo CRBN, (b) CRBN with thalidomide present in its binding pocket, and (c) CRBN with only the glutarimide ring of thalidomide present. (It should be noted that the CRBN geometry was held constant throughout Figure 22.2, ignoring any conformational changes that may occur upon MG binding). Figure 22.2a is the form of CRBN that is, under Approach 1, encountered by the second protein during ternary complex formation, whereas Figure 22.2b shows the surface of the CRBN complex (with thalidomide's contribution shown in lighter green) that is encountered by the second protein under the variant of Approach 1 just discussed. During the protein–protein docking phase of Approach 1, the MG binding pocket in the CRBN apo surface could conceivably – and artificially – interact with, e.g. an extended lysine sidechain. Conversely, the ridge formed by thalidomide in Figure 22.2b could perhaps occupy a small sub-pocket on the second protein in a predicted PPI. The surface in Figure 22.2c is intermediate between these two extremes, where the glutarimide ring of thalidomide fills the MG binding site but does not otherwise extend above it, thereby presenting a flat “plain” to the second protein during protein–protein docking.

The situation depicted in Figure 22.2c, although somewhat artificial, is interesting because it resembles many of the surfaces of the protein–ligand complexes used as inputs when modeling PROTAC-mediated ternary complexes in Method 4B [18]. Specifically, the binders in the pockets of the proteins used by Method 4B tend to fill in a cavity or cleft along the surface, but generally do not present much functionality extending beyond the protein surface. Moreover, although the protein–ligand complexes provided as inputs for Method 4B often actually do exist (i.e. as can be found in binary protein–ligand cocrystal structures), these complexes are in fact

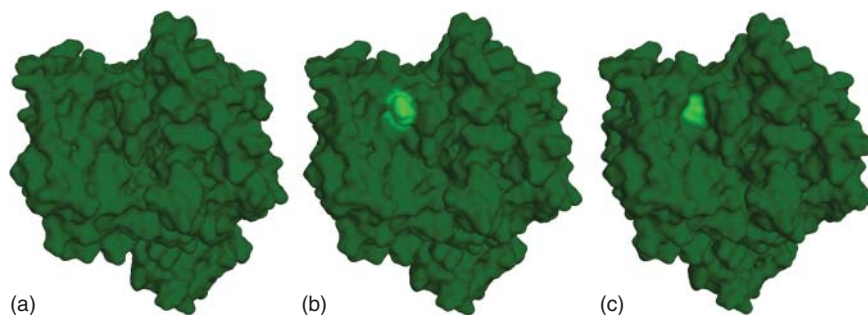


Figure 22.2 Molecular surfaces for cereblon with the IMiD-binding pocket filled with (a) nothing, (b) thalidomide (light green surface), and c) only the glutarimide ring of thalidomide.

imprecise building blocks for modeling PROTAC-mediated ternary complexes, as these binary complexes lack both the linker and the second binding moiety. Despite this approximation, however, Method 4B is capable of reproducing known experimental structures and degradation tendencies across a number of different situations [18]. Therefore, inspired by the past success of Method 4B, we developed a modeling approach, described below as Approach 2, where an MG-mediated ternary complex is assembled not from two apo proteins as in Approach 1, but instead from two protein+partial MG binder complexes, analogous to how ternary complexes are assembled in Method 4B.

22.3.2 Approach 2: Treating MGs as “linkerless PROTACs”

Procedure. When PROTAC-mediated ternary complexes are modeled using Method 4B, the user must supply two proteins with warheads (i.e. bound ligands similar to either end of the PROTACs) appropriately placed in the protein pockets, in addition to one or more PROTACs. This information is used throughout the simulation: the protein residues near the bound warheads provide site constraints to guide the protein–protein docking phase of Method 4B, and the positions of the bound warheads guide the placement of the full PROTACs into the pockets during ternary complex assembly. In Approach 2, where MGs are considered as “linkerless PROTACs” using a Method 4B-like approach, conceptually the required information is the same: two site constraint definitions are needed to guide protein–protein docking, and two pre-placed small molecule templates are required to guide the placement of the MGs in the pockets. The difference with Approach 2 for modeling MG-mediated ternary complexes is that the user need not explicitly provide all of this information – any missing information will be generated on-the-fly, as will be described below. Nevertheless, some minimum level of information is required to invoke Approach 2 rather than Approach 1, as described by Scenarios 4–9 in Table 22.2. In other words, to use Approach 2, at a minimum, the user must provide either (a) a site constraint on both proteins to guide protein–protein docking (Scenario 4) or (b) a site constraint that steers the protein–protein docking for just one of the proteins, as well as a pre-placed

small molecule template to guide the placement of the MGs into this same protein (Scenarios 5 or 6). Once this minimal level of information has been provided, Approach 2 is invoked by changing a setting (known as the Partitioning scheme) away from the default value of Do Not Partition. Whereas PROTACs are inherently partitioned into three parts in Method 4B (two binding moieties and a linker), MGs in Approach 2 are partitioned into just the two binding moieties, i.e. MGs are treated as linkerless PROTACs. There are, however, two decision points when MGs are partitioned: first, where (*i.e.*, at which bond) the MG should be partitioned, and second, which resulting “partial MG” is associated with which protein.

For the first point, an interface has been developed where a particular bond can be set at the split point. Only single (nonaromatic) bonds can be considered as potential split points, and both resulting partial MGs after any potential split point bond is broken must have more than one non-hydrogen atom. If an MG containing a macrocyclic ring is to be partitioned (and the ring itself is to be broken rather than fully assigned to one of the partial MGs), then it must be split twice, with the user selecting two bonds as split points. However, cyclohexyl rings (and smaller) cannot be split twice in this fashion, but instead must fully belong to one partial MG. Importantly, this manual assignment of partition split points can also be performed automatically, where the single bonds that most evenly divide the MGs are chosen as the split points. In fact, it was found (data not shown) that, for the 32 MGs in the validation set of this work (Table 22.1), this automated partitioning scheme provided the best results overall for reproducing known ternary complex crystal structures, and thus all data shown below for Approach 2 was generated using this automatic Partitioning scheme.

The second decision that has to be made in partitioning MGs is judging which partial MG should be “assigned” to which protein. In PROTAC-mediated ternary complexes, there is no ambiguity on this point, due to the modular nature of PROTACs: one binder is by definition an E3 ligase recruiter and should thus be assigned to the accompanying E3 ligase, with the other binder clearly responsible for binding to the protein-of-interest. Similarly, in some MG-mediated ternary complexes, there is also no ambiguity: for example, it is well-known [3] that the glutarimide ring of the IMiDs (Figure 22.1) binds to a tri-Trp pocket in CRBN, and so any partial MG generated via Partitioning that contains this moiety should be assigned to CRBN. However, in other cases (such as for the DCAF15-recruiting MGs in Table 22.1), it is less clear that one particular partial MG is primarily responsible for binding to one specific protein. In these situations, the user may manually assign a specific partial MG to a specific protein – or, as above, this assignment can be performed automatically. In this automatic assignment procedure, each partial MG is docked against each protein, keeping only poses where the atom that reconnects to the full MG is >20% solvent-exposed (relative to its solvent exposure in the partial MG without any protein present). The best scoring pose (using the GBVI/wSA dG scoring function) is taken as the score for each protein-partial MG combination, and the specific pairing of partial MG + protein that gives the best overall docking score is taken as the automatic assignment. This automatic procedure fully recapitulates known, unambiguous pairings, such as with CRBN as discussed above or in cases where

there is a clear spatial delineation between the two “ends” of an MG. In cases where this presumptive pairing is not known (e.g. for the DCAF15-containing complexes in Table 22.1), reversing this automatically assigned pairing generally gave poorer results (data not shown), and thus all data presented below was generated utilizing this automated procedure to decide which partial MG should be assigned to which protein.

After the automatic Partitioning and assignment procedures described above, the two resulting partial MGs must also be explicitly placed within the two proteins, ultimately generating two protein-binder complexes, wholly analogous to those used in Method 4B to model PROTAC-mediated ternary complexes. [18] The specifics of how these two protein-binder complexes are generated depend on the amount of information provided by the user, which is tabulated in the various Scenarios in Table 22.2. We begin with Scenario 9, which contains the most user-provided information: the structures of two proteins (which are always required), the specification of a site constraint on each protein to guide the protein-protein docking phase of Approach 2, and the specification of a small molecule Pose template bound to each protein. (*N.B.* that the complexes used by Method 4B for modeling PROTAC-mediated ternary complexes also meet the definition of Scenario 9). In Scenario 9, the partial MGs generated with the automatic Partitioning procedure described above are placed into their respective protein pockets using a maximum common substructure (MCS) algorithm [37] based on the user-provided Pose templates. For the validation data work of this study (Table 22.1), the fuzzy matching capability afforded by this MCS approach is unnecessary, as the binders used for the Pose templates exactly match the partial MGs generated with the automatic Partitioning procedure. However, in actual applications, this MCS approach facilitates the use of a single common Pose template for rapidly investigating MG variants that all contain a common scaffold.

As mentioned above, unlike with Method 4B for modeling PROTAC-mediated ternary complexes, Approach 2 for modeling MG-mediated ternary complexes does not require the specification of all of the information described by Scenario 9. Under Scenarios 7 and 8, one piece of information is missing relative to Scenario 9 – a small molecule Pose template on one of the proteins, used to guide the placement of the partial MG into its protein pocket. (It should be emphasized that Scenarios 7 and 8 exist as two separate Scenarios only for this validation study: Scenario 7 specifies the binding Pose template only for the larger protein of each ternary complex in Table 22.1, whereas Scenario 8 specifies the binding Pose template only for the smaller protein. In general terms, these two Scenarios both refer to a situation where it is known how a putative MG interacts with one protein but not the other. The validation results presented below consider Scenarios 7 and 8 together.) In order to generate this missing protein-partial MG binary complex, first the two partial MGs generated via automatic Partitioning are each matched, using the MCS algorithm, against the binder Pose that is provided; the partial MG that matches less completely is defined as belonging to the apo protein. This partial MG is then docked into the apo protein, and only poses where the atom on this partial MG that reconnects to the full MG is >20% solvent exposed are kept. By default, the single best scoring

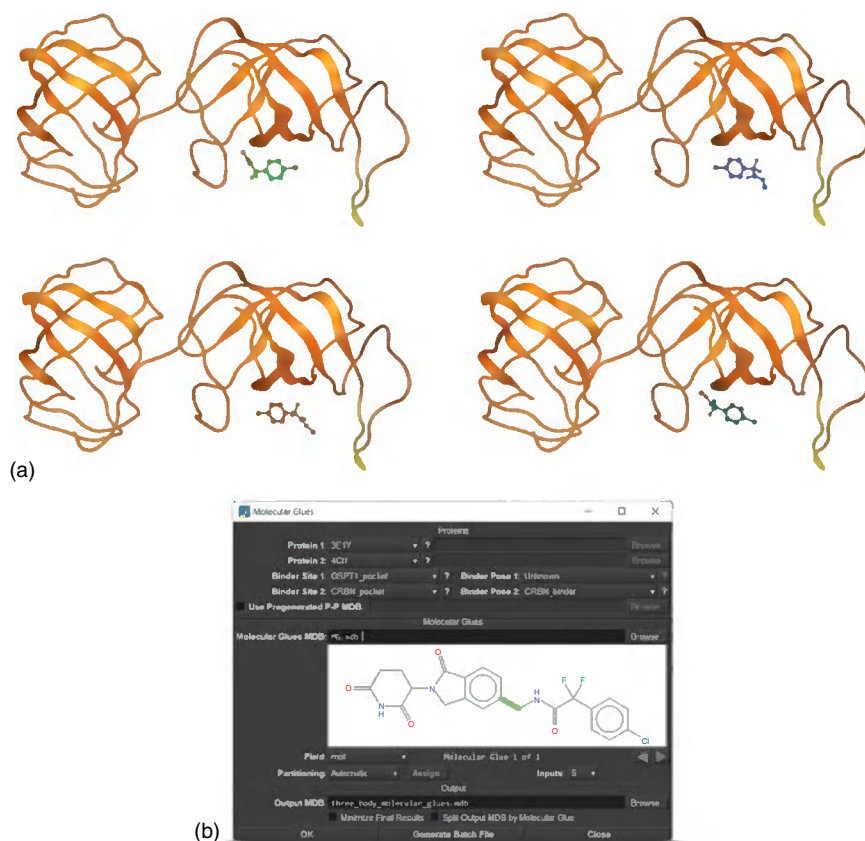


Figure 22.3 (a) Multiple input files can be automatically generated using the Inputs setting. In this example, as shown in (b), Binder Pose 1 was left at the default setting of Unknown, and the Inputs setting (lower-right) was adjusted to 5. Only four nonredundant poses were generated where the atom reconnecting to the rest of the MG is >20% solvent exposed.

(using GBVI/wSA dG) of these acceptable poses is defined as the protein-partial MG complex under Scenarios 7 and 8. However, as shown in Figure 22.3, it is also possible to consider multiple potential protein-partial MG complexes, by using up to five of these docked poses (as governed by the setting specified via the Inputs option on the GUI). As shown in Figure 22.3, these different poses, while all containing a (partially) solvent-exposed reconnection point, present different exit vectors for reassembling the full MG and will thus generate different ensembles of predicted ternary complexes. Each of these protein-partial MG input complexes will be used in separate Approach 2 simulations, with the results from all simulations collated and analyzed at the end (see below for further details). The effect of considering multiple *in silico*-generated protein-partial MG complexes on the accuracy of reproducing known MG-mediated ternary complexes will be discussed in further detail below after the details of the other Scenarios of Table 22.2 under Approach 2 are described.

In Scenarios 5 and 6, all required information is provided for one of the proteins – its structure, its interacting Site for constraining protein–protein docking, and a template small molecule Pose for defining how MGs fit into its pocket – but for the other protein, only the protein structure is provided. In order to generate the missing Site and Pose information, first a protein Site constraint must be defined. This site definition will serve not only as a constraint during protein–protein docking, but will also limit the region of the protein the partial MG will be docked into to generate the missing Pose. Thus, the *in silico* procedure used to generate this missing site constraint incorporates both small molecule and protein information. In particular, MOE's Site Finder tool is used to identify pockets on the unconstrained protein that can potentially accommodate small molecules, and hydrophobic protein surface patches are also evaluated. Rather than simply returning the pocket that gives the best Site Finder score [31], a lower-scoring pocket that is collocated with a hydrophobic protein surface patch may instead be returned. As in Scenarios 7 and 8, the Inputs option can be adjusted from its default value of 1 to generate up to 5 independent site constraint definitions. Once this missing site has been generated, the same protocol described above for Scenarios 7 and 8 is used to produce the missing Pose template definition, with the exception that only the single best-scoring Pose is returned after docking the partial MG into the protein, regardless of the value of the Inputs option (i.e. the extra Inputs have already been produced while establishing the site constraint).

As mentioned above, it is currently not possible to utilize Approach 2 under the limited knowledge of Scenarios 1–3, and thus the final Scenario to be considered with Approach 2 is Scenario 4, where both protein site constraints have been specified, but no information about the small molecule Pose templates used to guide partial MG placement has been provided. In this Scenario, the two partial MGs generated with the Partitioning procedure are each separately docked against both proteins, into the region of the protein described by the specified site constraint. Across the resulting four small molecule docking simulations, only poses where the reconnecting atom on the partial MG is >20% solvent exposed are kept, and the partial MG-protein pairing that gives the best summed docking scores is assigned as correct (e.g. partial MG A with protein 1 and partial MG B with protein 2). Once this correct pairing has been established, up to five separate pairs of partial MG-protein input complexes can be generated, as governed by the Inputs setting.

The steps outlined above for each Scenario ultimately result in two protein-partial MG complexes, and optionally with multiple proposed complexes, if the Inputs setting has been increased from its default value of 1. In Approach 2, these binary input complexes are then combined with the user-provided MGs to generate MG-mediated ternary complexes. The computational protocol to assemble these ternary complexes is similar to that published for Method 4B [18], with a few minor adjustments: there is no filtering of “acceptable” ternary complexes based on interfacial surface area; multiple independent protein–protein docking runs are always performed, to sample the PPI more effectively; and MG conformations are always generated using MOE's LowModeMD method [38], rather than any other conformational search algorithm. For this last step, the portions of the MGs that match (using the MCS algorithm) the

corresponding binder Poses are held rigid. In this validation study, there is always an exact match between the specified MGs (Table 22.1) and the binder Poses used (be they explicitly specified, as in Scenario 9, or computationally generated), and so this conformational search is simply a torsional scan of the single bond selected as the split point by the automatic Partitioning scheme. However, in more realistic simulations, a library of specified MGs may possess, for example, R group substituents that are not contained in the binder template Pose, and thus conformations of these unmatched components will also be sampled during the conformational search.

Finally, the ternary complexes produced using Approach 2 are always subjected to our double clustering protocol, as was previously used for modeling PROTAC-mediated ternary complexes [18]. If multiple Inputs are generated under Scenarios 4–8, then not only is each individual simulation independently clustered using this double cluster protocol, but also all ternary complexes produced in each independent simulation are collated into a single database, which itself is then also double clustered to generate a “pan” simulation double cluster.

Results and Discussion. Figure 22.4 shows the results of applying Approach 2 to the 32 ternary complexes in Table 22.1, utilizing different levels of information provided to the simulations, as defined by Scenarios 4 (purple bars), 5 and 6 (averaged as yellow bars), 7 and 8 (averaged as blue bars), and 9 (green bars). Figure 22.4a presents the results where the Inputs setting (see above) was left at the default value of 1, while Figure 22.4b shows the results when this setting was adjusted to 5 (i.e. Approach 2 automatically generated five independent sets of protein-partial MG input complexes, as needed). The green bars, where Two Sites and Two Poses were fully specified (i.e. Scenario 9), do not change between Figure 22.4a and b, as there is no “missing” information to generate, and thus the Inputs setting has no effect. The y-axis in both charts shows the hit rate for the largest (most populous) Double Cluster, i.e. the percent of ternary complexes in the largest Double Cluster with an RMSD for the protein alpha carbons $<10 \text{ \AA}$ (relative to the known crystal structures) and positions for the heavy atoms of the MG $<10 \text{ \AA}$ from their crystallographic positions. For Figure 22.4b, this Hit Rate corresponds to the largest “pan” Double Cluster, which was determined by collating and clustering all results for each complex in Table 22.1 generated by using multiple, automatically generated Input files. In order to easily evaluate the effects of both the different Scenarios and the use of multiple input files, the bars in each Scenario grouping were arranged from lowest to highest hit rate. Although this sorting does facilitate comparisons, it also complicates evaluations of Approach 2 as applied to specific ternary complexes of interest. Full numerical results are available upon request; additionally, ternary complexes containing CRBN have been highlighted with a red border, given their importance in targeted protein degradation and therapeutic applications.

In Figure 22.4a, where only one set of Inputs was generated when needed, clearly Scenarios 5 and 6 (yellow bars) show inferior performance compared to Scenario 4 (purple bars), which in turn is inferior to Scenarios 7 and 8 (blue bars) and Scenario 9 (green bars). Specifically, there is a 0% hit rate in 18, 22, 9, and 6 out of 32 ternary complexes (for Scenarios 4, 5/6, 7/8, and 9, respectively). It should also be noted that Scenario 4 could not be completed for 2P1Q, 6RX2, and 7BQU, and

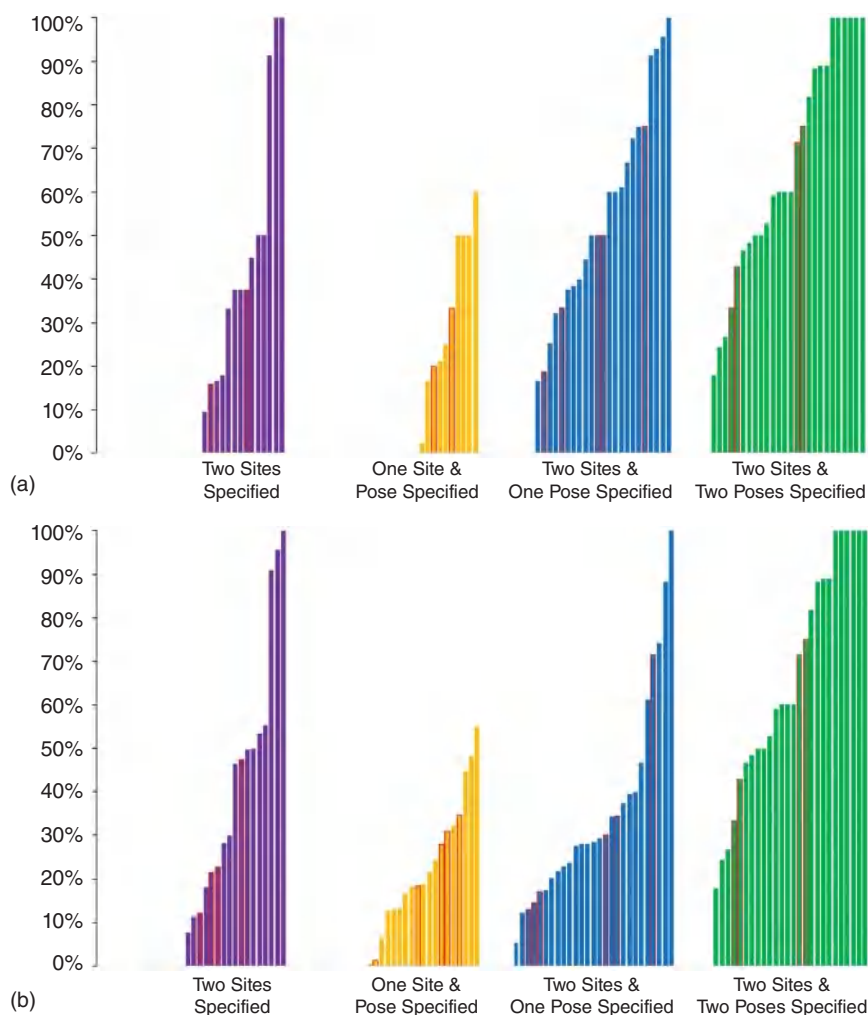


Figure 22.4 Hit Rates using Approach 2 with (a) only one Input automatically generated when needed or (b) with up to five nonredundant Inputs automatically generated. The purple bars correspond to Scenario 4, the yellow to the average of Scenarios 5 and 6, the blue to the average of Scenarios 7 and 8, and the green to Scenario 9. The bars with red borders indicate CRBN-containing ternary complexes.

similarly, Scenario 7 halted prematurely on 6RX2 and 7BQU, due to an inability of the automated Input creation workflows described above to generate any successful poses when docking the partial MGs into the smaller proteins; these errors are included in the 0% hit rate counts just provided. In comparison to Approach 1, where apo protein–protein docking is followed by whole MG docking at the PPI, the most equivalent data can be found in Table 22.3, in the Scenario 4 column using the “Best protein score, Per System,” and with both protein and MG $<10\text{ \AA}$ – which shows only 8/32 complexes correctly predicted (or 24 complete failures). Thus, even the

worst result of Approach 2 – Scenarios 5 and 6 using only one Input – yields superior performance compared to the best results of Approach 1.

The relatively poor performance for One Site and Pose Specified (Scenarios 5 and 6, yellow bars) is worthy of further discussion, particularly as a common application of Approach 2 is likely the modeling of CRBN and its well-known suite of MGs to form ternary complexes with potentially novel proteins-of-interest. The results of Figure 22.4a suggest that Approach 2 can indeed be successfully applied to CRBN-containing ternary complexes: the red bars in Figure 22.4 highlight these complexes, and there are nonzero hit rates in Figure 22.4a for 2, 2, 5, and 4 CRBN-containing systems (out of 8 total, proceeding left-to-right). This performance, while imperfect, stands in stark contrast to the results generated with Approach 1, where *none* of the CRBN-containing systems could be reproduced with $<10\text{ \AA}$ RMSD for both the proteins and the MGs. Nonetheless, the results of Figure 22.4a highlight the advantage of restricting the protein binding site on *both* proteins, i.e. the improved performance shown by the blue bars relative to the yellow. Without this extra site constraint, the PPI is wholly unconstrained for one of the proteins during protein–protein docking; as a consequence, most of the 0% hit rates generated in the One Site & Pose dataset are for ternary complexes where an entirely wrong face of the unconstrained protein is predicted to interact with the fully specified protein. Thus, in order to effectively model a protein like CRBN interacting with novel proteins of interest, additional biophysical data will be quite helpful, and possibly even necessary, to correctly predict the PPIs, as might be afforded, for example, via site-directed mutagenesis or hydrogen-deuterium exchange [39–41].

In lieu of performing additional experiments, however, the results of Figure 22.4b show that considering multiple possible Input structures for any unspecified protein-partial MG complex is also worthwhile. The most notable effect of performing these multiple simulations is that there are far fewer systems that give 0% hit rates relative to the single Input simulations: 15, 13, 5, and 6 for the purple, yellow, blue, and green bars of Figure 22.4b, respectively (out of 32 total). The benefit of performing multiple simulations with different Inputs is most pronounced for Scenarios 5 and 6 (yellow bars), where there were 0% hit rates for 22 systems in Figure 22.4a, but only 13 in Figure 22.4b. Considering only the eight CRBN-containing ternary complexes, using multiple Inputs generates successful results for 4, 5, 7, and 4 of the ternary complexes (proceeding from left-to-right in Figure 22.4b) – again an improvement relative to Figure 22.4a. However, it should be noted that if there is already a great deal of information about the system available to guide the simulation, such as in Scenarios 7 and 8 (blue bars), then considering multiple Inputs seems generally to “dilute” the quality of the results, as can be seen by the lower blue bars in Figure 22.4b relative to Figure 22.4a. As already noted, using multiple Inputs in this Two Sites & One Pose Specified situation *does* give only five failures, compared to nine when using only one automatically generated Input, but this improvement may not be worth the fivefold increase in simulation time that comes with setting Inputs to 5.

22.4 Conclusions

Two different computational Approaches have been developed and implemented in MOE to model MG-mediated ternary complexes – the first such tools available for these systems, to the best of our knowledge. A unified graphical interface for the two Approaches has been developed for the user to specify the identity of the three ternary complex components: the two proteins and one or more MGs. Additional information can optionally be provided via this interface to further refine the simulations, as detailed above. Moreover, the contents of the panel can be written to a batch file suitable for execution in parallel using high-performance or cluster computing resources.

Although both Approaches in this work rely on protein–protein docking to generate putative PPIs, it should be noted once again that Approach 1 does so using apo proteins, whereas Approach 2 docks proteins containing partial MGs bound to the proteins as well. These partial MGs are generated by breaking the user-supplied MGs into two parts, either via manual assignment of a split point or via an automated MG Partitioning protocol. After the MGs are partitioned, Approach 2 essentially treats MGs as PROTACs that lack a linker moiety, and thus previous computational tools [18] that can successfully model PROTAC-mediated ternary complexes can also be applied to MGs. The validation results detailed above show that this PROTAC-like Approach 2 clearly outperforms Approach 1, where MG-mediated ternary complexes are constructed from apo proteins.

The success of Approach 2 suggests that the traditional, stark delineation between a modular, multicomponent PROTAC and an indivisible, monovalent MG should be viewed instead as more of a continuum. Indeed, although details about ternary complex structure are often unknown or undisclosed for many of the MGs shown in Figure 22.1, it seems reasonable to conjecture, for example, that the benzyl-morpholine and piperidine-benzyl “tails” of CFT7455 and NVP-DKY709, respectively, have been developed to more effectively interact with particular proteins of interest. Mezigdomide is even more elaborate and can also be viewed as possessing at least some degree of the hallmark modularity of a PROTAC, which suggests that rational design of these more expansive MGs should be more amenable to rational design, in contrast to what has historically been the case for MGs [16]. Additionally, it is known that even slight changes in PROTAC composition can have dramatic effects on ternary complex structure and degradation behavior [41] – and thus a method like Approach 2, which generates and evaluates multiple PPI hypotheses, will likely prove quite useful in developing MGs with greater degrees of functionalization and modularity.

Finally, it should be noted that this work has been strictly concerned with recapitulating structural information, particularly in reproducing the structures of known MG-containing ternary complexes as determined with atomistic detail by X-ray crystallography. Of even greater use in MG design is the prediction of the relative efficacy of putative MGs. In Method 4B for modeling PROTAC-mediated ternary complexes, the size of the largest double cluster was shown [18] to be a useful score for rank-ordering potential PROTAC designs, as it was found to

generally correlate with experimental measurements of PROTAC efficacy such as DC50. Future research with the MG-modeling tools of this work will look to develop an analogous *in silico* metric to gauge the utility of potential MG designs – although it should be noted that there are far fewer published results comparing MGs [42] than there are for PROTACs [15]. Nonetheless, the protocols described in this work, particularly Approach 2, should prove useful in generating structural hypotheses that can be validated and then applied to rational, structure-based MG design.

References

- Schreiber, S.L. (1992). Immunophilin-sensitive protein phosphatase action in cell signaling pathways. *Cell* 70: 365–368.
- Huai, Q. et al. Crystal structure of calcineurin–cyclophilin–cyclosporin shows common but distinct recognition of immunophilin–drug complexes.
- Fischer, E.S. et al. (2014). Structure of the DDB1–CRBN E3 ubiquitin ligase in complex with thalidomide. *Nature* 512: 49–53.
- Sasso, J.M. et al. (2022). Molecular glues: the adhesive connecting targeted protein degradation to the clinic. *Biochemistry* <https://doi.org/10.1021/acs.biochem.2c00245>.
- Donovan, D.H., Luh, L.M., and Cromm, P.M. (2023). Targeted protein degradation – the story so far. In: *Inducing Targeted Protein Degradation* (ed. P. Cromm), 1–24. Wiley. <https://doi.org/10.1002/9783527836208.ch1>.
- Bussiere, D.E. et al. (2020). Structural basis of indisulam-mediated RBM39 recruitment to DCAF15 E3 ligase complex. *Nat. Chem. Biol.* 16: 15–23.
- Winter, G. E. et al. Phthalimide conjugation as a strategy for in vivo target protein degradation.
- Lu, J. et al. (2015). Hijacking the E3 ubiquitin ligase cereblon to efficiently target BRD4. *Chem. Biol.* 22: 755–763.
- Sakamoto, K.M. et al. (2001). Protacs: chimeric molecules that target proteins to the Skp1–Cullin–F box complex for ubiquitination and degradation. *Proc. Natl. Acad. Sci. U S A* 98: 8554–8559.
- Zengerle, M., Chan, K.-H., and Ciulli, A. (2015). Selective small molecule induced degradation of the bet bromodomain protein BRD4. *ACS Chem. Biol.* 10: 1770–1777.
- Che, Y., Gilbert, A.M., Shanmugasundaram, V., and Noe, M.C. (2018). Inducing protein-protein interactions with molecular glues. *Bioorg. Med. Chem. Lett.* 28: 2585–2592.
- Mullard, A. (2021). Targeted protein degraders crowd into the clinic. *Nat. Rev. Drug Discov.* 20: 247–250.
- Gadd, M.S. et al. (2017). Structural basis of PROTAC cooperative recognition for selective protein degradation. *Nat. Chem. Biol.* 13: 514–521.
- Sun, X. et al. (2019). PROTACs: great opportunities for academia and industry. *Signal Transduct. Target. Ther.* 4: 64.

- 15 Weng, G. et al. (2021). PROTAC-DB: an online database of PROTACs. *Nucleic Acids Res.* 49: D1381–D1387.
- 16 Dong, G., Ding, Y., He, S., and Sheng, C. (2021). Molecular glues for targeted protein degradation: from serendipity to rational discovery. *J. Med. Chem.* 64: 10606–10620.
- 17 Drummond, M.L. and Williams, C.I. (2019). In silico modeling of PROTAC-mediated ternary complexes: validation and application. *J. Chem. Inf. Model.* 59: 1634–1644.
- 18 Drummond, M.L., Henry, A., Li, H., and Williams, C.I. (2020). Improved accuracy for modeling PROTAC-mediated ternary complex formation and targeted protein degradation via new in silico methodologies. *J. Chem. Inf. Model.* 60: 5234–5254.
- 19 Molecular Operating Environment (MOE).
- 20 Petzold, G., Fischer, E.S., and Thomä, N.H. (2016). Structural basis of lenalidomide-induced CK1 α degradation by the CRL4CRBN ubiquitin ligase. *Nature* 532: 127–130.
- 21 Matyskiela, M.E. et al. (2016). A novel cereblon modulator recruits GSPT1 to the CRL4CRBN ubiquitin ligase. *Nature* 535: 252–257.
- 22 Sievers, Q.L. et al. (2018). Defining the human C₂H₂ zinc finger degrome targeted by thalidomide analogs through CRBN. *Science* 362, eaat0572.
- 23 Matyskiela, M.E. et al. (2020). Crystal structure of the SALL4–pomalidomide–cereblon–DDB1 complex. *Nat. Struct. Mol. Biol.* 27: 319–322.
- 24 Surka, C. et al. (2021). CC-90009, a novel cereblon E3 ligase modulator, targets acute myeloid leukemia blasts and leukemia stem cells. *Blood* 137: 661–677.
- 25 Furihata, H. et al. (2020). Structural bases of IMiD selectivity that emerges by 5-hydroxythalidomide. *Nat. Commun.* 11: 4578.
- 26 Du, X. et al. (2019). Structural basis and kinetic pathway of RBM39 recruitment to DCAF15 by a sulfonamide molecular glue E7820. *Structure* 27, 1625–1633.e3.
- 27 Faust, T.B. et al. (2020). Structural complementarity facilitates E7820-mediated degradation of RBM39 by DCAF15. *Nat. Chem. Biol.* 16: 7–14.
- 28 Guillory, X. et al. (2020). Fragment-based differential targeting of PPI stabilizer interfaces. *J. Med. Chem.* 63: 6694–6707.
- 29 Watson, E.R. et al. (2022). Molecular glue CELMoD compounds are regulators of cereblon conformation. *Science* 378: 549–553.
- 30 Jumper, J. et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596: 583–589.
- 31 Soga, S., Shirai, H., Kobori, M., and Hirayama, N. (2007). Use of amino acid composition to predict ligand-binding sites. *J. Chem. Inf. Model.* 47: 400–406.
- 32 Corbeil, C.R., Williams, C.I., and Labute, P. (2012). Variability in docking success rates due to dataset preparation. *J. Comput. Aided Mol. Des.* 26: 775–786.
- 33 Nowak, R.P. et al. (2018). Plasticity in binding confers selectivity in ligand-induced protein degradation. *Nat. Chem. Biol.* 14: 706–714.
- 34 Huang, S.-Y. (2014). Search strategies and evaluation in protein–protein docking: principles, advances and challenges. *Drug Discov. Today* 19: 1081–1096.

- 35 Huang, S.-Y. (2015). Exploring the potential of global protein–protein docking: an overview and critical assessment of current programs for automatic ab initio docking. *Drug Discov. Today* 20: 969–977.
- 36 Pierce, B.G., Hourai, Y., and Weng, Z. (2011). Accelerating protein docking in ZDOCK using an advanced 3D convolution library. *PLoS One* 6: e24657.
- 37 Bron, C. and Kerbosch, J. (1973). Algorithm 457: finding all cliques of an undirected graph. *Commun. ACM* 16: 575–577.
- 38 Labute, P. (2010). LowModeMD—implicit low-mode velocity filtering applied to conformational search of macrocycles and protein loops. *J. Chem. Inf. Model.* 50: 792–800.
- 39 Eron, S.J. et al. (2021). Structural characterization of degrader-induced ternary complexes using hydrogen–deuterium exchange mass spectrometry and computational modeling: implications for structure-based design. *ACS Chem. Biol.* 16: 2228–2243.
- 40 Zorba, A. et al. (2018). Delineating the role of cooperativity in the design of potent PROTACs for BTK. *Proc. Natl. Acad. Sci. U S A* 115.
- 41 Smith, B.E. et al. (2019). Differential PROTAC substrate specificity dictated by orientation of recruited E3 ligase. *Nat. Commun.* 10: 131.
- 42 Hansen, J.D. et al. (2021). CC-90009: a cereblon E3 ligase modulating drug that promotes selective degradation of GSPT1 for the treatment of acute myeloid leukemia. *J. Med. Chem.* 64: 1835–1843.

23

Free Energy Calculations in Covalent Drug Design

Levente M. Mihalovits, György G. Ferenczy, and György M. Keserű

Medicinal Chemistry Research Group, Drug Innovation Centre, Research Centre for Natural Sciences, Magyar tudósok körútja 2, Budapest 1117, Hungary

23.1 Introduction

Covalent inhibitors bind to their protein target by forming a chemical bond between the reactive electrophilic part (warhead) of the inhibitor and the targeted nucleophilic sidechain of the protein, the latter is typically cysteine, lysine, serine, threonine, or tyrosine. Drugs with covalent mechanisms possess potential advantages over their noncovalent counterparts, such as longer residence time, higher selectivity, and lower dosage requirements.

However, off-target reactivity and idiosyncratic toxicity are possible risks emerging from the electrophilic nature of the warhead. Historically, covalent inhibitors are present from the early years of medication; aspirin, penicillin, omeprazole, clopidogrel, and numerous other marketed drugs act via a covalent mechanism. Early covalent drugs were discovered serendipitously and therefore design principles, discovery tools, and development strategies were mostly missing. Consequently, there was a hesitance to develop covalent inhibitors owing to the abovementioned risks attributed to chemical reactivity. Methodological developments and the results of chemical biology converged to a paradigm shift which occurred during the early 2000s. The pharma industry re-evaluated the importance and possible advantages of covalent inhibitors and since then the covalent mechanism of action has become an essential drug-targeting approach, producing a number of new drugs, especially in oncology-related indications.

23.2 Mechanism of Covalent Inhibition

Covalent inhibition is typically described as a two-step process. In the first step, the ligand and the protein form a noncovalent complex. This process is governed by molecular recognition and leads to a complex where the reactive group of the ligand (warhead) and the targeted nucleophilic residue of the protein are in proximity. The

Computational Drug Discovery: Methods and Applications, First Edition.

Edited by Vasanthanathan Poongavanam and Vijayan Ramaswamy.

© 2024 WILEY-VCH GmbH. Published 2024 by WILEY-VCH GmbH.

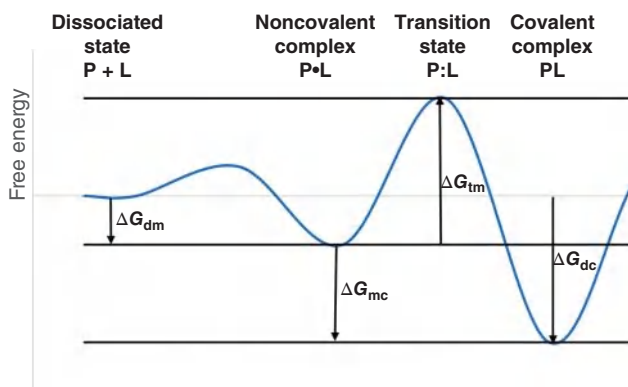
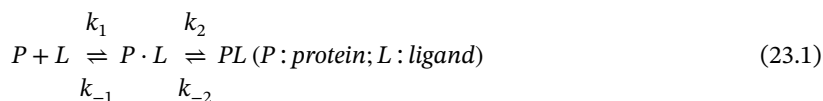


Figure 23.1 Schematic free energy profile of the two-step process of covalent inhibition. ΔG_{dm} is the binding free energy of the noncovalent complex, ΔG_{mc} and ΔG_{dc} are the free energy gain of covalent complex formation with respect to the noncovalent complex and the dissociated state, respectively, and ΔG_{tm} is the free energy barrier of the formation of the covalent complex from the noncovalent complex.

next step is the chemical reaction, where bond formation occurs, and the covalent inhibitor-protein complex is formed (Eq. 23.1).



Here k_1 and k_{-1} is the binding and dissociation rate constant, respectively, for the noncovalent complex formation, and k_2 and k_{-2} is the reaction rate constant of the covalent complex formation and decomposition, respectively. The schematic free energy profile of this two-step process is shown in Figure 23.1.

While the first step of the process is typically reversible, depending on the free energy profile, the second bond-forming step can be either reversible or irreversible. A modest reaction barrier (ΔG_{tm}) and reaction-free energy (ΔG_{mc}) allow not only the covalent bond formation but also the reverse process, the bond breaking. Then the reformulation of the noncovalent complex occurs with significant speed and the opposite processes lead to chemical equilibrium. This contrasts with chemical reactions with high barriers and low reaction energy that make k_{-2} of Eq. (23.1) negligible and the reaction mechanistically irreversible. In this case, the kinetic half-life of the covalent complex equals the re-synthesis rate of the target protein. The computational description of reversible versus irreversible inhibition requires different approaches, as discussed in the forthcoming sections.

23.3 Computational Characterization of Reversible Covalent Binding

Reversible covalent ligand binding can be described by equilibrium processes. The first step, namely the formation of the noncovalent complex, is typically a low-barrier

fast process both in the case of reversible and irreversible binders. The complex formation brings the reactive group of the ligand and the nucleophile residue of the protein in close proximity so that the chemical reaction can occur as a second step. In the present discussion, we follow the general approach and characterize the noncovalent complex formation with the equilibrium constant, thus focusing on the thermodynamics of the complex formation. However, the kinetics of the dissociation, k_{-1} of Eq. (23.1) often designated by k_{off} , may also affect the subsequent covalent step as fast-dissociating compounds might have less chance to form the covalent bond. This consideration is typically missing in the discussion of covalent inhibition, although it might be necessary to examine kinetics when the dissociation of the noncovalent complex is fast.

The characteristic difference between reversible and irreversible inhibitors is in the free energy profile of the covalent step. The bond formation of reversible ligands is associated with low barriers and modest reaction-free energy. The chemical equilibrium can be assumed between the dissociated state and the noncovalent complex, on one hand, and between the noncovalent complex and the covalent complex, on the other hand. The dissociation constant (K_d) for the total process can be expressed with the equilibrium constants of the two steps [1].

$$K_d = \frac{[P][L]}{[PL] + [P \cdot L]} = \frac{K_1 K_2}{K_2 + 1} \quad (23.2)$$

with

$$K_1 = \frac{k_{-1}}{k_1} = \frac{[P][L]}{[P \cdot L]} = \exp\left(\frac{\Delta G_{dm}}{RT}\right) \text{ and } K_2 = \frac{k_{-2}}{k_2} = \frac{[P \cdot L]}{[PL]} = \exp\left(\frac{\Delta G_{mc}}{RT}\right) \quad (23.3)$$

and K_d can be expressed with free energy changes [1]

$$K_d = \frac{1}{\frac{1}{K_1} + \frac{1}{K_1 K_2}} = \frac{1}{\exp\left(-\frac{\Delta G_{dm}}{RT}\right) + \exp\left(-\frac{\Delta G_{dc}}{RT}\right)} \quad (23.4)$$

where the $\Delta G_{dc} = \Delta G_{dm} + \Delta G_{mc}$ relation was used (*cf.* Figure 23.1).

Although Eq. (23.4) establishes the relationship between the experimentally measurable dissociation constant K_d , and binding free energy changes ΔG_{dm} and ΔG_{dc} , the calculation of the latter quantities is not straightforward. We recall that the affinity differences of noncovalent inhibitors are most often calculated by molecular dynamics (MD)-based alchemical transformations using thermodynamic cycles. However, noncovalent inhibition is a single-step process framed in Figure 23.2, and therefore the calculation of the $\Delta G_N - \Delta G_D$ difference gives the affinity difference $\Delta G_{dm}(2) - \Delta G_{dm}(1)$ directly. By contrast, the affinity of covalent inhibitors generally depends on both steps, and further considerations are needed to apply thermodynamic cycles to obtain affinity differences. When ΔG_{dc} , the reaction energy of the covalent complex formation is significantly lower than ΔG_{dm} , the noncovalent binding free energy, then Eq. (23.4) reduces to

$$K_d = \exp\left(\frac{\Delta G_{dc}}{RT}\right) \quad (23.5)$$

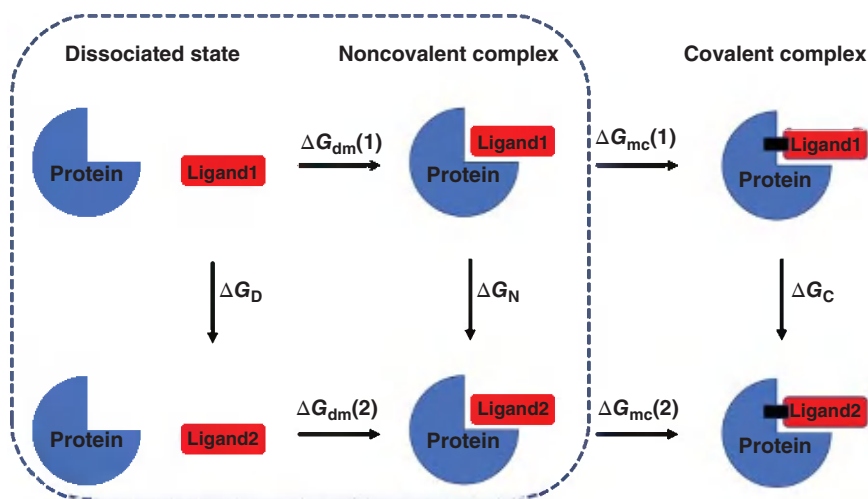


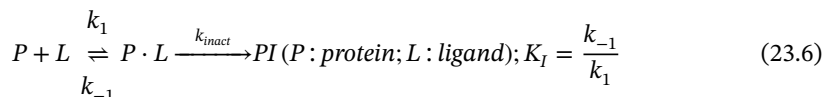
Figure 23.2 Thermodynamic cycles for the binding of two covalent ligands. The noncovalent complex formation, the step present in both noncovalent and covalent ligand binding, is framed.

and the calculation of $\Delta G_C - \Delta G_D$ formally gives the affinity difference $\Delta G_{dc}(2) - \Delta G_{dc}(1) \approx \Delta G_{mc}(2) - \Delta G_{mc}(1)$. It must be noted, however, that the calculation of ΔG_C includes the alchemical transformation of two covalently bound ligands and molecular mechanics (MM) force fields that are not expected to properly describe the free energy differences owing to ligand reactivity changes. Therefore, we should assume that ligand reactivities are unaltered. This assumption may be valid when the warhead is the same and structural differences of the ligands are restricted to regions distant from the warhead. However, even small changes in the ligand structure may alter the secondary interactions and the binding pose in the noncovalent complex, and this may affect reactivities. The various approaches and thermodynamic cycles applied in calculating the affinities of covalent inhibitors based on Eqs. (23.4) and (23.5) are presented in the section of case studies. A general discussion of calculating the binding free energy difference of two ligands is presented in connection with the noncovalent binding step of irreversible inhibitors that corresponds to the framed thermodynamic cycle in Figure 23.2.

23.4 Computational Characterization of Irreversible Covalent Binding

Irreversible covalent inhibition also follows the previously introduced two-step mechanism, however, the second, covalent step – unlike in reversible inhibition – contains a large activation barrier and the product state is in a deep minimum. Consequently, the reverse reaction is not likely to happen, and the

reaction becomes irreversible. Then the reaction described by Eq. (23.1) can be written in the irreversible case as follows.



The two main steps of irreversible covalent inhibition, namely molecular recognition and covalent labeling, can be described by the K_I equilibrium constant and the k_{inact} rate constant, respectively (Eq. (23.6)). The k_{inact} and K_I notations are typically used in irreversible enzyme inhibition, and they correspond to k_2 in Eq. (23.1) and K_1 in Eq. (23.3), respectively. Although covalent labeling is not restricted to enzymes, we will use this generally applied notation for the kinetic and thermodynamic characterization of irreversible covalent inhibition. These quantities can be derived from experiments; however, the separate determination of K_I and k_{inact} is ponderous [2, 3] and in many cases only the k_{inact}/K_I ratio is determined. The computation of K_I and k_{inact} is also feasible. Various computational chemistry methods are used to model the noncovalent and covalent binding events and to calculate the noncovalent binding free energy and the transition state free energy (ΔG_{dm} and ΔG_{tm} , respectively, on Figure 23.1) of the covalent bond formation. The relation between the ΔG_{dm} , ΔG_{tm} and K_I , k_{inact} are shown in Eqs. (23.7) and (23.8).

$$\Delta G_{dm} = RT \ln(K_I) \quad (23.7)$$

$$\Delta G_{tm} = -RT \ln \left(\frac{k_{inact}}{\frac{k_b T}{h}} \right) \quad (23.8)$$

where R is the universal gas constant, T is the absolute temperature, k_b is the Boltzmann constant and h is the Planck constant.

Computational evaluation of K_I and k_{inact} allows us to explore the structural details behind the experimental inhibitory activity and to make predictions on the affinity of ligand candidates against specific protein targets. The ideal scenario for an irreversible covalent inhibitor is having a low K_I and a moderate k_{inact} value. Low K_I corresponds to high target affinity, while moderate k_{inact} corresponds to suitable reactivity toward the targeted sidechain, thus avoiding potential off-target toxicity and resulting in a higher therapeutic index. Hence, the main objective of irreversible covalent drug design is to improve the k_{inact}/K_I ratio, describing the complete covalent inhibition process.

While K_I , k_{inact} and their ratio offers a proper characterization of the covalent inhibition, in specific cases only the IC_{50} value is measured, which are the ligand concentrations that halve the activity of the inhibited target. Its experimental determination is straightforward and less demanding than the equilibrium and rate constants; however, the IC_{50} value is less suitable for comparative computational and experimental analysis due to its time dependence [4].

23.4.1 Computation of the Dissociation Constant of the Noncovalent Step

The dissociation constant K_I , characterizing the molecular recognition in the non-covalent binding step is related to the noncovalent binding free energy according to Eq. (23.7). However, instead of evaluating the binding free energies of individual ligands, most often the binding free energy difference of ligands is calculated based on the framed thermodynamic cycle in Figure 23.2. The sum of free energy changes along the closed cycle is zero and this leads to Eqs. (23.9) and (23.10).

$$\Delta G_D + \Delta G_{dm}(2) - \Delta G_N - \Delta G_{dm}(1) = 0 \quad (23.9)$$

$$\Delta \Delta G = \Delta G_{dm}(2) - \Delta G_{dm}(1) = \Delta G_N - \Delta G_D \quad (23.10)$$

Note that the cycle contains alchemical transformations in which ligand 1 smoothly transforms into ligand 2 in both bulk solvent and in the protein environment. These transformations (vertical arrows) are more efficient than performing modification from the unbound to the bound states (horizontal arrows).

The two main methods to evaluate the free energy differences in these thermodynamic cycles are free energy perturbation (FEP) and thermodynamic integration (TI). FEP is based on the Zwanzig equation [5], which defines the free energy difference between states A and B

$$\Delta F(A \rightarrow B) = F_B - F_A = -k_B T \ln \left\langle \exp \left(-\frac{E_B - E_A}{k_B T} \right) \right\rangle_A \quad (23.11)$$

where T is the temperature, k_B is Boltzmann's constant, and the triangular brackets denote an average over a simulation run for state A . Once a configuration is generated for state A , the energy for state B is also computed and their difference contributes to the ensemble average according to Eq. (23.11). Since Eq. (23.11) assumes a small perturbation between the two states, the free energy difference between states A and B is typically calculated as the sum of free energy differences obtained in several small perturbations connecting states A and B . We note that the Helmholtz free energy difference (ΔF) appearing in Eq. (23.11) well approximates the Gibbs free energy difference (ΔG) related to the experimental observations in solution.

The free energy difference between states A and B can also be calculated by TI. Then a λ parameter is defined, and the potential of the system is made equal to that of state A when $\lambda = 0$ (V_A) and that of state B when $\lambda = 1$ (V_B). The free energy difference between the two states can be written as:

$$\Delta F(A \rightarrow B) = \int_0^1 \left\langle \frac{\partial V(\lambda)}{\partial \lambda} \right\rangle_\lambda d\lambda \quad (23.12)$$

where $\langle \rangle_\lambda$ denotes ensemble average and $V(\lambda)$ is a hybrid potential. When it is set to the linear combination of the potentials of the two states

$$V(\lambda) = (1 - \lambda)V_A + \lambda V_B \quad (23.13)$$

then

$$\Delta F(A \rightarrow B) = \int_0^1 \langle V_B - V_A \rangle_\lambda d\lambda \quad (23.14)$$

According to Eq. (23.14), ΔF can be obtained by first performing simulations at various λ values to calculate the ensemble averages $\langle V_B - V_A \rangle_\lambda$, and then integrating numerically by λ .

In practice, TI or FEP includes a series of MD or Monte Carlo (MC) simulations and the post-processing of the simulation trajectories. Once the $\Delta\Delta G$ values are determined they can be transformed into absolute binding free energies by applying a shift with a specific value. This value can be determined if at least one absolute binding free energy is available either from the experiment or from the calculation. Alternatively, when a set of experimental binding free energies are available then the relative values can be uniformly shifted to obtain the best fit between calculated and experimental values.

It is worth mentioning that the thermodynamic cycle depicted in Figure 23.2 is not the only option for the calculation of binding free energy differences. More elaborate cycles can be defined, partitioning the transformations into various sub-steps such as the elimination of atomic charges, the transformation of the van der Waals radii, and the reintroduction of the atomic charges for the mutated atoms. With carefully selected cycles, the binding free energy difference between the same ligand binding to two related proteins can also be evaluated. This includes sidechain rather than ligand mutational TI or FEP. With the aid of such calculations, ligand selectivity over related proteins can be investigated.

23.4.2 Computation of the Rate Constant of the Covalent Step

The reaction rate constant (k_{inact}) of the irreversible covalent reaction is related to the free energy barrier of the covalent bond formation according to Eq. (23.8). Owing to the chemical transformation, this step cannot be described with typical classical force fields, and a quantum chemical approach is needed. Owing to the size of the protein–ligand system and also to the large number of configurations needed for estimating free energies, a mixed quantum mechanical/molecular mechanics simulation (QM/MM MD) is well suited to calculate reaction barriers. However, the application of high-level quantum chemical methods is hindered by their high computational requirements and we are typically confined to using semiempirical density functional or wave-function-based approaches like variants of the DFTB [6] or NDDO-type schemes (AM1 [7], PM3 [8], PM6 [9]). These simulations typically use enhanced sampling methods, such as steered MD (SMD) [10], umbrella sampling (US) [11], and metadynamics [12], which enable visiting high-energy regions in a reasonable amount of simulation time. All these approaches introduce biasing potentials in order to surmount energy barriers. SMD applies a time-dependent potential, whose center smoothly moves forward along the reaction coordinate as the simulation progresses. In the context of covalent inhibition, SMD is most often used to produce starting structures for US simulations. US is typically applied in a series of simulations (windows) with constant biasing potentials centered along the reaction coordinate at various positions. The original free energy profile, often termed potential of mean force (PMF), is constructed by processing the trajectories

by the weighted histogram analysis method (WHAM) [13]. Metadynamics introduces Gaussian-like biases to fill the “free energy valleys” throughout the simulations. Once the free energy landscape is modified to be flat and the sampling is uniformized for the complete configurational space, the unbiased free energy profile can be constructed considering the applied biases.

Apart from the sampling method, other essential details of these simulations are the choice of the QM and MM regions, the level of QM theory, and the boundary-, embedding- and total energy schemes. Concerning the selection of the QM region, its increase is expected to improve the accuracy of the calculations, but it also increases the required computational resources. The QM region must contain the atoms directly involved in the covalent reaction and their close proximity. A common solution for the QM/MM separation is to include the targeted protein side chain up to its C_β atom and either the whole ligand or its warhead (reactive moiety) in the QM region. In the case of large ligands, a part of the ligand may be excluded from the QM region if a suitable nonpolar bond, most often a C-C bond, can be cut. The QM/MM separation inevitably leads to dangling bonds that are most often filled by hydrogen atoms (link atom approach) [14, 15], although other options are also available, such as the frozen localized orbital [16–18] and the boundary atom [19, 20] schemes.

The interaction between the QM and MM subsystems can be treated by the mechanical (ME) or electrostatic embedding (EE) schemes. The former calculates the interaction at the MM level while performing the QM calculations in the absence of the MM region. EE includes the MM charges – typically atom-centered point charges – in the QM Hamiltonian. The two main approaches to evaluating the total energy of the system are the additive and subtractive methods [21]. The former sums up the energy of the two regions (E_{QM} and E_{MM}) and adds the interaction energy ($E_{QM-MM \text{ interaction}}$) (Eq. 23.15), whereas the subtractive approach corrects the MM energy of the whole system ($E_{MM(total)}$) with the difference between the QM ($E_{QM(QM)}$) and MM ($E_{MM(QM)}$) energies of the QM subsystems (Eq. (23.16))

$$E_{total} = E_{QM} + E_{MM} + E_{QM-MM \text{ interaction}} \quad (23.15)$$

$$E_{total} = E_{MM(total)} + E_{QM(QM)} - E_{MM(QM)} \quad (23.16)$$

The choice of the QM theory is limited by the required computational resources. Semiempirical methods allow acceptable sampling even when calculations for several ligands are to be performed as is typical in drug discovery settings. Higher-level QM methods, either DFT or wave-function-based, require excessive resources and are less feasible even for the exploration of the reaction mechanism of a single covalent inhibitor. Therefore, free energy profile calculations for the covalent step of several covalent inhibitors are currently limited to the application of approximate quantum chemical methods. However, various attempts have been made to include more involved QM methods in QM/MM free energy calculations, both in the context of reaction mechanism exploration [22–25] and covalent inhibitory activity calculations [26, 27]. Corrections applied to the PMF are specific to the investigated

reaction, and they are typically estimated by the difference between the semiempirical and higher-level QM methods. Other approaches, like estimating the effect of the protein environment on the high-level QM region by FEP calculations, were also published [28–30]. Other attempts were reported to substitute computationally demanding QM methods with artificial intelligence-derived, computationally more feasible potentials. The Δ -machine learning approach [31] learns the difference between the low- and high-level QM methods for a specific reaction and applies a correction for the energy calculated with the cheaper QM method. Such an approach might find use in covalent inhibition-free energy calculations. MM regions can be treated by different MM force fields. The balanced parametrization of proteins and organic molecules is realized in several force fields including AMBER, CHARMM, GROMOS, and OPLS.

23.5 Case Studies of Reversible Covalent Inhibition

To the best of our knowledge, the first calculation of covalent inhibitor affinities was part of a study that compared four approaches for prioritizing building blocks of cathepsin L inhibitors with an activated nitrile warhead [32]. Starting from the experimental structure of the protein–inhibitor complex the authors investigated its possible extensions toward an apolar pocket. FEP was applied to calculate ΔG_C and ΔG_D free energies (see Figure 23.2) that were used to obtain affinity differences of the ligands. This scheme assumes that the free energy gain of the covalent complex formation dominates the inhibition, and the ligand reactivities are not influenced by the structural changes separated by several bonds from the nitrile warhead. The results supported these hypotheses as FEP-based prioritization of building blocks was found to be superior to other approaches including selection by a medicinal chemist, manual modeling, and docking followed by manual filtering.

Chatterjee et al. [1] investigated α -ketoamide derivatives as reversible inhibitors of cysteine proteases calpain-1 and calpain-2 and gave a detailed rationalization of computational schemes and approximations useful in quantifying inhibitor affinities. Based on Eqs. (23.2)–(23.5), they suggested neglecting the formation of the noncovalent complex when its free energy is higher by at least 5.5 kcal/mol than that of the covalent complex. This corresponds to an about 10,000-fold difference between the two terms in Eq. (23.4) and justifies the use of the simplified Eq. (23.5). They applied a special thermodynamic cycle that included the transformation of the warhead, containing a common core structure, to the various ligands. FEP/ λ -exchange MD (FEP/ λ -REMD) simulations were performed to obtain free energy differences. Experimental affinity differences between calpain-1 and calpain-2 were compared to the calculated free energy differences. While the free energy of noncovalent complex formation had a weak negative correlation with experimental selectivities, free energy of covalent complex formation accounted well for observed selectivities. In contrast, experimental affinity changes were not reproduced either for calpain-1 or calpain-2, which might be explained by smaller variations in affinity as compared to selectivity.

Another set of α -ketoamide derivatives overlapping with those investigated in ref. [1] were studied against calpain-1 [33]. FEP was applied to calculate free energy changes of dissociated state, the noncovalent and covalent complex (see vertical transformations in Figure 23.2). In addition, the absolute binding free energy of the noncovalent complex formation for a reference compound (ΔG_{dm}) was also calculated. Together with the experimental affinity (K_d) it was used to obtain the absolute binding free energy (ΔG_{dc}) for the covalent complex formation of the reference compound using Eq. (23.4). The availability of ΔG_{dm} and ΔG_{dc} for a reference compound made it possible to calculate the analogous quantities for any ligand using the thermodynamic cycles in Figure 23.2. The experimental affinities of the ligands were compared to the computed affinities obtained as the free energy of noncovalent complex formation (ΔG_{dm}), the covalent complex formation (ΔG_{dc}), and their combination according to Eq. (23.4). The weak correlation found for the noncovalent complexes improved significantly for the covalent complexes and gave the best result when combined. This observation points out that although the chemical modifications of the ligands are distant from the warhead, their varying effect on the noncovalent complex formation is not always negligible.

The affinity variation of nitrile-containing reversible cathepsin L inhibitors was investigated by FEP calculations [34] to obtain the relative binding free energy of five ligands in noncovalent ($\Delta\Delta G_{dm}$) and covalent complexes ($\Delta\Delta G_{dc}$). In contrast to the calculations in Ref. [1], the thermodynamic cycles included the transformation among the various ligands rather than transformations from a core structure to ligands. A special treatment of halogen bonds was provided by introducing an extra point charge to represent the anisotropic charge distribution of iodine, bromine, and chlorine substituents. The authors intuitively formed the sum of ΔG_{dm} and ΔG_{dc} that provided an improved description of the experimental affinities. Another set of nitrile-containing cathepsin L inhibitors was investigated in a similar manner [35]. Again, the thermodynamic cycles were applied to calculate the free energies of the vertical transformations shown in Figure 23.2. Noncovalent (ΔG_{dm}) and covalent (ΔG_{dc}) free energies were correlated with experimental affinities, and the latter was found to better account for experimental data. These data suggest that noncovalent complex formation has limited influence on the inhibitory activity of nitrile-containing cathepsin L inhibitors.

The mechanism of the chemical reaction between α -ketoamide inhibitors and the main protease of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) was investigated by the combined empirical valence bond and PDL/SLRA/ β methods [36]. The most probable mechanism starts with the concerted cysteine attack at the carbonyl carbon atom of the bound ligand and a proton transfer to the histidine residue of the catalytic core. The reaction completes with ligand protonation by the histidine. The computed reaction barrier of ~ 13.5 kcal/mol and reaction energy of ~ -9 kcal/mol correspond to a reversible reaction that is in line with the reaction energy estimated from the experimental IC_{50} value of $0.67 \mu M$.

The irreversible and reversible inhibition of SARS-CoV-2 main protease was investigated by an acrylate and an α -ketoamide, respectively. Absolute noncovalent binding free energies were calculated using alchemical transformations, and the covalent

reaction free energy was determined by QM/MM optimizations together with frequency analysis [37]. Significant differences were found between the noncovalent and covalent binding free energies of the two compounds. The computed covalent binding free energy of the irreversible compound was found to be greater than that of the reversible inhibitor.

The reaction mechanism of reversible nitrile-containing cruzain inhibitors was investigated [26] by QM/MM MD simulations with AM1/d-phot QM Hamiltonian and CHARMM/TIP3P force field. Cysteine thiolate attack on the nitrile carbon atom and the proton transfer between a histidine residue and the nitrile N-atom was found to occur simultaneously. Similar observations were published in ref. [38], where free energy calculations of a reversible and an irreversible cruzain inhibitor were carried out by DFTB3/FF14SB potential with adiabatic mapping MP2/FF14SB corrections. Reversible nitriles were reacted via the concerted mechanism, while a consecutive mechanism was found for the irreversible inhibitor. Calculated reaction-free energies reflected well the difference between the reversible and irreversible inhibitors. Similar studies were performed for alkyne- and nitrile-based inhibitors of cathepsin K [39]. The nucleophilic attack of the active site cysteine and the proton transfer from the catalytic histidine to the inhibitor occurred simultaneously. Reaction-free energies were calculated using AM1/d-phot QM Hamiltonian and CHARMM/TIP3P force field, either with M06-2X/6-31++G(d,p) or B3LYP-D3/6-31+G(d) corrections that distinguished the reversible nitrile and irreversible alkyne inhibitors.

Da Costa and co-workers [40] calculated the free energy profiles of reversible heteroaryl nitrile inhibitors of the cysteine protease rhodesain. Reaction energies were calculated at the PM6/CHARMM level using QM/MM MD with US. The computed free energy profiles showed low transition state energies and a strong correlation was found between calculated reaction energies, and measured binding affinities for the nine examined compounds.

23.6 Case Studies of Irreversible Covalent Inhibition

Irreversible covalent inhibition has been studied extensively; however, many publications focus solely on the covalent step and neglect the formation of the noncovalent complex. The primary objective of many studies is mechanistic; they aim to explore the potential or free energy profile of the chemical reaction that occurs at the covalent step. These investigations may provide information relevant to drug discovery, and the interested reader is referred to the recent review of computational studies using quantum chemical methods to investigate the molecular mechanism of covalent inhibitors [41]. Here we review studies going beyond the exploration of reaction mechanism and assess the binding free energy profile of either multiple ligands or multiple proteins.

Active-inactive separation of potential fragment-sized MurA inhibitors was carried out successfully by performing US preceded by SMD simulations at the DFTB3/FF14SB QM/MM level [27]. The constructed PMFs showed larger activation energies for inactive compounds compared to active ones. QM correction at

ω B97XD/cc-aug-pVTZ level was applied for Michael-acceptors owing to the poorer performance of DFTB3. The free energy change associated with the noncovalent complex formation did not contribute significantly to the inhibitory activity of these small compounds.

Carbapenemase activity of class A β -lactamases was investigated by QM/MM MD simulations [42]. The deacylation reaction of the tetrahedral intermediate of carbapenem in eight different class A β -lactamases was examined using SCC-DFTB/FF12SB US simulations. Obtained activation-free energies showed a good correlation with the experimental barriers. Fritz and co-workers [43] predicted the inhibition of further class A β -lactamases by clavulanate, a drug used in combination with antibiotics. QM/MM US simulations were performed to investigate the free energy profile along the multistep reaction and the complex responsible for the irreversible inhibition was identified. Furthermore, free energy barriers were able to discriminate between clinically relevant and less effective compounds.

Inhibition of human tissue transglutaminase by acrylamide derivatives was investigated by PM3 and DFTB3 potentials with FF99SB force field [44]. The activation-free energies obtained from US simulations for both consecutive and concerted mechanisms of six selected acrylamides were correlated with experimental IC_{50} s. The transition barriers of the two-step consecutive mechanism showed better agreement with the experimental results.

The free energy surface of the reaction between N3 peptidyl Michael-acceptor and SARS-CoV-2 main protease was studied [45] using AM1/FF03 US corrected by M06-2X/6-31+G(d,p) level quantum chemical calculations. Based on the calculated reaction profile, the authors designed two potential main protease inhibitors. Both were equipped with different recognition units and one with a different warhead than the parent peptidyl compound.

Case studies discussed above focus solely on the covalent reaction and neglect the noncovalent recognition step. Other studies calculate both the noncovalent binding free energy (with the corresponding K_I) and the free energy barrier of the reaction (with the corresponding k_{inact}) thus providing a more complete description of the process of irreversible covalent inhibition. Nevertheless, such a description does not consider the binding kinetics of the noncovalent complex that might influence the covalent bond formation. The evaluation of K_I and k_{inact} requires different computational approaches, TI or FEP simulations for K_I and PMF calculations for k_{inact} . Moreover, while classical force fields are well suited for the quantitative description of the noncovalent complex formation, a QM/MM approach is typically applied for the covalent bond formation.

Differential inhibitory activity of osimertinib toward L718Q (susceptible) and T790M/L718Q EGFR (resistant) mutants was investigated by US simulations at SCC-DFTB/FF99SB level of theory [46]. Noncovalent binding free energies were obtained by WaterSwap [47] simulations and were found to be similar for the two mutants. By contrast, the conformational space of the osimertinib-EGFR complex as obtained by PMF calculations was found to be affected by the L718Q mutation resulting in the stabilization of a conformation that prevents covalent binding to the cysteine.

The inhibition of KRAS, EGFR, and Tec-kinases by several covalent binders was investigated in ref [48]. The noncovalent binding free energy differences for 10 KRAS and 5 EGFR inhibitors were evaluated using TI with the FF14SB force field. The thermodynamic cycles contained three substeps; discharging of the softcore atoms, transformation of the neutral atoms, and reintroduction of charges for the modified softcore atoms. The $\Delta\Delta G$ values were shifted with a constant to obtain the best fit to the experimental binding free energies, and the obtained energies were then transformed to K_I s using Eq. (23.7). The reaction-free energy profile for the nucleophilic substitution reaction of the KRAS inhibitors and the Michael-addition of the EGFR inhibitors was computed with US at the DFTB3/FF14SB level. Finally, the reaction activation free energies were derived from the free energy profiles and converted to k_{inact} values according to Eq. (23.8). Computed k_{inact} and K_I values and k_{inact}/K_I ratios showed fair correlation with experimental data. In case of Tec-kinases free energy calculations were used to evaluate selectivity of a Michael-acceptor ligand toward the three selected kinases, ITK, BTK, and BMX. The binding free energies of the same ligand toward the different active sites were evaluated applying the corresponding sidechain mutations in a thermodynamic cycle to obtain binding free energy differences. While the experimental selectivity between BTK and BMX was well accounted for, no sensible results for the ITK to BTK mutations were obtained that was attributed to the significant difference (five mutations) between their active sites. The Michael addition between the acrylamide ligand and the active site cysteine residue provided reasonable reaction barriers in good agreement with the experimental values.

This methodology was applied for the inhibition of immunoproteasome by a series of oxathiazolones [49]. Binding free energy differences were estimated using TI, while k_{inact} values were computed after evaluating the PMF of the rate-determining step at DFTB3/FF14SB level. The study included the exploration of the two proposed alternative mechanisms, one proceeding through a carbonate and the other through a carbonthioate intermediate. According to the PMF constructed, the carbonate route was found more feasible and the rate-determining step turned out to be a synchronous reaction comprising a nucleophilic attack on the central carbon of the oxathiazolone ring and the proton transfer between the ligand and the Thr1 residue. Selectivity difference of two oxathiazolones toward the constitutive proteasome and immunoproteasome was also evaluated by applying residue mutations to convert the binding site of one protein to the other. Experimental selectivity trends were reproduced, and the structural background of selectivity differences was identified. The Ser53Gln mutation was found to differently affect the binding pose of the examined ligands that influenced the free energy barrier of the covalent step significantly.

An alternative method estimated the k_{inact}/K_I ratio rather than its components for the inhibition of BTK by acrylamides, FAAH by carbamides, and KRAS by acrylamides (ARS series) [50]. The covalent reaction was calculated for a model system at QM (B3LYP-D3/6-311+G*) level, and the noncovalent binding term was evaluated as the effect of the enzyme environment on the transition state applying FEP with modified force-field parameters. A good correlation with the experimental results was found.

A full free energy profile for the binding of a cyanoacrylamide ligand to BTK has been generated by calculating the absolute binding free energy of the noncovalent complex and the PMF of the covalent reaction [51]. The covalent step was modeled with QM/MM MD simulations using ω B97X-D3/def2-TZVP level DFT calculations for the QM region. The absolute binding free energy was obtained by alchemical free energy transformations and was found to be close to the experimental value of a closely related ligand.

23.7 Summary

Accurate description of covalent ligand binding requires complex computational tools that are not generally applied in noncovalent drug design applications. A complete characterization of covalent ligands can be achieved by determining the free energy profile of the two-step mechanism of the binding. Computed free energy profiles can be contrasted with data derived from experiments and are useful for interpreting and predicting inhibitory activities. Mechanistic differences between reversible and irreversible covalent ligand binding suggest different approaches for the estimation of the free energy profiles. The equilibrium process of reversible inhibitors is well described by the free energy differences among the three states of ligand dissociation, noncovalent, and covalent complex formation. This characteristic of covalent binding prevents the calculation of affinity differences by thermodynamical cycles similar to those widely used for the modeling of two-state noncovalent ligands. The situation simplifies when there is a significant free energy difference between the noncovalent and covalent complexes and the system essentially contains two states, the dissociation state, and the covalent complex. The calculation of absolute binding free energies allows us to consider the presence of all three states and the complete characterization of covalent ligand binding at the expense of increased computational effort. A quantitative description of irreversible binding requires the computation of the free energy difference between the dissociation state and the noncovalent complex (as in noncovalent and reversible covalent binding) and the rate constant of the chemical reaction. While the equilibrium constant can be obtained by classical molecular simulations, the calculation of the rate constant needs quantum mechanics to properly describe bond breaking and forming. QM/MM calculations preferably combined with MD simulations are typically used to determine the free energy barrier of the reactions.

An increasing number of applications report free energy calculations for supporting covalent inhibitor design. Quantitative reproduction of the inhibitory activities has been shown in multiple publications. Further improvements are expected from approaches developed for the accurate description of the three-state equilibrium for reversible inhibition and with the availability of computational technologies that provide higher-level quantum chemical description without sacrificing sampling efficiency for irreversible ligands.

References

- 1 Chatterjee, P., Botello-Smith, W.M., Zhang, H. et al. (2017). Can relative binding free energy predict selectivity of reversible covalent inhibitors? *J. Am. Chem. Soc.* 139 (49): 17945–17952.
- 2 Strelow, J.M. (2017). A perspective on the kinetics of covalent and irreversible inhibition. *J. Biomol. Screen.* 22 (1): 3–20.
- 3 Harris, C.M., Foley, S.E., Goedken, E.R. et al. (2018). Merits and pitfalls in the characterization of covalent inhibitors of bruton's tyrosine kinase. *SLAS Discov.* 23 (10): 1040–1050.
- 4 Krippendorff, B.-F., Neuhaus, R., Lienau, P. et al. (2009). Mechanism-based inhibition: deriving K_I and k_{inact} directly from time-dependent IC_{50} values. *J. Biomol. Screen.* 14 (8): 913–923.
- 5 Zwanzig, R.W. (1954). High-temperature equation of state by a perturbation method. I. Nonpolar gases. *J. Chem. Phys.* 22 (8): 1420–1426.
- 6 Gaus, M., Cui, Q., and Elstner, M. (2011). DFTB3: Extension of the self-consistent-charge density-functional tight-binding method (SCC-DFTB). *J. Chem. Theory Comput.* 7 (4): 931–948.
- 7 Dewar, M.J.S., Zoebisch, E.G., Healy, E.F., and Stewart, J.J.P. (1985). Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* 107 (13): 3902–3909.
- 8 Stewart, J.J.P. (1989). Optimization of parameters for semiempirical methods. I. Method. *J. Comput. Chem.* 10 (2): 209–220.
- 9 Stewart, J.J.P. (2007). Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. *J. Mol. Model.* 13 (12): 1173–1213.
- 10 Grubmüller, H., Heymann, B., and Tavan, P. (1996). Ligand binding: molecular mechanics calculation of the Streptavidin-Biotin rupture force. *Science* (80-) 271 (5251): 997–999.
- 11 Torrie, G.M. and Valleau, J.P. (1977). Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* 23 (2): 187–199.
- 12 Laio, A. and Parrinello, M. (2002). Escaping free-energy minima. *Proc. Natl. Acad. Sci. U. S. A.* 99 (20): 12562–12566.
- 13 Kumar, S., Rosenberg, J.M., Bouzida, D. et al. (1992). THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.* 13 (8): 1011–1021.
- 14 Singh, U.C. and Kollman, P.A. (1986). A combined ab initio quantum mechanical and molecular mechanical method for carrying out simulations on complex molecular systems: Applications to the $CH_3Cl^+ Cl^-$ exchange reaction and gas phase protonation of polyethers. *J. Comput. Chem.* 7 (6): 718–730.
- 15 Field, M.J., Bash, P.A., and Karplus, M. (1990). A combined quantum mechanical and molecular mechanical potential for molecular dynamics simulations. *J. Comput. Chem.* 11 (6): 700–733.

- 16 Théry, V., Rinaldi, D., Rivail, J.-L. et al. (1994). Quantum mechanical computations on very large molecular systems: the local self-consistent field method. *J. Comput. Chem.* 15 (3): 269–282.
- 17 Warshel, A. and Levitt, M. (1976). Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J. Mol. Biol.* 103 (2): 227–249.
- 18 Pu, J., Gao, J., and Truhlar, D.G. (2004). Generalized hybrid orbital (GHO) method for combining ab initio Hartree–Fock wave functions with molecular mechanics. *J. Phys. Chem. A* 108 (4): 632–650.
- 19 Zhang, Y. (2006). Pseudobond ab initio QM/MM approach and its applications to enzyme reactions. *Theor. Chem. Acc.* 116 (1–3): 43–50.
- 20 Antes, I. and Thiel, W. (1999). Adjusted connection atoms for combined quantum mechanical and molecular mechanical methods. *J. Phys. Chem. A* 103 (46): 9290–9295.
- 21 Cao, L. and Ryde, U. (2018). On the difference between additive and subtractive QM/MM calculations. *Front. Chem.* 6: 89. 1–15.
- 22 Lence, E., van der Kamp, M.W., González-Bello, C., and Mulholland, A.J. (2018). QM/MM simulations identify the determinants of catalytic activity differences between type II dehydroquinase enzymes. *Org. Biomol. Chem.* 16 (24): 4443–4455.
- 23 Bowman, A.L., Grant, I.M., and Mulholland, A.J. (2008). QM/MM simulations predict a covalent intermediate in the hen egg white lysozyme reaction with its natural substrate. *Chem. Commun.* 37: 4425.
- 24 Ruiz-Pernía, J.J., Silla, E., Tuñón, I. et al. (2004). Hybrid QM/MM potentials of mean force with interpolated corrections. *J. Phys. Chem. B* 108 (24): 8427–8433.
- 25 Wang, X., Bakanina Kissanga, G.M., Li, E. et al. (2019). The catalytic mechanism of S -acyltransferases: acylation is triggered on by a loose transition state and deacylation is turned off by a tight transition state. *Phys. Chem. Chem. Phys.* 21 (23): 12163–12172.
- 26 Dos Santos, A.M., Cianni, L., De Vita, D. et al. (2018). Experimental study and computational modelling of cruzain cysteine protease inhibition by dipeptidyl nitriles. *Phys. Chem. Chem. Phys.* 20 (37): 24317–24328.
- 27 Mihalovits, L.M., Ferenczy, G.G., and Keserű, G.M. (2019). Catalytic mechanism and covalent inhibition of UDP-N-acetylglucosamine enolpyruvyl transferase (MurA): implications to the design of novel antibacterials. *J. Chem. Inf. Model.* 59 (12): 5161–5173.
- 28 Wei, D., Lei, B., Tang, M., and Zhan, C.G. (2012). Fundamental reaction pathway and free energy profile for inhibition of proteasome by epoxomicin. *J. Am. Chem. Soc.* 134 (25): 10436–10450.
- 29 Wei, D., Fang, L., Tang, M., and Zhan, C.G. (2013). Fundamental reaction pathway for peptide metabolism by proteasome: insights from first-principles quantum mechanical/molecular mechanical free energy calculations. *J. Phys. Chem. B* 117 (43): 13418–13434.
- 30 Wei, D., Tang, M., and Zhan, C.G. (2015). Fundamental reaction pathway and free energy profile of proteasome inhibition by syringolin A (SylA). *Org. Biomol. Chem.* 13 (24): 6857–6865.

- 31 Ramakrishnan, R., Dral, P.O., Rupp, M., and Von Lilienfeld, O.A. (2015). Big data meets quantum chemistry approximations: The Δ -machine learning approach. *J. Chem. Theory Comput.* 11 (5): 2087–2096.
- 32 Kuhn, B., Tichý, M., Wang, L. et al. (2017). Prospective evaluation of free energy calculations for the prioritization of cathepsin L inhibitors. *J. Med. Chem.* 60 (6): 2485–2497.
- 33 Zhang, H., Jiang, W., Chatterjee, P., and Luo, Y. (2019). Ranking reversible covalent drugs: from free energy perturbation to fragment docking. *J. Chem. Inf. Model.* 59 (5): 2093–2102.
- 34 Lameira, J., Bonatto, V., Cianni, L. et al. (2019). Predicting the affinity of halogenated reversible covalent inhibitors through relative binding free energy. *Phys. Chem. Chem. Phys.* 21 (44): 24723–24730.
- 35 Bonatto, V., Shamim, A., Rocho, F.D.R. et al. (2021). Predicting the relative binding affinity for reversible covalent inhibitors by free energy perturbation calculations. *J. Chem. Inf. Model.* 61 (9): 4733–4744.
- 36 Mondal, D. and Warshel, A. (2020). Exploring the mechanism of covalent inhibition: simulating the binding free energy of α -ketoamide inhibitors of the main protease of SARS-CoV-2. *Biochemistry* 59 (48): 4601–4608.
- 37 Awoonor-Williams, E. and Abu-Saleh, A.A.-A.A. (2021). Covalent and non-covalent binding free energy calculations for peptidomimetic inhibitors of SARS-CoV-2 main protease. *Phys. Chem. Chem. Phys.* 23 (11): 6746–6757.
- 38 Silva, J.R.A., Cianni, L., Araujo, D. et al. (2020). Assessment of the Cruzain cysteine protease reversible and irreversible covalent inhibition mechanism. *J. Chem. Inf. Model.* 60 (3): 1666–1677.
- 39 Santos, A.M.D., Oliveira, A.R.S., da Costa, C.H.S. et al. (2022). Assessment of reversibility for covalent cysteine protease inhibitors using quantum mechanics/molecular mechanics free energy surfaces. *J. Chem. Inf. Model.* 62 (17): 4083–4094.
- 40 da Costa, C.H.S., Bonatto, V., dos Santos, A.M. et al. (2020). Evaluating QM/MM free energy surfaces for ranking cysteine protease covalent inhibitors. *J. Chem. Inf. Model.* 60 (2): 880–889.
- 41 Mihalovits, L.M., Ferenczy, G.G., and Keserű, G.M. (2021). The role of quantum chemistry in covalent inhibitor design. *Int. J. Quantum Chem.* qua.26768.
- 42 Chudyk, E.I., Limb, M.A.L., Jones, C. et al. (2014). QM/MM simulations as an assay for carbapenemase activity in class A β -lactamases. *Chem. Commun.* 50 (94): 14736–14739.
- 43 Fritz, R.A., Alzate-Morales, J.H., Spencer, J. et al. (2018). Multiscale simulations of clavulanate inhibition identify the reactive complex in class A β -lactamases and predict the efficiency of inhibition. *Biochemistry* 57 (26): 3560–3563.
- 44 Jasim, M.H. and Rathbone, D.L. (2018). Reaction profiling of a set of acrylamide-based human tissue transglutaminase inhibitors. *J. Mol. Graph. Model.* 79: 157–165.
- 45 Arafet, K., Serrano-Aparicio, N., Lodola, A. et al. (2021). Mechanism of inhibition of SARS-CoV-2 MprobyN3peptidyl Michael acceptor explained by QM/MM

- simulations and design of new derivatives with tunable chemical reactivity. *Chem. Sci.* 12 (4): 1433–1444.
- 46 Callegari, D., Ranaghan, K.E., Woods, C.J. et al. (2018). L718Q mutant EGFR escapes covalent inhibition by stabilizing a non-reactive conformation of the lung cancer drug osimertinib. *Chem. Sci.* 9 (10): 2740–2749.
- 47 Woods, C.J., Malaisree, M., Hannongbua, S., and Mulholland, A.J. (2011). A water-swap reaction coordinate for the calculation of absolute protein–ligand binding free energies. *J. Chem. Phys.* 134 (5): 054114.
- 48 Mihalovits, L.M., Ferenczy, G.G., and Keserű, G.M. (2020). Affinity and selectivity assessment of covalent inhibitors by free energy calculations. *J. Chem. Inf. Model.* 60 (12): 6579–6594.
- 49 Mihalovits, L.M., Ferenczy, G.G., and Keserű, G.M. (2021). Mechanistic and thermodynamic characterization of oxathiazolones as potent and selective covalent immunoproteasome inhibitors. *Comput. Struct. Biotechnol. J.* 19: 4486–4496.
- 50 Yu, H.S., Gao, C., Lupyan, D. et al. (2019). Toward atomistic modeling of irreversible covalent inhibitor binding kinetics. *J. Chem. Inf. Model.* 59 (9): 3955–3967.
- 51 Awoonor-Williams, E. and Rowley, C.N. (2021). Modeling the binding and conformational energetics of a targeted covalent inhibitor to Bruton's tyrosine kinase. *J. Chem. Inf. Model.* 61 (10): 5234–5242.

Part VIII

Computing Technologies Driving Drug Discovery

24

Orion® A Cloud-Native Molecular Design Platform

Jesper Sørensen, Caitlin C. Bannan, Gaetano Calabrò, Varsha Jain, Grigory Ovanesyan, Addison Smith, She Zhang, Christopher I. Bayly, Tom A. Darden, Matthew T. Geballe, David N. LeBard, Mark McGann, Joseph B. Moon, Hari S. Muddana, Andrew Shewmaker, Jharrod LaFon, Robert W. Tolbert, A. Geoffrey Skillman, and Anthony Nicholls

OpenEye, Cadence Molecular Sciences, 9 Bisbee Ct., Santa Fe, NM 87508, USA

24.1 Introduction

The role of computation in drug discovery is continually evolving, with many advancements in theories, methodologies, and hardware. However, it is when these innovations are integrated into one that the real benefits emerge. It is commonly known that computer chips are becoming increasingly faster; from the mid-1960s, integrated circuits were, correctly, predicted to double in density every 18 months over the following two decades [1], a trend that continued well past its claim. Recently, novel chip architectures like graphics processing units (GPUs) and ARM chip technologies, like the Graviton chip series on Amazon Web Services (AWS) or Apple Silicon chips, are greatly boosting compute capabilities. These improvements in semiconductor technologies are making complex calculations tractable [2], but improvements in scientific calculations have not only come from improved hardware. Improved (scientific) algorithms, also play a pivotal role in increasing speed and accuracy, such as the development of the particle mesh Ewald (PME) method [3], Fast Fourier Transforms (FFT) [4], and Dijkstra's Algorithm [5]. It is in aggregate that these separate advances have enabled computational chemistry to become increasingly generative, predictive, and, consequently, important to drug discovery. To this end, OpenEye's cloud platform, Orion, enables the integration of elastic access to the expansive hardware resource at AWS, cutting-edge scientific algorithms and methodologies, along with tools for data analysis and visualization of results, which are then effortlessly shareable between colleagues.

Orion has enabled ligand-based virtual screening (LBVS) at an unprecedented scale of 10^{10} virtual compounds [6, 7], and this elastic resource offered by the cloud has enabled structure-based virtual screening (SBVS) at the gigascale (10^9). These virtual screening methods have been used for both the discovery of novel compounds for drug targets and for leveraging the raw data to design more intelligent

and cost-efficient virtual screening methodologies. Further advancements in this area will enable the screening of even larger chemical libraries (discussed further in Section 24.4). Orion is also a powerful tool for molecular and macromolecular simulation. We present here the integration of scientific methods and technologies such as WESTPA (Weighted Ensemble Simulation Toolkit with Parallelization and Analysis) [8] and NES (Non-Equilibrium Switching) [9] that perfectly match Orion's architecture. In the case of NES, relative binding free energy (RBFE) calculations have become fast and cost-effective, while retaining predictive accuracy (discussed further in Section 24.5). The ability to make predictions in just a few hours on large sets of virtual compounds empowers decision-making in drug design. Furthermore, in the field of small molecule crystal structure prediction (CSP), Orion has enabled us to completely re-engineer the problem to make highly accurate predictions in one to two days, rather than weeks or months [10] (discussed further in Section 24.7). The cloud-native infrastructure described in more detail in Section 24.2 enables parallelizable calculations at a scale that was unheard of just a few years ago.

In the sections that follow, we will introduce Orion (Section 24.2) and highlight some key computational methodologies developed for the platform: target data organization (Section 24.3); hit finding (Section 24.4); lead-optimization (Sections 24.5 and 24.6); drug formulation (Section 24.7).

24.2 The Platform

The Orion Platform (see Figure 24.1) integrates state-of-the-art cloud technology with scientific expertise, providing an on-demand data center customized for computer-aided drug design (CADD). With Orion, scientists can take advantage of the mega-scale computational power provided by AWS without having to be concerned about the technical details of securely managing data, orchestrating computational resources, and enabling visualization and collaboration through the

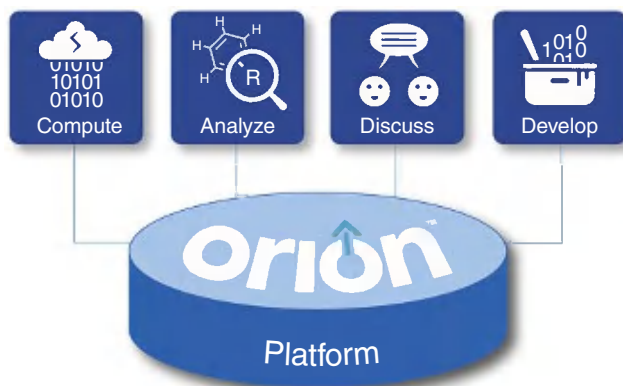


Figure 24.1 Schematic illustrating the features of the Orion platform as a compute engine with workflow development, and a place to analyze, visualize, and review data.

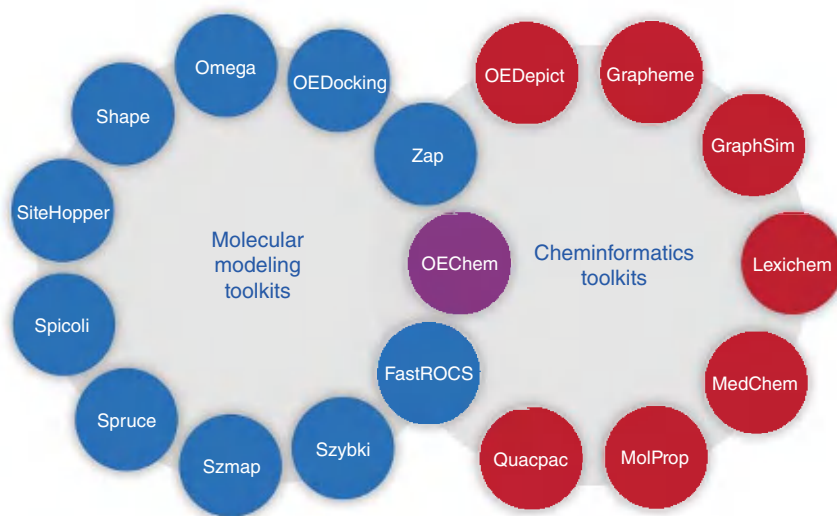


Figure 24.2 Illustration of the OpenEye toolkits with OEChem as the foundational center for both the cheminformatics and molecular modeling toolkits.

web browser, freeing scientists to concentrate on the actual problems they want to solve.

Orion has pioneered a *unified* computing and modeling environment. It provides a massive cloud-powered compute and data engine, coupled with traditional tools set within a web browser interface. This interface follows a client-server architecture that relies on native graphics acceleration to provide a low-latency, interactive 3D experience. A key advantage of building a browser interface is that it simplifies delivery, updating, and, hence, adoption.

Orion's goal to improve CADD comes with a unique set of challenges when visualizing and representing molecular systems. The first challenge for CADD software is defining the data model and parsing relevant classes of data. In Orion, this is accomplished by leveraging OpenEye's Cheminformatics and Molecular Modeling toolkits (Figure 24.2). These resources define and implement core chemistry handling, 2D rendering and depiction, 3D shape and optimization, molecular surface generation and processing, molecular grid generation and processing, and other general-purpose data handling with strongly typed data records.

Visualization challenges arise during virtual screening when discovery teams are seeking to identify candidate ligands (often referred to as hits, discussed in Section 24.4). The hit identification process can involve searching and exploring massive compound libraries, sometimes containing billions of molecules, by means of 2D (graph), 3D molecular similarity, or molecular docking. Static reports and data plots are not sufficient to drive this stage of the discovery process. The process is often interactive, sometimes requiring the ability to sketch or edit a molecular search query (in 2D or 3D) or prepare a protein–ligand system in 3D. Once a search query has been prepared and the search results have been computed, the results are often

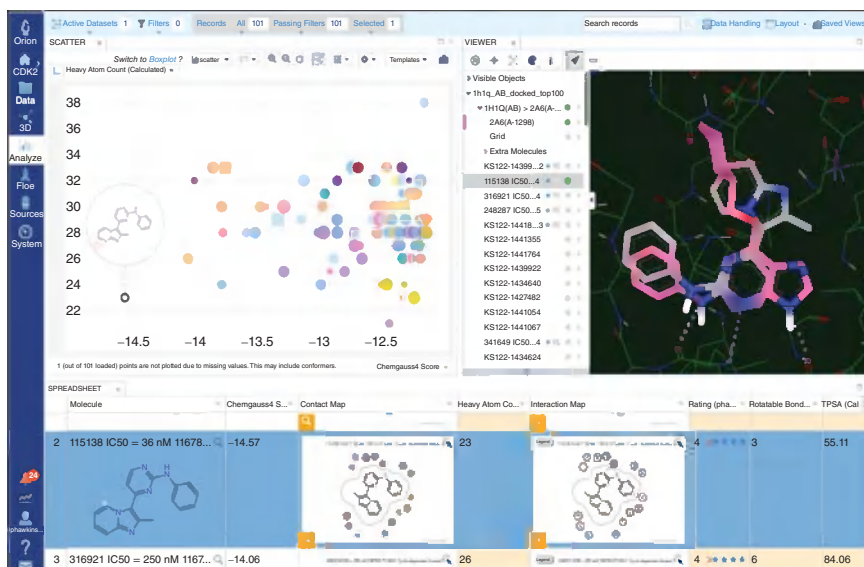


Figure 24.3 A screenshot of the Orion analysis page illustrates the interconnectivity of several ways to visualize the data: spreadsheet, 3D viewer, and plotting. In this particular instance, we can visualize results of a docking calculation. Displayed are the interactions between the hit molecule and the protein environment.

triaged to classify or prioritize the hits and identify compounds that warrant further exploration. Molecules resulting from such a search are typically augmented with useful information to help inform the triage process; they may also be clustered and ranked. For example, scores from the well-known ROCS approach [11] are used to rank molecules based on the probability that they share relevant (biological) properties with the query molecule. The example in Figure 24.3 shows how Orion visualizes the search results from molecular docking in a way that combines the 3D superposition of query and result conformations, the perceived interactions between protein and candidate ligand, and the other data associated with the query and results. Additionally, the analyze page allows the user to easily add additional chemical properties not already calculated and stored in the dataset. The raw datasets are easily shared with members of the project in Orion. More importantly, once a user has filtered and reached a stage at which they want to share their observations, Orion “discussion boards” can be created, which save the state of desired views along with any comments project members wish to save. Boards can be updated in sequence to show the progression of a project. As storage on AWS is like having one giant hard disk, sharing is simple and avoids any need to send large packets of data around company email servers. In addition, researcher access to identical views means it is easy for a scientist to convey their observations without their interpretations being lost along the way.

The preceding description of visualization and analysis challenges is not exhaustive, but hopefully sufficient to convey the complexity and technical depth of

the software required to solve such a problem. Computational resources are accessed in Orion using “Flow-Based” programming [12]. This programming paradigm is well suited to elastic and heterogenous cloud resources and can allow nonprogrammers to modify workflows by adding or replacing high-level reusable components. These components are referred to as “Cubes.” When multiple cubes are connected, these computational workflows in Orion are called “Floes” (Figure 24.4).

Floes are orchestrated in Orion with a proprietary scheduler and runtime system that is responsible for scaling up the appropriate type and quantity of compute instances. It fairly shares the computing hardware between Floes, efficiently packs Cubes onto large computers, and minimizes costs by taking advantage of less expensive resources and scaling them down when they are not required [13–15]. *Parallel* Cube allocations operate on independent items of work. Such allocations balance work among themselves via a “work stealing” algorithm, and if a process fails, they automatically retry, while avoiding issues of unnecessary scale-up in response to such failure [16]. As a consequence, floe developers are able to bring parallelism to Floes with minimal effort.

When a relatively small number of data records are being processed or viewed in the Orion UI, they are stored in a relational database management system (RDBMS). However, as the number of data records reaches millions or higher, the spreadsheet-like UIs cannot take advantage of all the data and a RDBMS is pushed beyond its limits. For these large systems, Orion provides a different data model, called *Shard Collections*, that is purpose-built for massively parallel IO. Shard Collections are made up of file objects (i.e. shards) and are a thin wrapper over Amazon S3. This massively parallel IO allows, for instance, searches over 10^{10} or more virtual compounds [6, 7], docking of compounds at the gigascale (10^9), and large-scale quantum chemistry calculations of torsions for torsion strain analyses [17, 18].

Sets of cubes and floes, form suites and modules (Figure 24.5) that leverage our toolkits as well as third-party Open-Source software for statistics, quantum mechanics (QM), machine learning, bioinformatics, and molecular dynamics simulations [19–26]. Customers can add new Floes that either leverage provided resources or can add new cubes that provide access to unique or even proprietary methods, along with any required software or hardware.

In the next sections, we will describe how we leverage and utilize Orion for various aspects of the drug discovery cycle.

24.3 Target Preparation and Structural Data Organization

Solving the structure of proteins and nucleic acids has revolutionized drug discovery [27, 28] and created the field of structure-based drug design [29]. Structure has allowed for the visualization and understanding of how candidate drug molecules bind, enabling novel designs, while also assisting in the validation of modes of

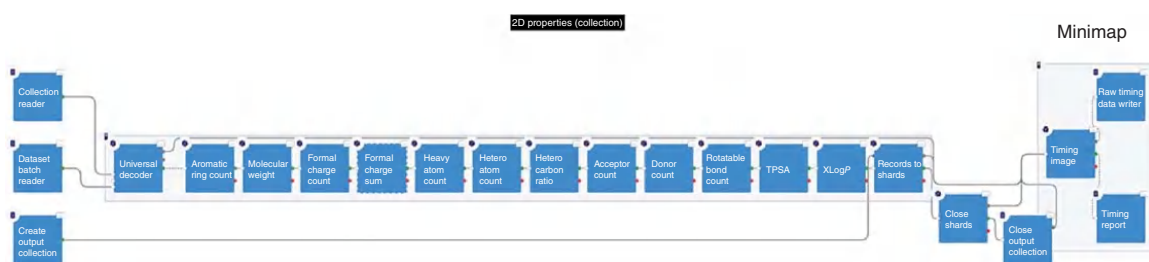


Figure 24.4 Screenshot of a workflow in Orion, with lines connecting different input/output ports on the cubes.

The Orion® platform



Figure 24.5 Illustration of the Suite and Modules that are delivered on top of the Orion platform. Orion is fully capable of using other third party software and in-house tools to take advantage of all the features the platform provides.

action. Experimental structures come primarily from X-ray crystallography, but in recent years, the advancement of cryo-electron microscopy [30–32] has further improved our understanding of how biomolecules interact in more complex environments [33, 34]. Lastly, cryo-electron tomography holds even greater potential to understand biomolecules in a cellular context [35]. These experimental techniques have been complemented by modeling efforts such as homology modeling [36–38], and more recently machine learning approaches to structure prediction [39].

The quantity of protein structural data now available and the accelerating rate of structure determination, have presented challenges in the preparation, storage, and organization of protein structural data. The Protein Data Bank (PDB) [40] stores structural data with annotations, cross references to other databases, and statistics of a given structure in the context of the database [40, 41]. However, such data is not prepared for typical modeling applications. In Orion protein structures are stored and queried using the MacroMolecular Data Service (MMDS). Structures are stored on a per-target basis, pre-prepared for modeling tasks, and are typically aligned onto a selected reference structure. This provides project teams with easy access to how a single molecule binds to the target, and also how that compares to other relevant drug candidates (Figure 24.6). Furthermore, as targets are organized by family (*vide infra*) and have a common superposition reference, it is possible to compare binding modes of different ligands into the same or related protein targets. This is particularly relevant for off-target effects in families that have highly similar folds and binding site structure, e.g. kinases [43–45].

The structure preparation process uses SPRUCE [46], an OpenEye protein preparation tool. It provides not only structures prepared for modeling applications but also a quality assessment of the structural data using Iridium [42]. Iridium

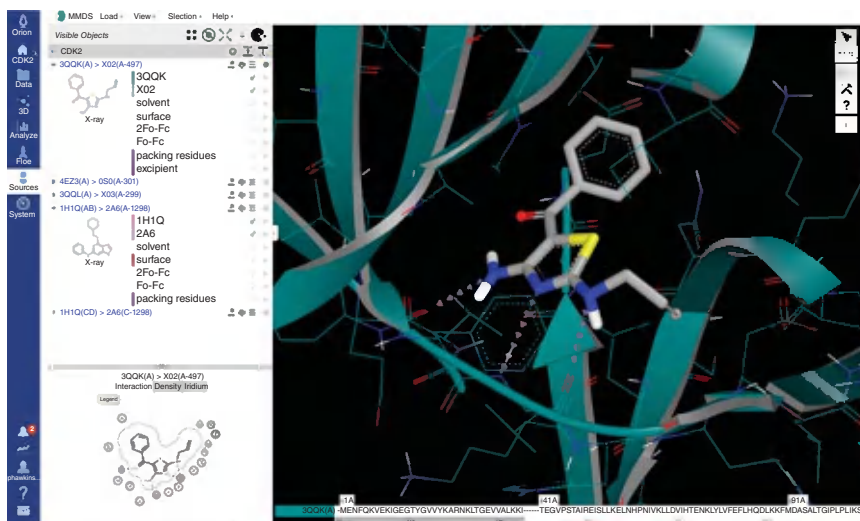


Figure 24.6 Screenshot from MMDS, showing a protein binding site, with a ligand bound to the cyclin-dependent protein kinase 2 target (PDB ID 3QOK) with hydrogen bonds shown in 3D. On the left, a list of additional structures prepared in the same reference frame (view) for this target are shown. At the bottom left, depictions illustrate the protein–ligand interactions, with tabs to show the electron density overlay on the ligand and the Iridium protein structure quality classification. Source: Warren et al. [42]/with permission from Elsevier.

highlights potential issues with the experimental data that warrant investigation and, in some cases, correction before use. As part of the Iridium classification, structures are flagged for crystal packing artifacts that potentially influence a molecule’s binding modes or binding site configurations. A recent example where corrective measures had to be taken is the EG5 target provided in a well-known FEP benchmark by Merck KGaA [47]. Additionally, an Orion Floe provides depictions that enable a quick overview of the binding mode and structural data (Figure 24.7). Inspired by work at Pfizer in their attempt to organize their internal and relevant public structural data, SPRUCE produces what is termed a design unit (DU) [48]. A DU, in addition to being a prepared structure, organizes components by protein, ligand, cofactors, solvent, excipients, and more. Thus, the DU data structure makes the retrieval of each component or set of components very tractable for common modeling tasks.

With structural data, the target definition itself is usually established, although the location of the binding site may differ depending on how a target is being prosecuted, i.e. whether interest is in an orthosteric or allosteric site [29]. Well studied targets such as kinases have an established convention for how they are organized into sub-families. However, not all protein families are organized by evolutionary relationships, but are instead organized by therapeutic area. MMDS does not impose a hierarchy on protein structural data; the interface is flexible

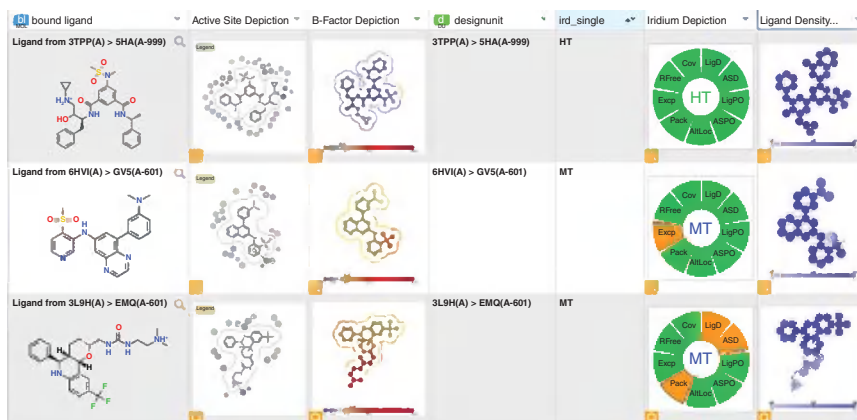


Figure 24.7 Screenshot of Orion's analyze page of a SPRUCE-prepared dataset, where the design unit title is shown, along with the bound small molecule (if relevant), as well as other depictions similar to those from MMDS.

enough to build any relational tree, the only requirement being whether two targets are considered superposable inside a family branch or “node.” In preparing the entire PDB, we chose to leverage the resource from “Guide to Pharmacology” [49], which has established a tree system for most targets of relevance to the pharmaceutical industry. Beyond those target descriptions, we adopted a flat structure for all remaining targets that SPRUCE processed. There are multiple alternative choices we could have adopted, and might in the future, based on function or evolutionary relationship, like Enzyme Classification (EC) [50], GPCR-db [51], PANTHER [52], or Superfamily [53]. MMDS also allows for multiple root nodes with structure duplication between trees, but this has been beyond our initial goals.

It is possible for users to introduce their own hierarchy alongside our own, geared more directly to their working project teams and with their proprietary structures. Using “Guide to Pharmacology,” we have mapped targets using IDs from the UniProtKB [54] for the human forms to map to the relevant PDB entries, and we used sequence alignment to incorporate additional species into a target if they had a high enough sequence similarity to the human variant. One of the main challenges that needed to be solved for this work was mapping PDB entries to targets, particularly around multi-protein entries in UniProtKB. As an example, the PDB entries 4E92, 2M3Z, and 7LRY all map to UniProtKB entry P12497; however, the structures are of three different targets, i.e. capsid proteins, nucleocapsid proteins, and reverse transcriptase. This is because HIV-1 is a multi-protein entry in UniProtKB, and it was necessary to parse the feature information along with the structure-to-property data from UniProtKB to correctly separate the targets and correctly map their structures based on the sequence. Additionally, it is becoming more common that a PDB entry contains chains from multiple proteins in larger

assemblies, which is also a complicating factor when trying to correctly map the structures without doing redundant and costly structure preparation.

At the time of writing, MMDS contains 103 667 PDB structures and 1274 AlphaFold2 models from DeepMind and EMBL-EBI [55] (*vide infra*). This expands into around 195 490 design units, primarily from generating the proper biological forms from asymmetric units and enumerating alternate locations as relevant to drug candidate binding modes. These cover around 16 250 different targets. Ideally, we would be able to prepare all 194 259 (as of Aug. 2022) structures in the PDB, however, to add a target to MMDS, there needs to be a known binding site. As such, we devised an algorithm that aims to detect common binding sites in protein structures of the same target. This is how we picked reference structures for each of the targets we incorporated. For some targets, a reference structure could not be automatically detected. Even so, there are several targets in the PDB where no small or peptidic molecule is bound to the target (apo structures), so we could not designate a common binding site. This will be augmented with pocket detection algorithms in the future. There were also targets where small molecules were bound but not to a consistent binding site. And there are structures, which are clearly not protein drug targets but, i.e. assemblies of peptide aggregates (e.g. PDB ID 1YJP [56]).

For structures from AlphaFold to be incorporated into our preparation pipeline, we needed a PDB reference structure with a similar structure and with an accessible binding pocket. This was problematic for larger structures, because of AlphaFold's limit of 2700 or 1280 residues per structure (depending on the source and species). Some of this limited the utility of AlphaFold in this iteration, but pocket detection algorithms could improve this in the future.

Our primary objective was to make this data available in Orion and MMDS for project teams. However, it also became valuable as a database for our SiteHopper [57] tool. SiteHopper is a search tool built using OpenEye's ROCS technology [11], but instead of comparing (bound) ligand conformations, it compares protein binding sites. Searching a database of protein binding pockets with a given binding site as a query allows a researcher to find binding sites in the database that look similar, which can be useful for predicting potential off-target effects. Orion hosts several SiteHopper databases: one containing the 195 490 protein–ligand binding sites, mentioned above, but we have also employed pocket detection tools on this dataset, including an in-house method, OEPocket, and the published F-pocket [58], and in aggregate have generated around 2.2M potential binding sites (excluding the known sites) that can now be searched.

The combination of an automated biomolecule preparation tool like SPRUCE and the cloud resources in Orion, improves the rigor and reliability of the structure preparation processes and makes the prepared structures readily accessible. Including annotations and depictions makes the results easily digestible and actionable. The prepared protein structures are suitable for a variety of structure-based calculations, including binding mode evaluation and docking, and new structures can easily be prepared and incorporated as they become available (from either public, or internal sources), accelerating the pace of structure-based drug discovery.

24.4 Virtual Screening

Virtual screening is a computational form of hit identification that occurs early in the drug discovery process [59]. It is expedient and economical to evaluate the putative activity of potentially active compounds *in silico* (via a computer calculation) rather than an *in vitro* assay, provided the computation required is fast enough to allow for rapid and inexpensive evaluation of often millions and sometimes billions of molecules [6, 60]. Virtual screening algorithms generally aim to quickly score or rank molecules using some method that is a proxy for their probability of activity in an *in vitro* assay rather than endeavoring to quantitatively predict activity or binding, as is often the goal later in the lead optimization stages of drug discovery (see Section 24.5). Virtual screening methods are often evaluated with metrics such as the Area Under the receiver operating characteristic (ROC) curve (AUC) or enrichment factor (EF), both of which focus on the retrieval of items classified as active from inactive compounds [61, 62].

Virtual screening approaches can be broadly categorized into two camps based on the primary source of knowledge used to score the screened molecules. In the case of LBVS that source is one or more drug-like molecules, most often molecules known to be active, whether against the desired target in an *in vitro* assay, efficacious in an *in vivo* experiment, or identified more directly from biological research. These known actives then provide the “knowledge base” that LBVS algorithms use to compare molecules and assess their potential for activity. Because of the utility of direct comparison, many LBVS methods are similarity-based. Common and well-known tools from cheminformatics, such as 2D graph-based similarity comparison or using molecular fingerprints [63], can be performant and effective LBVS techniques [64–66]. Similarity can also be used for 3D molecular comparisons, which then require the conformational space available to each molecule to be assessed and sampled. If conformer space can be sampled efficiently, finding a molecule that can take a similar shape and chemical features to a known active compound is an effective virtual screening approach. Shape and chemical feature comparison is central to the widely used ROCS application [11] as well as other virtual screening tools. Other virtual screening approaches may choose or combine particular features from 3D alignments of multiple active molecules, attempting to build a “pharmacophore” of the presence and spatial arrangement of critical characteristics to which other molecules can be compared.

The counterpart to LBVS is (SBVS), which relies either solely or predominantly on knowledge of the protein structure to which active molecules must bind (discussed above in Section 24.3). SBVS is most often carried out using docking, where a molecule is placed into the protein binding site and is evaluated and/or optimized via a scoring function. Because these scoring function evaluations need to be relatively fast and efficient, they often include radically simplified estimators of the complex physics that underly protein–ligand binding, e.g. combine approximate electrostatics and van der Waals treatment with rough corrections for (or complete exclusion of) effects such as entropy, desolvation, or protein flexibility. However, the result of a docking calculation can also provide a hypothesis of the binding mode of the

molecule in question, which itself can provide a jumpstart to hit-to-lead or lead optimization. Additionally, as the docking evaluation is based solely on the complementarity of the molecule in question to both the shape and interactions of the binding site, there is no intrinsic bias toward discovering molecules that are similar to known actives, potentially leading to the discovery of more diverse hits than through LBVS approaches.

Part of the recent growth of virtual screening literature is due to the increase in the size of commercial libraries of available compounds [60]. While virtual libraries of compounds could always be enumerated through generative or reaction-based mechanisms, the synthetic viability of such molecules was often a challenge, either due to reaction planning or the sourcing of starting materials. Through more recent advances in reagent management, availability, and high-throughput reliable reactions, many commercial vendors now offer searchable libraries of make-on-demand compounds well in excess of a billion molecules, with synthetic success rates anecdotally reported above 85%. Large pharma companies can often employ similar efforts to marshal their available compounds and reaction knowledge to build very large virtual libraries with good rates of synthetic success [6, 7]. Nevertheless, the widespread availability, price, success rates, and speed of delivery of these commercially available libraries level the playing field for small biotechs and pharma startups. For relatively little investment, new projects and targets can be jump-started through a large-scale virtual screen covering billions of molecules. However, the compute costs of screening a library scale directly with its size, so the advent of very large libraries that are worth screening has brought with it a commensurate increase in the computational resources needed to execute these virtual screens. Happily, the advent of cloud computing has provided a computing resource that is well-suited to many of the demands of large-scale virtual screening. Cloud computing platforms provide ample storage to hold these large libraries, even in stereo-chemically expanded (when necessary) and 3D conformationally sampled forms; this allows a molecular database to be prepared once and then searched many times, an important efficiency when the database is so large that the computational cost of preparing the database is itself significant. In Orion, OpenEye provides access to several enumerated and conformationally sampled libraries from a variety of chemical vendors. In total, Orion currently provides access to around eight billion molecules prepared for large-scale virtual screening efforts.

The independent nature of each compound's evaluation makes most virtual screening approaches amenable to divide-and-conquer approaches, and the availability of renting large quantities of relatively inexpensive compute resources from cloud computing platforms allows for very wide parallelization of virtual screens across hundreds, thousands, or tens of thousands of compute nodes. Finally, the dynamic nature of cloud computing resources allows for compute nodes to be used only when calculations require resources and then scaled down and turned off when the calculation is complete, encouraging the widest scaling of virtual screens to return results in the shortest time relative to the total amount of compute used. These features combine to make the capability of large-scale virtual screens

of billions of purchasable compounds more cost-efficient, impactful, and widely available than ever before.

The cost of running a virtual screen in the cloud can be estimated by the following equation, where $\text{cost}_{\text{SBVS}}$ is the total cost of the screen in US Dollars (\$), N_{mol} is the total number of molecules to dock, r_{CPU} is the cost of renting a CPU in the cloud in \$/h, t_{dock} is the mean time to process a molecule in seconds, and t_{io} is the meantime a molecule spends in I/O operations in seconds.

$$\text{cost}_{\text{SBVS}} = N_{\text{mol}} * r_{\text{CPU}} * \frac{(t_{\text{dock}} + t_{\text{io}})}{60 \text{ s} * 60 \text{ min}}.$$

This yields a cost of around \$8300 per billion molecules for a structure-based virtual screen assuming $r_{\text{CPU}} = \$0.03/\text{h}$, $t_{\text{dock}} = 1 \text{ sec}$ and $t_{\text{io}} = 0 \text{ sec}$. As of this writing $r_{\text{CPU}} = \$0.03/\text{h}$ is typical for spot instances of AWS c5 CPUs, which are standard modern CPUs. A t_{dock} of one second requires an efficient docking program, in our case FRED or HYBRID [67, 68], but is entirely feasible, particularly for small binding sites. The overhead from moving the molecules from their database to the CPU running the docking program and back, t_{io} , can be minimized to the extent that it is negligible compared to the docking cost (t_{dock}), but it is a challenging problem that is often overlooked. A typical cloud-based SBVS run on billions of molecules will typically utilize tens of thousands of CPUs in parallel, during which all processors must be kept saturated with molecules to dock, and then the results must be stored. While recruiting tens of thousands of CPUs in the cloud is clearly doable, the Orion platform does this cost-efficiently, feeding instances that are running and rapidly shutting down unused instances. Furthermore, in Orion's orchestration layer, as described in Section 24.2, we have built-in tolerance to instances failing or being taken away in the AWS spot market; such pieces of work are retried in a manner that is invisible to the user, without loss of work even at these extreme scales.

To validate our large-scale SBVS approach, the docking of 1.4 billion Enamine Real molecules to HSP90 was performed in Orion using the Gigadock Floe, which cost approximately \$14K total, or around \$10K per billion molecules docked. The top scoring 120 molecules from this run were ordered and assayed. About one-third of the molecules assayed showed activity. The top-scoring molecule out of the entire 1.4 billion docked molecules was a 4 μm inhibitor shown in Figure 24.8. As can be seen, the hit molecule has a different scaffold and binding mode than the original ligand the protein was crystallized with, showcasing the strength of SBVS to find novel leads.

With commercial libraries on the scale of tens of billions of compounds, the cost of performing a full SBVS on of these libraries is significant. The Enamine REAL collection is around eight billion compounds as of this writing, and a SBVS on this library with Orion's Gigadock is likely to cost between \$50K and \$100K. This is a significant sum, although not outside the budget of many serious drug discovery efforts. This is particularly true when compared to the cost of robotic high throughput screening

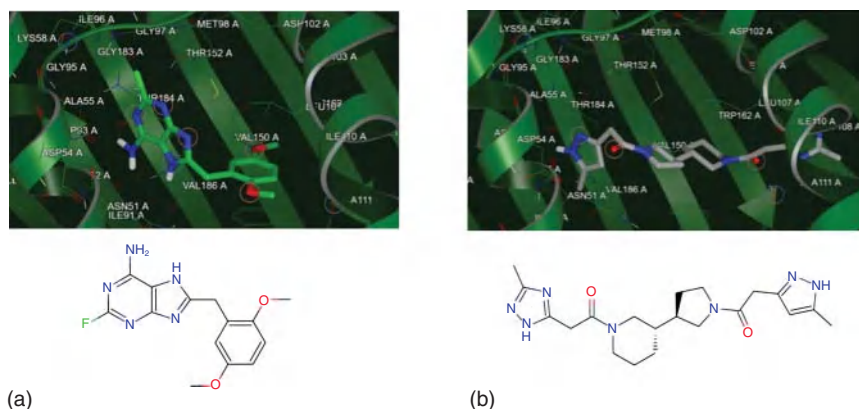


Figure 24.8 (a) Co-crystal ligand, a 53 μm inhibitor, bound to the HSP90 active site. (b) Top scoring database compound, a 4 μm inhibitor, docked to HSP90.

(HTS), which has significantly higher costs, often in the millions of dollars per million compounds. Nevertheless, SBVS costs are large when docking billions, which is a reason to investigate further optimizations to reduce costs.

One method of reducing the costs of large SBVS runs is to create a machine learning model that predicts the score of compounds much more rapidly than the docking algorithm. These models are trained per target and are generally of the following form:

1. A small fraction of the molecules is docked to the target using the normal docking algorithm.
2. The structure of these molecules, usually encoded as fingerprints, is fed to the machine learning algorithm as training data along with the docking scores.
3. The machine learning algorithm predicts the docking scores of the entire set of billions of molecules.
4. A fraction of the molecules with the highest predicted docking scores are docked.
5. The top-scoring molecules from step #4 are output to a hit list.

There are many variations of the general procedure outlined above, e.g. clever ways to pick the molecules that go into the training set, or using go/no-go classification rather than a regression model to determine which molecules progress to full docking. The overall theme, however, remains the same: to use a fingerprint-based model to predict which molecules are most likely to have good scores and then perform actual docking on those molecules [65, 69–72].

An alternative approach is to create a model to predict the *binding mode* of high-scoring compounds, rather than the docking score itself, and then to calculate such a score (which, for a single pose, is very fast). This is the approach taken by Gigadock Warp, a recent Orion Floe (see Figure 24.9). It takes the following form:

1. Perform an initial complete docking of 2% of the molecules.
2. Select the top 50 highest-scoring molecule poses.

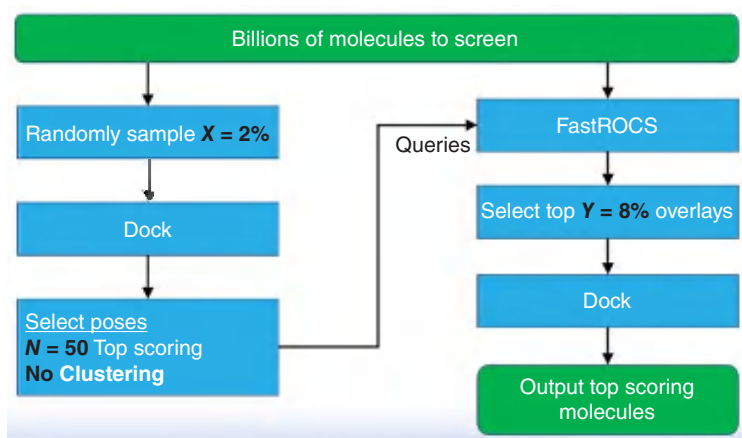


Figure 24.9 Gigadock Warp approximate docking algorithm.

3. Use these poses to search the entire set of molecules with FastROCS for those that have the highest 3D similarity to the pose queries.
4. The best molecules from step #3 are then docked.
5. The top-scoring molecules from step #4 are output to a hit list.

Gigadock Warp searched the same 1.4 billion Enamine molecules and HSP90 target described above and produced a top ten thousand hitlist with 70% of the same molecules in the top 10 000 as from a full Gigadock at 1/8th the time and cost. This cost savings, while retaining good hit performance, is important as the size of virtual (but chemically accessible) molecule libraries continually increases. Further research into protocols that leverage machine learning is ongoing as we keep in mind that the estimated size of chemical space is much larger and perhaps as large as 10^{60} molecules [73].

24.5 Predicting Small-Molecule Binding Affinity

With a small-molecule lead in hand from virtual screening, the next stage in drug discovery is lead optimization. There is a vast array of computational chemistry methods that can be brought to bear on structure-based lead optimization; here we will restrict our focus to the role of molecular dynamics simulations in the context of Orion. We are using the paradigm depicted in Figure 24.10, where these comparatively expensive simulations are placed downstream of methods that generate a large set of candidate ligands, posed in the receptor site, that various refinement, scoring, and filtering steps winnow down to a starting set of compounds of particular interest. The ligand binding model will have been based on, at best, a minimized structure of the bound protein/ligand complex. At this point, we propose two distinct approaches involving biosimulations: a relatively short MD run, optionally followed by a more expensive RBFE calculation seeking a better prediction of binding affinity.

MD in structure-based lead optimization



Figure 24.10 Role of MD in structure-based lead optimization. Based on previous iterations, a large list of candidate ligands for the next round of synthesis is generated, posed in the receptor, and triaged. After further filtering and clustering (based in part on force field energy minimizations), another round of triage further reduces the list. A quick MD screening allows further winnowing to a set of good candidates worth more expensive free energy calculations.

The initial short MD run accomplishes several goals. Even a short-timescale simulation is a good test of the stability of the bound ligand pose generated by the upstream docking/minimization approaches. Adding explicit water solvation and configurational sampling at physiological temperature allows the initial protein–ligand complex to relax away from the highly localized bias of a minimized docked pose. Favorable-looking protein–ligand interactions from these earlier stages can prove to be unstable in competition with solvent interactions or ligand strain, and even a brief simulation can be enough to show an initially promising synthesis candidate to not be worth carrying forward. This stage also allows for approximate affinity scoring methods such as MMPBSA [74, 75] to be applied as ensemble averages to the short trajectory. While often less accurate than explicit RBFE scores, if such cheaper methods yield a good enough prediction, they can at least be used to remove poorer-ranked compounds. Or if they score really well, the RBFE stage might not be necessary at all. A final benefit of an initial short MD calculation is that this data can be reused in an RBFE method such as NES [9], as described in detail below.

The workflow for this stage, called Short Trajectory MD (STMD), is shown in Figure 24.11. While fairly typical in form [76], each stage consists of multiple cubes targeting different AWS instance types, in series or in parallel, suited to the particular computational needs of the task at hand, minimizing turnaround time for the user. The inputs are Orion datasets from the upstream workflows: an MD-ready prepared protein receptor, typically from the Spruce Floe (see Section 24.3), and a set of posed ligands, typically from the energy minimization of docked structures generated from, e.g. SMILES strings. The bound complex with each ligand is formed and solvated in a periodic box of explicit water and counterions to form the “flask,” the whole system for simulation. Each flask is submitted to the “Run MD” stage in parallel; if multiple bound poses for the same ligand are supplied or if multiple starts

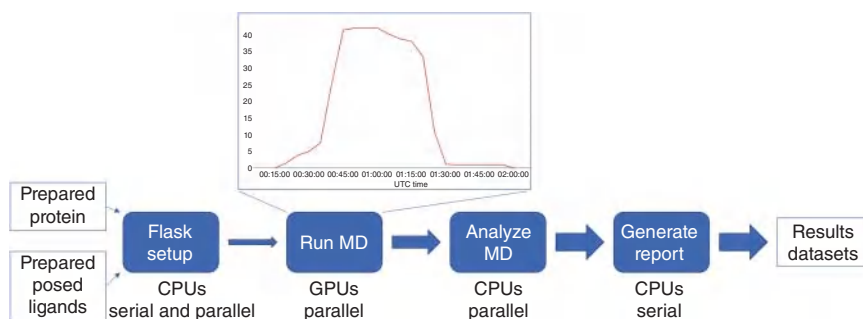


Figure 24.11 Stages in Orion's Short Trajectory MD Workflow. Each stage contains a number of "cubes" as described above. The computationally demanding stage "run MD" is done in parallel on GPUs; the other stages are done on CPUs with a mixture of serial and parallel cubes. The inset above the "run MD" stage shows the time course of the Orion scheduler's recruitment of GPUs for a set of 42 ligands. Data records accumulate results as they progress through each stage, ultimately being written into high-content results datasets. A report is generated to summarize key results. All stages are run in Orion, including visual ingestion of such results.

are desired for the same bound pose as recommended by Bhati et al. [77], they are also run in parallel. The entire workflow is completed within just a few hours.

The results are analyzed by ligand, grouping together multiple starts or multiple poses of the same ligand, and clustering by ligand configuration. Clusters are scored by ensemble MMPBSA ($\langle \text{MMPBSA} \rangle$) [74, 75] and BintScore, an internally developed knowledge-based score monitoring protein–ligand interactions, assessing how close the ligand stayed to the starting pose. These scores can be used directly to rank ligands for synthesis or to assess the stability of the initial pose. If there are already synthesized ligands with measured activities, these can be correlated with the scores from STMD to see if an adequately predictive model can be established. The variable and target-dependent accuracy of these endpoint scores means these models will not be useful for all targets, but they can be useful for some. Even when the endpoint scores do not give accurate models, STMD still serves a valuable role in validating the initial pose. If the pose has substantively changed after even a short simulation, this might suggest not pursuing that compound. At this point, a subset of ligands can be selected to carry forward to the next stage, RBEF with NES.

Though NES [9] is not as mature a method for RBEF as the well-established free energy perturbation (FEP) [78] or thermodynamic integration (TI) [79] methods, NES is much more efficiently parallelizable and, as such, a natural fit for Orion. As shown in Figure 24.12, all three methods employ the same basic elements in alchemically changing one ligand into another: the starting ligand (ligand A) is gradually morphed into the final ligand (ligand B) along a transformation variable λ , measuring a key energy difference along the way. That energy difference needs to reflect the average behavior of the ligand at each step in the simulation. FEP and TI work by running a brief equilibrium simulation, typically a few ns, to collect statistics for the ensemble average, over a typical range of 20 λ values or 20 windows. Thus, a typical total simulation time is 50–100 ns per

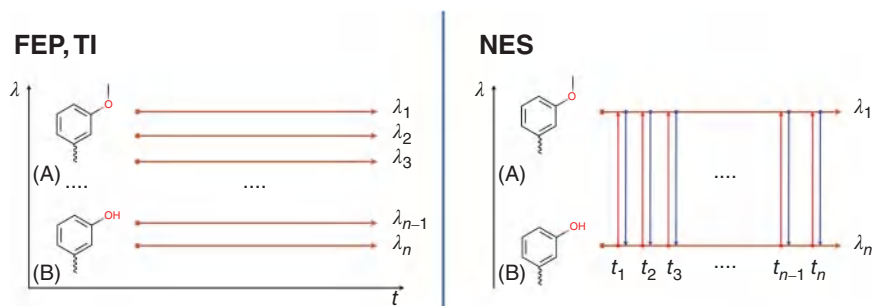


Figure 24.12 Conceptual comparison of Non-Equilibrium Switching (NES) with Free Energy Perturbation (FEP) and Thermodynamic Integration (TI) methods in how they carry out an alchemical transformation or edge. On the left, FEP and TI step through n values of λ , carrying out short equilibrium runs at each value. On the right, NES runs n independent short simulations quickly and continuously changing λ ; the runs are carried out in both directions, i.e. starting from each ligand endpoint, and then the results are analyzed together. The starting point for each of the n simulations in NES is drawn from a prior equilibrium simulation of the ligand.

transformation, in principle parallelizable over each λ value, i.e. around 20-fold. In contrast, NES carries out a fast and continuous transformation of one ligand into the other (called a “switch”) along λ , typically in 0.1 ns or less (we currently use 0.05 ns following Ref. [9]), evaluating the overall nonequilibrium work done along the way. The average behavior is obtained by examining the distribution of many such switches; we currently use 80 following Ref. [9]. A typical total simulation time is 4–8 ns per switch, straightforwardly parallelizable over the 80 separate transformations. Therefore, NES is both more efficient and more parallelizable than FEP or TI. The catch is that the starting point for each switch must be drawn from an equilibrium sampling of the ligand, so a prior simulation must be done for each ligand. This is why prior STMD calculations can be reused to give equilibrium starting points for the NES switches.

As with the earlier STMD stage, the workflow for the NES stage, shown in Figure 24.13, follows a Setup→Compute→Analysis→Output paradigm, using serial or parallel cubes as appropriate. Given multiple alchemical transformations within an NES run (often dozens), each having many independent parallel switches within each edge (e.g. 320 per edge, given 80 switches starting from each ligand, for both the bound and unbound ligands), there are thousands of tasks to run in parallel. To meet this demand, the Orion scheduler can routinely recruit thousands of GPUs for this stage (cube); with the short runtime of each of these tasks, the entire NES calculation can be finished in a few hours, with final results immediately available to the scientist.

In contrast to the light, fast STMD stage, which gives both qualitative insight and quantitative models to the discovery process, the RBFE calculation is focused on producing a quantitative model for affinity. The free energy estimates for the various ligand transformations are used to generate a direct estimate of the ligand binding free energies, with one or more measured affinities used as a reference to

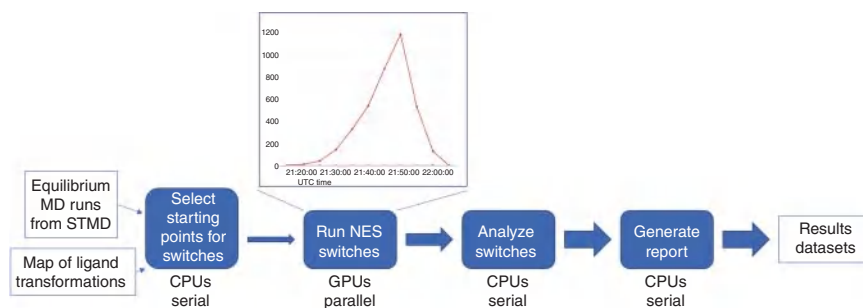


Figure 24.13 Stages in Orion's Non-Equilibrium Switching (NES) workflow. Each stage contains a number of "cubes" (not shown). The computationally demanding stage "Run NES Switches" is done in parallel on GPUs; the other stages are done by serial cubes on CPUs. The inset above the "Run NES Switches" stage shows the time course of the Orion scheduler's recruitment of GPUs for several thousand NES switches. The result datasets are written out, and several reports are generated to summarize key results in Orion.

establish an absolute scale. Ideally, these direct predictions would exhibit a unit slope with measured affinities, but frequently a good correlation is found but with a slope deviating from unity. In this situation, a robust linear model (as used with <MMPBSA> and BintScore) can be used to good effect, given enough experimental measurements. Figure 24.14 shows the results of the affinity models from both NES and STMD stages for 10 protein–ligand datasets [80], both in terms of rank order correlation (Kendall's τ) and mean absolute error (MAE). The primary value of the models lies in the correlation, i.e. rank-ordering the ligands by affinity to prioritize synthesis. While NES results show a clear advantage overall, there are some targets where either <MMPBSA> or BintScore show equivalent or better performance. With good correlations, lower MAEs should be expected, and that holds generally true for <MMPBSA> and BintScore, for which the models are robust linear regressions with experimental data. Interestingly, for the direct predictions of binding ΔG with NES, several datasets (PTP1B, p38, Thrmb, and MCL1) show much worse MAEs than the endpoint models even though the correlations are comparable or better; this is due to deviations from unit slope for the direct predictions from NES. Invariably, linear *models* of affinity based on the NES results improve the MAE dramatically in these cases.

Overall, the power and flexibility of Orion make it possible to integrate the use of biosimulations as a routine part of structure-based lead optimization, starting with a candidate set of ligands already triaged using static energy calculations or energy minimization approaches. Relatively fast STMD simulations allow for an initial assessment of a set of ligands in a few hours, allowing simpler methods the opportunity to produce a useful model of affinity, while at the same time performing the MD equilibrium prework necessary to set up the computationally intensive second stage of NES. The massive parallelism of NES in Orion also makes it possible to complete the entire RBFE calculation in just a few hours, with a high likelihood of generating good models for affinity to prioritize ligand synthesis. Looking ahead, in addition to steady ongoing refinements in the physics-based methods described

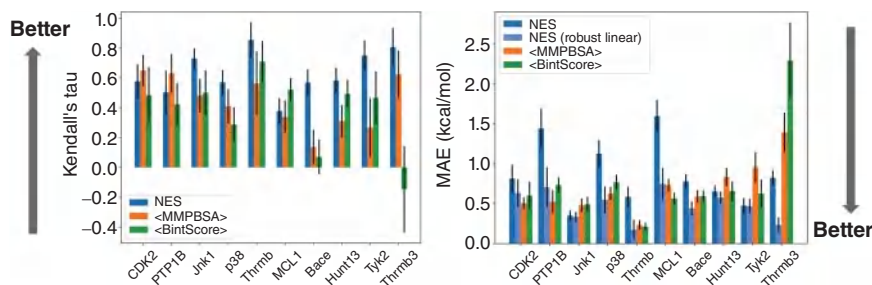


Figure 24.14 Kendall's tau correlations and mean absolute error (MAE) shown for NES and end-point analyses of ensemble MMPBSA and BintScore for 10 targets, all based on the same short (6 ns) MD trajectory. The standard error from bootstrapping is shown as a black line. On the right is the MAE for the robust linear models based on endpoint methods <MMPBSA> and BintScore. For NES, the MAEs for both the direct prediction of ΔG from NES (dark blue) as well as the robust linear model of ΔG from NES (light blue) are shown; the two differ when the linear model deviates from unit slope.

here, we also see complementarity with machine-learning methods in improving affinity predictions.

24.6 ADMET Prediction and Permeability in Drug Discovery

With one or several lead series in hand that bind to the desired protein target with high affinity and specificity, predicting absorption, distribution, metabolism, excretion, and toxicity (ADMET) liabilities can assist in prioritizing among series, or even assisting in rationally designing compounds to minimize ADMET liabilities. Pharmacokinetic (PK) properties can be cast into an ADMET profile, which describes the ability of a drug-like molecule to perform its intended pharmacological function [81]. An early report on attrition rates suggested that 39% of all new chemical entities at that time were withdrawn from clinical trials due to PK liabilities [82]. Considering the costs associated with clinical trial failures, finding a method to optimize ADMET profiles to prevent attrition remains a challenge for the industry, despite recent improvements in the area of bioavailability [81].

ADMET profiles are influenced by physiochemical properties, primarily permeability [83], which quantifies the ability of a drug-like molecule to traverse cellular membranes. The mathematical expression for permeability, P_m , comes from Fick's first law of diffusion, where it connects the membrane flux of the molecule, J_m , to the concentration gradient, $C_D - C_A$:

$$J_m = P_m(C_D - C_A). \quad (24.1)$$

From Fick's Law's perspective, permeability is just a mathematical coefficient lacking obvious physical insight to help guide drug development for ADMET optimization. To address this discrepancy, a model of the permeability

coefficient is required. The first such model was developed by Overton, who related permeability to the oil–water partition coefficient [84]. In the 1960s, the homogenous solubility-diffusion (HSD) model was introduced [54, 55, 85, 86], which connects permeability to the membrane–water partition coefficient and the membrane diffusion constant. More recently, Marrink and Berendsen [87] developed the inhomogeneous solubility-diffusion (ISD) model, where permeability is related to the free energy and diffusion profiles across the membrane. Although each of these models provides some general insight into factors affecting permeability, there is little mechanistic information to help guide ADMET profile optimization.

To provide the detailed mechanistic information needed for ADMET optimization, we developed a new kinetic model of permeability [88]. This model is implemented in Orion using our WESTPA toolkit [8], which provides fully continuous permeation pathways of membrane crossing events along with permeability coefficient estimates using the weighted ensemble (WE) enhanced sampling strategy [8, 88]. The model is described briefly next, with more details provided in Ref. [88].

Passive membrane permeation can be a complicated process, involving an ensemble of conformations of a drug-like molecule and the membrane environment. The permeability coefficient for a given molecule can be obtained from the kinetic rate constant of membrane crossing for the reaction,



where D/A denotes the molecule species in the aqueous donor/acceptor compartment. Under steady-state conditions, permeability depends only on the forward rate constant $k_{D \rightarrow A}$ and the size of the “unstirred layer” of the donor compartment, l_D ,

$$P_m = k_{D \rightarrow A} l_D. \quad (24.3)$$

Finally, the forward rate constant, $k_{D \rightarrow A}$, can be shown to be equivalent to the steady-state probability fluxes from the donor to acceptor states, $\langle f_{D \rightarrow A}^{SS} \rangle$, calculated from the WE simulations [88]. Therefore, the permeability coefficient using this model is calculated by

$$P_m = \langle f_{D \rightarrow A}^{SS} \rangle l_D. \quad (24.4)$$

The OpenEye Permeability Floes perform the following functions: system preparation, MD equilibration, WE simulation, and permeability analysis of the membrane-permeate system using the kinetic model presented above (see Figure 24.15 for an example of the flow relationship diagram of the compute kernels). The system preparation takes a molecule of interest and readies it for simulation. The input can be any representation that can be read by OpenEye’s OEChem Toolkit [89]. All stereochemistry is handled by the Omega Toolkit [90], which will respect predefined stereochemistry if such information is provided. A diverse set of conformers is generated using Omega, and the top 20 conformers are selected and solvated by a 2 nm layer of water (compartment D) using PACKMOL [91] at a density of 1 g/cm³.

Each of the 20 solvated molecules is then combined with a pre-equilibrated lipid bilayer and subjected to energy minimization and equilibration in the NPT

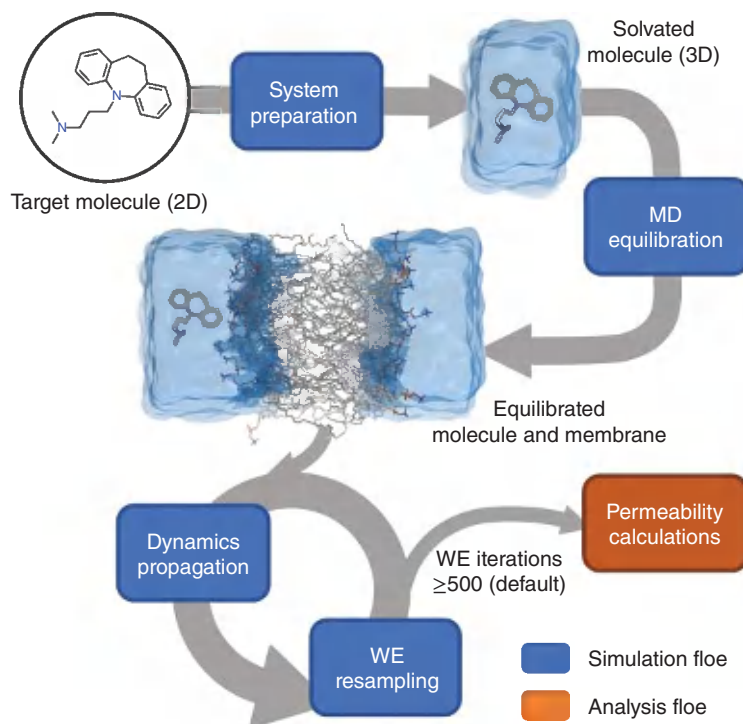


Figure 24.15 Schematic of the layout of the OpenEye Permeability Floes. The main components of the Floes are shown in *rectangles*, each of which contains a series of Cubes to perform its function. Functions of the Simulation or Analysis Floes are shown in *blue* or *orange*, respectively. The connectivity of these components is indicated by the *gray* arrows. The initial input is either a 2D or 3D molecule.

ensemble. WE logic and MD propagation must be performed iteratively; therefore, such cubes are connected in a cyclic fashion. Various parameters for system preparation and the WE logic are also exposed, including the option to turn on or off a “minimal adaptive” binning scheme [92], or whether to use a reweighting scheme for faster convergence to the steady state [93]. WE simulations are automatically parallelized to be run on GPUs from all AWS instance types available in Orion. Typically, a simulation will be scaled up by Orion to several hundred GPUs per WE iteration.

Permeability analysis is performed by a set of two Floes. A basic analysis floe estimates the permeability coefficient and extracts the highest weighted permeation trajectories. A second floe is used to calculate a set of auxiliary (i.e. non-WE) coordinates. The results of these analyses are illustrated using an actual drug-like molecule in the next section.

The model and method are illustrated using imipramine, a weakly basic drug (see Figure 24.16a). A set of WE trajectories was generated along with probability fluxes into the acceptor state, $\langle f_{D \rightarrow A}^{SS} \rangle$, to produce a permeability coefficient (P_m) that roughly converges to the MDCK-LE experimental value [94] after ~30 ns

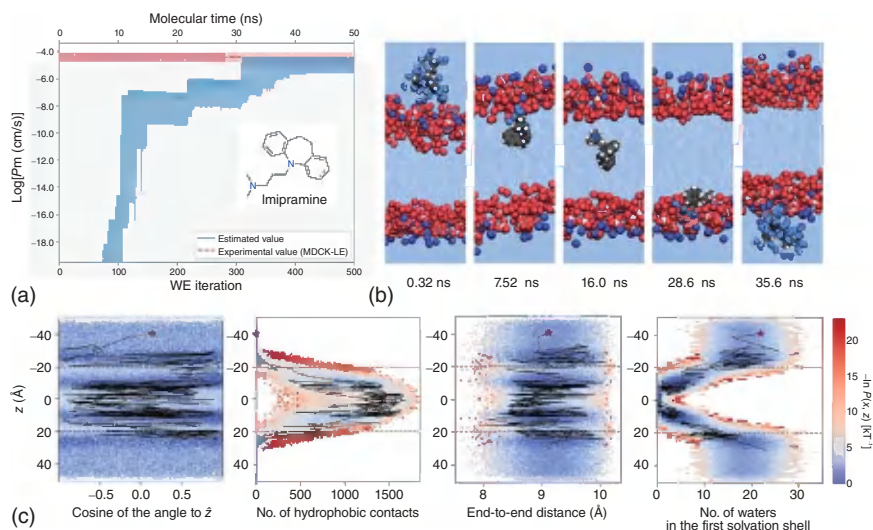


Figure 24.16 (a) Estimate of the permeability for imipramine at each WE iteration. The shaded region indicates the 95% confidence interval (CI) computed using the Monte Carlo bootstrapping procedure. Source: Adapted from Zhang et al. [88]. The final estimated log permeability is -4.86 (95% CI: $[-5.59, -4.51]$), which compares well to the MDCK-LE experiment (-4.42 ± 0.16). (b) Snapshots of the imipramine molecule (black) passing through a neat POPC bilayer (red) at selected molecular times. Atoms are represented by *van der Waals spheres* colored as follows: carbon – black, hydrogen – white, nitrogen – dark blue, phosphorus – white, and oxygen – red, except the carbons (white lines) and hydrogens, which are hidden for better visibility of the imipramine molecule. Water molecules in the first solvation shell of the imipramine molecule are shown in light blue. The molecular time at which the snapshots were taken is shown below their respective panel. (c) Free energy profile (in units of $k_B T$) along the bilayer normal, z (ordinate), and the cosine of the angle of the molecule with respect to the normal, hydrophobic contacts between the molecule and the membrane, the end-to-end distance of the molecule, and the number of waters in the first solvation shell of the molecule (abscissa, blue: $<5k_B T$, red: $>5k_B T$). The black line represents the top-weighted trajectory (probabilistic weight: 2.0×10^{-7}), with a purple star indicating the starting location. The approximate range of the membrane region is indicated by black dashed lines ($-20 \text{ \AA} < z < 20 \text{ \AA}$). The probabilities are symmetrized across the membrane to obtain the free energy profiles.

(Figure 24.16a). Considering the simplicity of our membrane, absolute agreement between model and experimental permeabilities should not be expected, yet the final estimated permeability coefficient of -4.86 within a 90% confidence interval of $[-5.59, -4.51]$ in log-scale is comparable to the experimental observation, which is -4.42 ± 0.16 . The Permeability Floe in Orion required ~ 5 days to complete and generated $\sim 12 \mu\text{s}$ of simulation data. We also extracted the top-weighted trajectory to highlight the most probable pathway of imipramine crossing the membrane (Figure 24.16b).

Several “auxiliary” coordinates (i.e. coordinates not used to drive the WE simulation) are also calculated to provide mechanistic insight into the permeation process. A selection of such coordinates is shown for imipramine in Figure 24.16c, which describes the following molecular features: orientation via the cosine of the

angle between the z -axis and the electric dipole moment; the local lipophilicity through the number of hydrophobic contacts of the molecule; the molecular length from the largest 3D inter-atomic distance; and a description of local solvation through the number of waters within the first solvation shell. From the dipole analysis (left panel), it is apparent there is no preferred orientation of the molecule in either the bulk water ($z < 20$ Å) or inside membrane (-20 Å $< z < 20$ Å) since only small free energy barriers exist. However, an orientational preference does appear for the highest weighted “walker” (black lines) near the headgroup/water interface (i.e. $\cos(\theta) = 0.5$ at $z = \pm 20$ Å). This suggests a molecule may typically use the same orientation to enter and exit the membrane. The linear combination of the number of hydrophobic contacts and the relative distance along z has been suggested to be the primary reaction coordinate for lipid insertion into a membrane [95], which is shown for imipramine in the second panel from left in Figure 24.16c. Here, the highest-weighted imipramine trajectory (black) samples a narrow low-energy pathway in the U-shaped distribution of hydrophobic contacts. The molecular length auxiliary coordinate shown in the second to the right panel in Figure 24.16c suggests the molecule can only undergo small (~ 2 Å) transitions in molecular length during permeation. Finally, the solvation descriptor shown in the right panel in Figure 24.16c suggests that imipramine can potentially stay solvated within the membrane, but the highest weighted trajectory (in black) mostly desolvates near the center of the membrane bilayer (see also Figure 24.16b). Taken together, these results suggest that optimizing the hydrophobic contacts and desolvation within the membrane could potentially aid the passive permeability process for imipramine.

ADMET liabilities can be costly because they may force a compound to be withdrawn from clinical trials. Here, we presented a kinetic model of permeability built as a floe in Orion to provide pathways and unbiased estimates of the permeation rate to aid in rational optimization of ADMET profiles, while allowing for co-optimization for high-affinity target binders. The results from the OpenEye Permeability Floe demonstrate how rich insight into the permeation process can be obtained through an extracted description of the orientation, local hydrophobic environment, molecular length, and (de)solvation ability. Through such mechanistic descriptions of the membrane permeation process, we hope to help identify and correct PK liabilities before a candidate is ever sent to the clinic.

24.7 Predicting Drug Crystal Forms

Typically, once a lead molecule has been established and optimized, it is moved forward to the formulation stage, which typically involves processing the molecule from its crystal form into a tablet. Predicting the accessible crystal structures (polymorphs) of a molecule can be invaluable, particularly in cases where there are multiple stable polymorphs. There are documented cases where the formulation of an efficacious drug has had to be revisited, simply because the polymorph in the final product changed even after years of using the same production processes. Reformulation and furthermore reproving efficacy is an expensive endeavor,

which necessitates a clear understanding of the polymorph landscape, so that a formulation can be designed to ensure a long-term stable specific polymorph in the final product. Experimental characterization of the crystal form landscape is time-consuming and not guaranteed to realize all low-energy polymorphs. Therefore, computational approaches for predicting the various stable and meta-stable crystal forms have received significant attention in recent years [96].

CSP of a molecule consists of two main challenges: (i) sampling – thorough enumeration of conformers, space groups, and orientations of the molecule forming a crystal lattice, and (ii) ranking – accurate ordering of the crystal packings by energy [97]. Usually, many promising trial packings are generated during the sampling phase using a fast molecular mechanics force field, which is further optimized and ranked using an accurate QM energy model such as dispersion-corrected density functional theory [96]. The high dimensionality of the sampling space and poor parallelizability of QM methods pose unique challenges to CSP. A typical CSP calculation of a flexible drug-like molecule can take months using conventional supercomputers or compute clusters. Many in the field have turned to the use of cloud computing to accelerate the workflow [10]. At OpenEye, with our cloud-native Orion platform, we are able to predict the crystal polymorph landscape of drug-like molecules within days.

The overall workflow for predicting the crystal polymorph landscape of a small molecule is schematically shown in Figure 24.17. This workflow has been automated through a series of Orion Floes. In the following, we describe the key steps in our CSP workflow using the compound Gestodene as an example (see insert in Figure 24.18).

There are four main stages in the CSP workflow. First, we identify all low-energy tautomers and perform torsional energy scans on all rotatable bonds. We then generate 3D conformers of each of the low-energy tautomers through exhaustive fragment

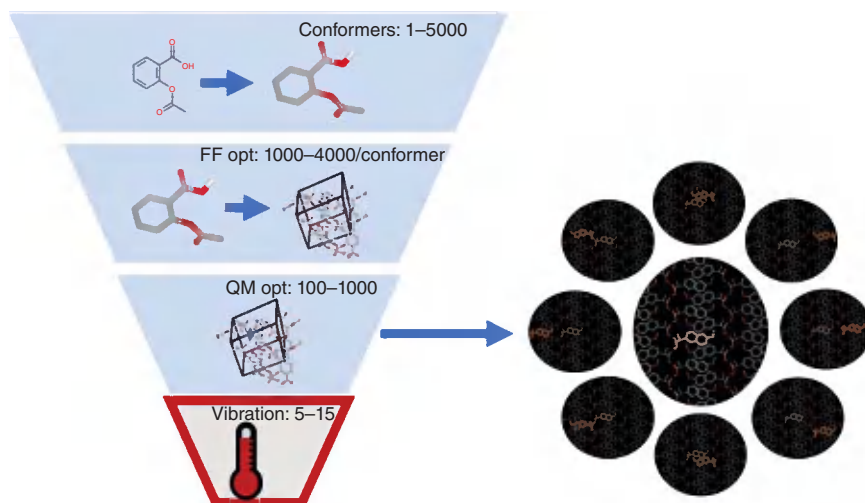


Figure 24.17 CSP protocol stages: Conformer ensemble generation, random packing and IEFF optimization, QM optimization of selected crystal structures using dimer expansion approach and finite temperature corrections.

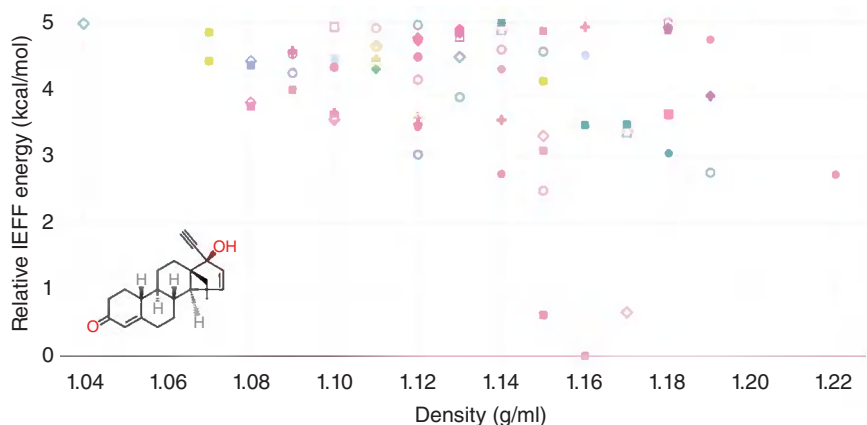


Figure 24.18 Results for Gestodene packings after IEFF optimization in Orion. The graph shows relative IEFF energy (in kcal) vs. density (g/ml) with the points colored by space group and shaped by conformer unique id. By default, this floe has a 5 kcal/mol energy window. For Gestodene, there are valid packings in 8 out of 20 chiral space groups.

and torsion sampling using Omega [90]. The geometries of these 3D conformers are then optimized using the quantum package Psi4 [19] with constrained torsions, at a low level of QM theory (e.g. HF3c [98]). The energies of the optimized conformers are evaluated at a higher level of theory (e.g. DFT-D). Conformers with high-strain energy are then filtered out. For Gestodene, there is a single dominant tautomer with 2 rotatable bonds, resulting in 18 low-energy conformers.

In the second stage, each 3D conformation is randomly packed into a specified list of space groups (e.g. the most frequent space groups in the CCDC database [99]) and rigidly optimized using our intermolecular energy force field (IEFF), a multipole-based force field [96]. The optimized crystal structures are then ranked by the sum of conformer strain energy and IEFF lattice energy, and the low-energy packings are selected for further analysis. For Gestodene, we packed the 18 low-energy conformers in the 20 most common chiral space groups, resulting in 72 low-energy crystal structures. The IEFF crystal energy landscape of Gestodene is shown in Figure 24.18, where each of the low-energy packings is marked by its space group and the specific 3D conformer.

The lowest energy structures from IEFF calculations are further optimized with QM methods. Optimization of the structures at the QM level is performed in two stages: (i) Loose – all low-energy IEFF crystal structures are optimized with a loose convergence criterion, and (ii) Tight – low-energy structures from loose optimization are further optimized with tight convergence criteria. We expect the error between loose and tight optimization to be approximately 1 kcal/mol. Typically, we use a low-level QM method such as HF3c for crystal geometry optimization and evaluate the energies of the optimized geometries using a high-level QM method such as DFT-D. For Gestodene, we optimized all 72 low-energy IEFF structures with both loose and tight convergence criteria using HF3c and evaluated the energies using B3LYP-D3MBJ6-31G*. The QM crystal energy landscape of Gestodene is shown in

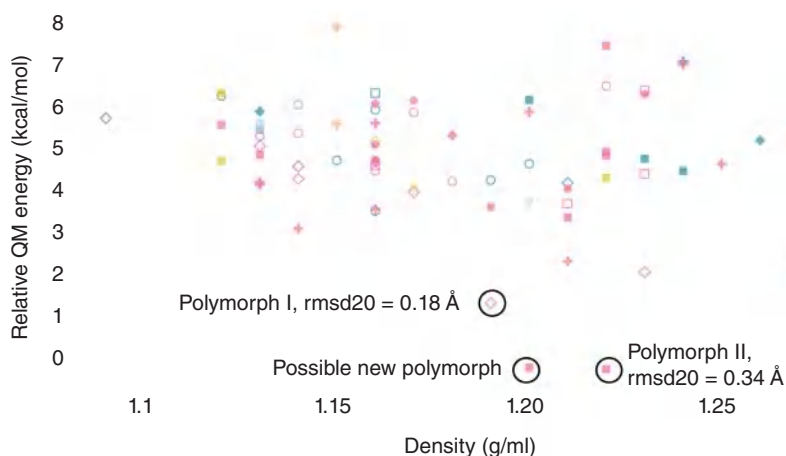


Figure 24.19 Results for Gestodene tight optimized crystal structures in Orion. The graph shows relative QM energy (in kcal/mol) vs. density (g/ml) with the points colored by space group and shaped by conformer unique id. Crystal RMSD20 (in Å) values are calculated using experimental polymorphs I and II.

Figure 24.20 Results for Gestodene predicted polymorphs in Orion. The 3D overlay shows (a) the experimental polymorph I (gray) and (b) polymorph II (gray) compared to the predicted crystal structure with the lowest RMSD20 (pink).

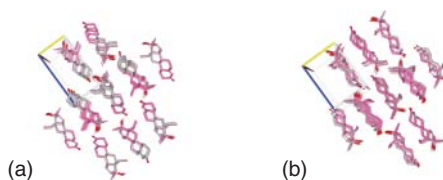


Figure 24.19. To evaluate our predictions, we compared the low-energy predicted structures to the two known crystal polymorphs of Gestodene. We found matching predictions for both experimental structures, as shown in Figure 24.20, and the structures were predicted to be highly stable (i.e. low energy), as indicated in Figure 24.19.

Finally, the entropy and free energies of the low-energy crystal forms at a finite temperature (298 K) are estimated using a harmonic approximation using HF-3c level calculations to compute a finite temperature correction to QM energies at 0 K. For small, rigid molecules such as Gestodene, we do not expect finite temperature corrections to be significant, and, as expected, the relative energies did not change significantly. However, these results do suggest that there is a possibility for a third polymorph of Gestodene, that is yet to be observed experimentally.

To exploit the massive scaling afforded on Orion by AWS, we developed novel methods and algorithms that are highly parallelizable at each stage of the above CSP workflow. For example, we developed a cluster expansion approach for optimizing and ranking crystal structures at the QM level that is highly parallelizable. In this approach, the total energy and gradients are calculated as the sum of all dimers in the crystal. Since all the dimer calculations are performed in parallel, the wall clock time for optimizing crystal structures is comparable to that of a single dimer calculation. Table 24.1 summarizes the total compute time and wall clock time for CSP of a

Table 24.1 Total compute and wall clock times for CSP of a variety of drug-like molecules.

Molecule	# rot. bonds	# heavy atoms	#Confs	Total compute time (CPU hours)	Wall clock time (hours)
O-Acetamidobenzamide	1	13	107	76K	40
Gestodene	2	23	18	59K	21
Diflunisal	2	18	110	216K	97
Proprietary 1	4	24	200	62K	22
Proprietary 2	4	29	376	290K	24
Proprietary 3	11	29	4000	562K	60
Proprietary 4	5	27	1317	597K	116
Proprietary 5	5	30	5341	639K	139
Proprietary 6	3	24	7001	552K	104
Proprietary 7	5	27	3234	702K	70

variety of drug-like molecules. Furthermore, we use the cluster expansion approach when calculating the vibrational entropy of the crystal, all with a wall clock time of hours instead of days and weeks. The cluster expansion approach appears as accurate as the conventional plane-wave approach but is significantly faster and more parallelizable.

24.8 Summary

The advent of cloud computing has dramatically changed the role of computation and modeling in drug discovery. Traditionally, calculations were performed on large in-house clusters or desktop machines. Computations for medicinal chemistry can consume massive amounts of time and hardware resources, which historically have been compounded by the fact that waiting times in a cluster queue could be substantial. The result is that even short calculations might have a slow turnaround time, which would limit their utility or even their use despite their inherent predictive powers. In many cases, it could be slower to perform a computation than to simply do the synthesis and assay work in the laboratory. On the other hand, those same large compute clusters could also sit unused, starved for work for long periods of time to the frustration of IT departments who want to see constant returns on the hardware investment. Cloud computing offers an excellent match for computational drug discovery, because it offers immediate elastic access to large heterogeneous resources, fitting both the burst-like nature of computation and the complex hardware requirements of different calculations required for a given project. Furthermore, advances in physics-based methods in combination with cloud computing have also made certain predictive calculations accurate enough to be a daily tool in the drug discovery process. Examples include free energy calculations to help prioritize compounds

for synthesis and generative modeling to suggest novel compounds that could be synthesized. In this chapter, we have described and illustrated some of the exciting advances in physics-based computation and data analysis powered by our cloud platform at each stage of the drug discovery cycle. We have also introduced the Orion cloud computing platform, which is a general-purpose computation engine where many new technologies like weighted ensemble MD, NES for free energy prediction, and CSP can be routinely performed. Additionally, Orion provides elastic access to hardware in a robust and fault-tolerant manner, removing the need for technical expertise in cloud computing. This seemingly unlimited compute resource offers a preview as to what computation will eventually deliver: on-demand compute resources with advanced data storage, analysis, sharing, and communication channels for the imported or generated data are a new paradigm for computer-aided drug discovery. Orion is an instant, almost infinite, datacenter optimized for CADD and available to all.

References

- 1 Moore, G.E. (1965). Cramming more components onto integrated circuits. *Electronics (Basel)* 38 (8): 114–117.
- 2 le Grand, S., Götz, A.W., and Walker, R.C. (2013). SPFP: speed without compromise – a mixed precision model for GPU accelerated molecular dynamics simulations. *Comput. Phys. Commun.* 184 (2): 374–380.
- 3 Darden, T., York, D., and Pedersen, L. (1993). Particle mesh Ewald: an $N \cdot \log(N)$ method for Ewald sums in large systems. *J. Chem. Phys.* 98 (12): 10089–10092.
- 4 Cooley, J.W. and Tukey, J.W. (1965). An algorithm for the machine calculation of complex Fourier series. *Math. Comput.* 19 (90): 297–301.
- 5 Dijkstra, E.W. (1959). A note on two problems in connexion with graphs. *Numer. Math. (Heidelb.)* 1 (1): 269–271.
- 6 Grebner, C., Malmerberg, E., Shewmaker, A. et al. (2020). Virtual screening in the cloud: how big is big enough? *J. Chem. Inf. Model.* 60 (9): 4274–4282.
- 7 Petrović, D., Scott, J.S., Bodnarchuk, M.S. et al. (2022). Virtual screening in the cloud identifies potent and selective ROS1 kinase inhibitors. *J. Chem. Inf. Model.* 62 (16): 3832–3843.
- 8 Russo, J.D., Zhang, S., Leung, J.M.G. et al. (2022). WESTPA 2.0: high-performance upgrades for weighted ensemble simulations and analysis of longer-timescale applications. *J. Chem. Theory Comput.* 18 (2): 638–649.
- 9 Gapsys, V., Pérez-Benito, L., Aldeghi, M. et al. (2020). Large scale relative protein ligand binding affinities using non-equilibrium alchemy. *Chem. Sci.* 11 (4): 1140–1152.
- 10 Zhang, P., Wood, G.P.F., Ma, J. et al. (2018). Harnessing cloud architecture for crystal structure prediction calculations. *Cryst. Growth Des.* 18 (11): 6891–6900.
- 11 Hawkins, P.C.D., Skillman, A.G., and Nicholls, A. (2007). Comparison of shape-matching and docking as virtual screening tools. *J. Med. Chem.* 50 (1): 74–82.

- 12 Morrison, J.P. (2010). *Flow-Based Programming: A New Approach to Application Development*, 2e. Scotts Valley, CA: CreateSpace.
- 13 Garey, M.R. and Johnson, D.S. (1990). *Computers and Intractability; A Guide to the Theory of NP-Completeness*. USA: W. H. Freeman & Co.
- 14 Ghodsi, A., Zaharia, M., Hindman, B. et al. (2011). Dominant resource fairness: fair allocation of multiple resource types. In: *Proceedings of the 8th USENIX Conference on Networked Systems Design and Implementation*, 323–336.
- 15 Wang, W., Li, B., Liang, B., and Li, J. (2016). Multi-resource fair sharing for datacenter jobs with placement constraints. In: *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*.
- 16 Rudolph, L., Slivkin-Allalouf, M., and Upfal, E. (1991). A simple load balancing scheme for task allocation in parallel machines. In: *Proceedings of the Third Annual ACM Symposium on Parallel Algorithms and Architectures – SPAA '91*, 237–245.
- 17 Rai, B.K., Sresht, V., Yang, Q. et al. (2022). TorsionNet: a deep neural network to rapidly predict small-molecule torsional energy profiles with the accuracy of quantum mechanics. *J. Chem. Inf. Model.* 62 (4): 785–800.
- 18 Rai, B.K., Sresht, V., Yang, Q. et al. (2019). Comprehensive assessment of torsional strain in crystal structures of small molecules and protein–ligand complexes using *ab initio* calculations. *J. Chem. Inf. Model.* 59 (10): 4195–4208.
- 19 Turney, J.M., Simmonett, A.C., Parrish, R.M. et al. (2012). Psi4: an open-source *ab initio* electronic structure program. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 2 (4): 556–565.
- 20 Abraham, M.J., Murtola, T., Schulz, R. et al. (2015). GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* 1–2: 19–25.
- 21 Virtanen, P., Gommers, R., Oliphant, T.E. et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* 17 (3): 261–272.
- 22 Pedregosa, F., Varoquaux, G., Gramfort, A. et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12 (null): 2825–2830.
- 23 Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Jia, Y., Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng (2015) *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*.
- 24 Eastman, P., Friedrichs, M.S., Chodera, J.D. et al. (2013). OpenMM 4: a reusable, extensible, hardware independent library for high performance molecular simulation. *J. Chem. Theory Comput.* 9 (1): 461–469.
- 25 Frisch, M.J., Trucks, G.W., Schlegel, H.B., Scuseria, G.E., Robb, M.A., Cheeseman, J.R., Scalmani, G., Barone, V., Petersson, G.A., Nakatsuji, H., Li, X., Caricato, M., Marenich, A. V, Bloino, J., Janesko, B.G., Gomperts,

- R., Mennucci, B., Hratchian, H.P., Ortiz, J. V, Izmaylov, A.F., Sonnenberg, J.L., Williams-Young, D., Ding, F., Lipparini, F., Egidi, F., Goings, J., Peng, B., Petrone, A., Henderson, T., Ranasinghe, D., Zakrzewski, V.G., Gao, J., Rega, N., Zheng, G., Liang, W., Hada, M., Ehara, M., Toyota, K., Fukuda, R., Hasegawa, J., Ishida, M., Nakajima, T., Honda, Y., Kitao, O., Nakai, H., Vreven, T., Throssell, K., Montgomery Jr., J.A., Peralta, J.E., Ogliaro, F., Bearpark, M.J., Heyd, J.J., Brothers, E.N., Kudin, K.N., Staroverov, V.N., Keith, T.A., Kobayashi, R., Normand, J., Raghavachari, K., Rendell, A.P., Burant, J.C., Iyengar, S.S., Tomasi, J., Cossi, M., Millam, J.M., Klene, M., Adamo, C., Cammi, R., Ochterski, J.W., Martin, R.L., Morokuma, K., Farkas, O., Foresman, J.B., and Fox, D.J. (2016) *Gaussian 16 Revision C.01*.
- 26 Erasmus, M.F., D'Angelo, S., Ferrara, F. et al. (2021). A single donor is sufficient to produce a highly functional in vitro antibody library. *Commun. Biol.* 4 (1): 350.
 - 27 Rondeau, J.-M. and Schreuder, H. (2015). Protein crystallography and drug discovery. In: *The Practice of Medicinal Chemistry*, 511–537. Elsevier.
 - 28 Blundell, T.L., Jhoti, H., and Abell, C. (2002). High-throughput crystallography for lead discovery in drug design. *Nat. Rev. Drug Discovery* 1 (1): 45–54.
 - 29 Anderson, A.C. (2003). The process of structure-based drug design. *Chem. Biol.* 10 (9): 787–797.
 - 30 Scapin, G., Potter, C.S., and Carragher, B. (2018). Cryo-EM for small molecules discovery, design, understanding, and application. *Cell Chem. Biol.* 25 (11): 1318–1325.
 - 31 Yip, K.M., Fischer, N., Paknia, E. et al. (2020). Atomic-resolution protein structure determination by cryo-EM. *Nature* 587 (7832): 157–161.
 - 32 Nakane, T., Kotecha, A., Sente, A. et al. (2020). Single-particle cryo-EM at atomic resolution. *Nature* 587 (7832): 152–156.
 - 33 Nogales, E. and Scheres, S.H.W. (2015). Cryo-EM: a unique tool for the visualization of macromolecular complexity. *Mol. Cell* 58 (4): 677–689.
 - 34 Verbeke, E.J., Zhou, Y., Horton, A.P. et al. (2020). Separating distinct structures of multiple macromolecular assemblies from cryo-EM projections. *J. Struct. Biol.* 209 (1): 107416.
 - 35 Baumeister, W. (2022). Cryo-electron tomography: a long journey to the inner space of cells. *Cell* 185 (15): 2649–2652.
 - 36 Waterhouse, A., Bertoni, M., Bienert, S. et al. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* 46 (W1): W296–W303.
 - 37 Webb, B. and Sali, A. (2016). Comparative protein structure modeling using MODELLER. *Curr. Protoc. Bioinformatics* 54 (1).
 - 38 Song, Y., DiMaio, F., Wang, R.Y.-R. et al. (2013). High-resolution comparative modeling with RosettaCM. *Structure* 21 (10): 1735–1742.
 - 39 Jumper, J., Evans, R., Pritzel, A. et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596 (7873): 583–589.
 - 40 Berman, H.M. (2000). The Protein Data Bank. *Nucleic Acids Res.* 28 (1): 235–242.

- 41 Burley, S.K., Berman, H.M., Bhikadiya, C. et al. (2019). Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.* 47 (D1): D520–D528.
- 42 Warren, G.L., Do, T.D., Kelley, B.P. et al. (2012). Essential considerations for using protein–ligand structures in drug discovery. *Drug Discovery Today* 17 (23–24): 1270–1281.
- 43 Wynn, M.L., Ventura, A.C., Sepulchre, J.A. et al. (2011). Kinase inhibitors can produce off-target effects and activate linked pathways by retroactivity. *BMC Syst. Biol.* 5 (1): 156.
- 44 Antolin, A.A., Ameratunga, M., Banerji, U. et al. (2020). The kinase polypharmacology landscape of clinical PARP inhibitors. *Sci. Rep.* 10 (1): 2585.
- 45 Hantschel, O. (2015). Unexpected off-targets and paradoxical pathway activation by kinase inhibitors. *ACS Chem. Biol.* 10 (1): 234–245.
- 46 *OpenEye Scientific Software* (www.eyesopen.com) (2022) Spruce Toolkit 2022.1.1.
- 47 Schindler, C.E.M., Baumann, H., Blum, A. et al. (2020). Large-scale assessment of binding free energy calculations in active drug discovery projects. *J. Chem. Inf. Model.* 60 (11): 5457–5474.
- 48 Gehlhaar, D.K., Luty, B.A., Cheung, P.P. et al. (2022). The Pfizer Crystal Structure Database: an essential tool for structure-based design at Pfizer. *J. Comput. Chem.* 43 (15): 1053–1062.
- 49 Harding, S.D., Armstrong, J.F., Faccenda, E. et al. (2022). The IUPHAR/BPS guide to PHARMACOLOGY in 2022: curating pharmacology for COVID-19, malaria and antibacterials. *Nucleic Acids Res.* 50 (D1): D1282–D1294.
- 50 Bairoch, A. (2000). The ENZYME database in 2000. *Nucleic Acids Res.* 28 (1): 304–305.
- 51 Pándy-Szekeres, G., Esguerra, M., Hauser, A.S. et al. (2022). The G protein database, GproteinDb. *Nucleic Acids Res.* 50 (D1): D518–D525.
- 52 Thomas, P.D., Campbell, M.J., Kejariwal, A. et al. (2003). PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.* 13 (9): 2129–2141.
- 53 Gough, J. (2002). SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Res.* 30 (1): 268–272.
- 54 Bateman, A., Martin, M.-J., Orchard, S. et al. (2021). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* 49 (D1): D480–D489.
- 55 Varadi, M., Anyango, S., Deshpande, M. et al. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* 50 (D1): D439–D444.
- 56 Nelson, R., Sawaya, M.R., Balbirnie, M. et al. (2005). Structure of the cross- β spine of amyloid-like fibrils. *Nature* 435 (7043): 773–778.
- 57 Batista, J., Hawkins, P.C., Tolbert, R., and Geballe, M.T. (2014). SiteHopper – a unique tool for binding site comparison. *J. Cheminf.* 6 (S1): P57.
- 58 le Guilloux, V., Schmidtke, P., and Tuffery, P. (2009). Fpocket: an open source platform for ligand pocket detection. *BMC Bioinf.* 10 (1): 168.

- 59 Walters, W.P., Stahl, M.T., and Murcko, M.A. (1998). Virtual screening – an overview. *Drug Discovery Today* 3 (4): 160–178.
- 60 Walters, W.P. and Wang, R. (2020). New trends in virtual screening. *J. Chem. Inf. Model.* 60 (9): 4109–4111.
- 61 Nicholls, A. (2008). What do we know and when do we know it? *J. Comput.-Aided Mol. Des.* 22 (3–4): 239–255.
- 62 McGann, M., Nicholls, A. and Enyedy, I. (2015). The statistics of virtual screening and lead optimization. *J. Comput.-Aided Mol. Des.* 29: 923–936.
- 63 Rogers, D. and Hahn, M. (2010). Extended-connectivity fingerprints. *J. Chem. Inf. Model.* 50 (5): 742–754.
- 64 Ewing, T., Baber, J.C., and Feher, M. (2006). Novel 2D fingerprints for ligand-based virtual screening. *J. Chem. Inf. Model.* 46 (6): 2423–2431.
- 65 Martin, L.J. and Bowen, M.T. (2020). Comparing fingerprints for ligand-based virtual screening: a fast and scalable approach for unbiased evaluation. *J. Chem. Inf. Model.* 60 (10): 4536–4545.
- 66 Hert, J., Willett, P., Wilton, D.J. et al. (2004). Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.* 44 (3): 1177–1185.
- 67 McGann, M. (2011). FRED pose prediction and virtual screening accuracy. *J. Chem. Inf. Model.* 51 (3): 578–596.
- 68 McGann, M. (2012). FRED and HYBRID docking performance on standardized datasets. *J. Comput.-Aided Mol. Des.* 26 (8): 897–906.
- 69 Briem, H. and Lessel, U.F. (2000). In vitro and in silico affinity fingerprints: finding similarities beyond structural classes. *Perspect. Drug Discovery Des.* 20 (1): 231–244.
- 70 Morrone, J.A., Weber, J.K., Huynh, T. et al. (2020). Combining docking pose rank and structure with deep learning improves protein–ligand binding mode prediction over a baseline docking approach. *J. Chem. Inf. Model.* 60 (9): 4170–4179.
- 71 Jastrzębski, S., Szymczak, M., Pocha, A. et al. (2020). Emulating docking results using a deep neural network: a new perspective for virtual screening. *J. Chem. Inf. Model.* 60 (9): 4246–4262.
- 72 Li, X., Xu, Y., Yao, H., and Lin, K. (2020). Chemical space exploration based on recurrent neural networks: applications in discovering kinase inhibitors. *J. Cheminf.* 12 (1): 42.
- 73 Raymond, J.-L. (2015). The chemical space project. *Acc. Chem. Res.* 48 (3): 722–730.
- 74 Kollman, P.A., Massova, I., Reyes, C. et al. (2000). Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc. Chem. Res.* 33 (12): 889–897.
- 75 Aldeghi, M., Bodkin, M.J., Knapp, S., and Biggin, P.C. (2017). Statistical analysis on the performance of molecular mechanics Poisson–Boltzmann surface area versus absolute binding free energy calculations: bromodomains as a case study. *J. Chem. Inf. Model.* 57 (9): 2203–2221.

- 76 Loeffler, H.H., Michel, J., and Woods, C. (2015). FESetup: automating setup for alchemical free energy simulations. *J. Chem. Inf. Model.* 55 (12): 2485–2490.
- 77 Bhati, A.P., Wan, S., Wright, D.W., and Coveney, P.V. (2017). Rapid, accurate, precise, and reliable relative free energy prediction using ensemble based thermodynamic integration. *J. Chem. Theory Comput.* 13 (1): 210–222.
- 78 Zwanzig, R.W. (1954). High-temperature equation of state by a perturbation method. I. Nonpolar gases. *J. Chem. Phys.* 22 (8): 1420–1426.
- 79 Mitchell, M.J. and McCammon, J.A. (1991). Free energy difference calculations by thermodynamic integration: difficulties in obtaining a precise value. *J. Comput. Chem.* 12 (2): 271–275.
- 80 Hahn, D.F., Bayly, C.I., Bobby, M.L. et al. (2022). Best practices for constructing, preparing, and evaluating protein-ligand binding affinity benchmarks [Article v1.0]. *Living J. Comput. Mol. Sci.* 4 (1): 1497.
- 81 Di, L. and Kerns, E. (2015). *Drug-Like Properties: Concepts, Structure Design and Methods from ADME to Toxicity Optimization*. Academic Press.
- 82 Prentis, R., Lis, Y., and Walker, S. (1988). Pharmaceutical innovation by the seven UK-owned pharmaceutical companies (1964-1985). *Br. J. Clin. Pharmacol.* 25 (3): 387–396.
- 83 Lipinski, C.A. (2004). Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discov. Today Technol.* 1 (4): 337–341.
- 84 Overton, C.E. (1895). *Über die osmotischen Eigenschaften der lebenden Pflanzen-und Tierzelle*. Fäsi & Beer.
- 85 Hanai, T. and Haydon, D.A. (1966). The permeability to water of bimolecular lipid membranes. *Journal of Theoretical Biology.* 11 (3): 370–382. [https://doi.org/10.1016/0022-5193\(66\)90099-39:43](https://doi.org/10.1016/0022-5193(66)90099-39:43).
- 86 Finkelstein, A. (1976). Water and nonelectrolyte permeability of lipid bilayer membranes. *Journal of General Physiology.* 68 (2): 127–135. <https://doi.org/10.1085/jgp.68.2.127>.
- 87 Marrink, S.J. and Berendsen, H.J.C. (1996). Permeation process of small molecules across lipid membranes studied by molecular dynamics simulations. *J. Phys. Chem.* 100 (41): 16729–16738.
- 88 Zhang, S., Thompson, J.P., Xia, J. et al. (2022). Mechanistic insights into passive membrane permeability of drug-like molecules from a weighted ensemble of trajectories. *J. Chem. Inf. Model.* 62 (8): 1891–1904.
- 89 (2022) *OEChem Toolkit*.
- 90 (2022) *Omega Toolkit*.
- 91 Martínez, L., Andrade, R., Birgin, E.G., and Martínez, J.M. (2009). PACKMOL: a package for building initial configurations for molecular dynamics simulations. *J. Comput. Chem.* 30 (13): 2157–2164.
- 92 Torrillo, P.A., Bogetti, A.T., and Chong, L.T. (2021). A minimal, adaptive binning scheme for weighted ensemble simulations. *J. Phys. Chem. A* 125 (7): 1642–1649.
- 93 Bhatt, D., Zhang, B.W., and Zuckerman, D.M. (2010). Steady-state simulations using weighted ensemble path sampling. *J. Chem. Phys.* 133 (1): 014110.

- 94 Dickson, C.J., Hornak, V., Bednarczyk, D., and Duca, J.S. (2019). Using membrane partitioning simulations to predict permeability of forty-nine drug-like molecules. *J. Chem. Inf. Model.* 59 (1): 236–244.
- 95 Rogers, J.R. and Geissler, P.L. (2020). Breakage of hydrophobic contacts limits the rate of passive lipid exchange between membranes. *J. Phys. Chem. B* 124 (28): 5884–5898.
- 96 Elking, D.M., Fusti-Molnar, L., and Nichols, A. (2016). Crystal structure prediction of rigid molecules. *Acta Crystallogr. B Struct. Sci. Cryst. Eng. Mater.* 72 (4): 488–501.
- 97 Oganov, A.R. (2018). Crystal structure prediction: reflections on present status and challenges. *Faraday Discuss.* 211: 643–660.
- 98 Sure, R. and Grimme, S. (2013). Corrected small basis set Hartree-Fock method for large systems. *J. Comput. Chem.* 34 (19): 1672–1685.
- 99 Groom, C.R., Bruno, I.J., Lightfoot, M.P., and Ward, S.C. (2016). The Cambridge structural database. *Acta Crystallogr. B Struct. Sci. Cryst. Eng. Mater.* 72 (2): 171–179.

25

Cloud-Native Rendering Platform and GPUs Aid Drug Discovery

Mark Ross¹, Michael Drummond², Lance Westerhoff³, Xavier Barbeau⁴, Essam Metwally⁵, Sasha Banks-Louie⁶, Kevin Jorissen⁶, Anup Ojah⁶, and Ruzhu Chen⁶

¹GridMarkets, San Francisco, CA, USA

²Chemical Computing Group, Montreal, QC, Canada

³QuantumBio, State College, PA, USA

⁴NuChem Sciences, Montreal, QC, Canada

⁵Merck, San Francisco, CA, USA

⁶Oracle, Austin, TX, USA

25.1 Introduction

Since computer-aided drug design (CADD) emerged as a method of modeling and analyzing medicinal compounds more than four decades ago [1], drug researchers have been able to screen larger numbers of molecules and identify the most promising drug candidates faster and cheaper than they could in a lab. While computational chemists have largely embraced molecular mechanics as the “best-practice modeling technique” in CADD for understanding how drug candidates bind to proteins, thanks to MM’s balance of time, cost, and accuracy, recent advances in processing hardware, modeling software, and cloud-native rendering platforms have placed more complex CADD simulations within reach.

The mathematical complexity of CADD derives from (i) the scale of the problem: It can take a huge number of atoms to represent a drug molecule interacting with a target cell in a medium; (ii) the physics: Atomic-scale interactions are governed by quantum mechanics and have many-body characteristics, which are incredibly demanding to compute; and (iii) the dynamics: The molecules individually are not in a permanent state, but rapidly fluctuate between states due to thermal energy and interactions with their environment, which can markedly affect their chemical properties.

In practice, one must compromise accuracy along one or more of these dimensions due to limited time and available compute resources. It’s no wonder then that generations of chemists and physicists have worked on theories that capture as much of the behavior of interest for a given class of problems at a reasonable computational cost. For example, instead of calculating a system “ab initio” using only the basic laws of physics, many successful theories use empirical or

semi-empirical (fitted to experimental or *ab initio* data) “effective” parameters that can approximate many-body electron–electron interactions within a much cheaper single-body theory. These methods may also hide problematic effects in highly correlated systems, represent thermal effects in a ground-state formulation; or approximate quantum behavior as best as possible within a classical formulation.

In particular, molecular mechanics models use classical force fields to produce approximations of simulated atomic reactions. These models have won favor among drug researchers because they can be built more quickly and are much less expensive to run than their quantum mechanics counterparts. MM “scales well,” which means that if you simulate a system that’s twice as large as the previous one, the needed compute cycles will increase by a reasonably proportional factor (whereas some of the most accurate methods might require more than eight times the number of cycles). This makes MM particularly well suited for simulating very large problems.

But such approximations offered by molecular mechanics models may soon be replaced by probabilistic outcomes, as advances in high-performance computing (HPC), molecular modeling software, and cloud platforms have placed more advanced molecular dynamics, quantum mechanics, and other complex CADD simulations within reach. These advances offer better modeling accuracy, and therefore improved scientific discovery or commercial development, within a short period of time and a relatively small budget. While these more complex methods were focused mostly on computing small molecule properties in the past, these methods are increasingly being applied to larger systems as increasing computational power becomes more accessible.

Meanwhile, there have also been great advances in experimental technology, which is increasingly able to resolve protein structures and chemical properties with great spatial and time resolution, using, for example, X-Ray diffraction, NMR study, or cryo-EM analysis, with the number of published structures increasing almost exponentially each year [2].

The advent of HPC systems becoming significantly faster and more efficient over time was driven primarily by Moore’s Law for much of the history of computational chemistry and biology. But in the last decade or so, as Moore’s Law famously topped out, new trends have emerged that continue to push the field forward. These trends are the focus of this paper.

First, GPUs have become more widely available, and an increasing number of HPC applications have been ported by their developers to run their most expensive parts (e.g. solvers) on the GPU, often speeding them up by orders of magnitude even within a single GPU workstation. GPU-accelerated algorithms are now widely applied to molecular modeling, allowing drug researchers to simulate millions of atoms in a wide variety of complex systems, compared to just thousands of atoms in the 1990s [3].

Second, since machine learning techniques have been introduced to CADD, breakthroughs such as AlphaFold-2 have used neural networks (NN) to predict the 3D structure of nearly every protein made by humans [4]. Third, cloud computing

has dramatically increased the availability of compute resources (CPU or GPU) to CADD researchers.

Cloud computing presents a utility model for computing, where, instead of buying and maintaining a fixed workstation or compute cluster over several years, a researcher can connect to a cloud data center and use any number of compute resources on demand, or for a prescribed amount of time.

As the cost is typically only for the volume of compute cycles consumed (also termed a “utility model”), this allows “bursting” to a much larger compute cluster, and hence solves a much bigger scientific question than could be afforded in the traditional “op-ex” model. The cloud has also democratized CADD capabilities, which are now accessible to anyone with a credit card and basic computational skills rather than being restricted to those operating an in-house HPC data center. However, this democratizing potential is often contingent on the availability of management tools, such as a user-friendly CADD portal that lets a researcher specify the problem in a familiar way, and then handles most of the cloud deployment aspects.

This new paradigm is illustrated in the remainder of this paper. We will describe three innovative use cases with a particular implementation of the paradigm: the pharma.GridMarkets.com CADD platform deployed on Oracle Cloud Infrastructure.

25.2 Complex Molecular Mechanics at Scale

For the past two decades, the transcription factor STAT3 has been a tempting target, implicated in multiple disease pathologies, including cancer. Success with both small molecule and peptidomimetic campaigns has proven elusive due to poor affinity, selectivity, and bioavailability. Recently, researchers at the University of Michigan approached this challenge using a different approach: PROteolysis Targeting Chimeras (PROTACs), which require an E3 ligase such as cereblon to degrade STAT3 (rather than simply inhibiting it) to halt tumor growth (Figure 25.1) [5].

While exploratory work is quite expensive in terms of time and resources, an accurate computational model can represent a significant advantage. Researchers at Chemical Computing Group (CCG) dove into the researchers’ library of two dozen initial PROTACs to see what lessons could be learned.

Using CCG’s molecular operating environment (MOE) software, the researchers broke down each of the initial PROTACs into four parts and then exhaustively pieced together all possible combinations back together. This produced 509 novel PROTACs, which were uploaded using GridMarkets’ Envoy plugin and solved using 32-core virtual machines in the cloud. The CADD process took three days and five hours (versus three to four weeks on a typical on-premises workstation).

In total, 594 potential STAT3-cereblon interfaces and 71 136 PROTAC conformations were found, resulting in 42 254 784 possible ternary complexes. With some filtering, this led to 1 347 654 ternary complexes, 95 821 of which were saved for further analysis after considering complex suitability [6].

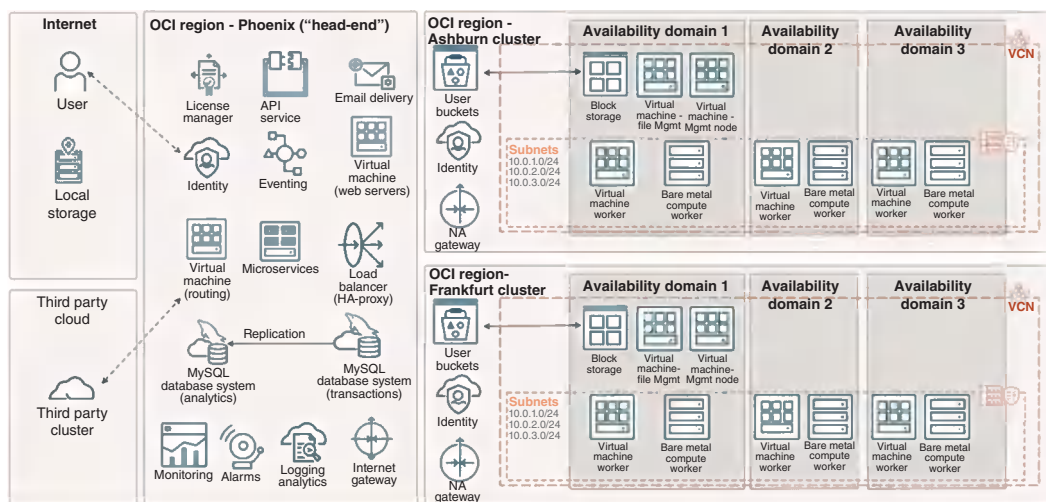


Figure 25.1 The pharma.GridMarkets.com platform runs in three regions on Oracle Cloud Infrastructure. The primary site is in the Phoenix cloud region, which runs all of the microservices, databases, event management, user authentication, logging, and API services. The Ashburn and Frankfurt regions run the HPC clusters, in which molecular modeling simulations are run on 32-core virtual machines or Bare Metal GPU servers.

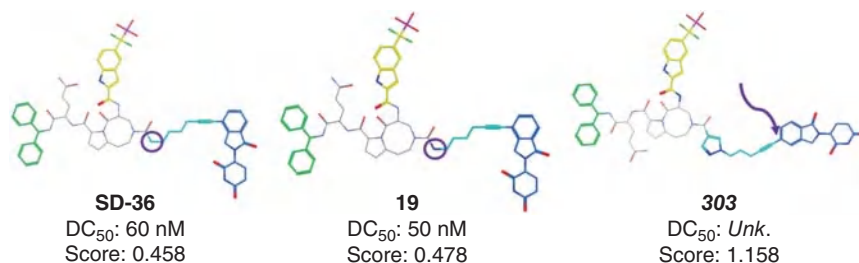


Figure 25.2 The Chemical Computing Group's scoring method used in this model is to divide the "maximum double cluster population" by the number of PROTAC conformations.

The entire STAT3 simulation screened 291 ternary complexes each minute (~5 per second), generating a total of 32 GB of data. The final analysis predicted 95 new PROTACs (making up 18.7% of the total novel set) to be better than the initial lead, Compound 14 ("SD-36"). Of those, 19 new PROTACs (3.7%) gave better-predicted scores than any compound in the initial set, some of which (see Figure 25.2) resulted from subtle structural variations that would have been otherwise difficult to discover among the initial set of 24 investigated PROTACs. And, while the final activities have yet to be reported, the value of more exhaustively sampling and exploring the druggable space cannot be underestimated. This is the basis of rational drug design.

25.3 Modeling Billions of Molecules in a Day

In the spring of 2020, as the coronavirus began its infectious rampage, many drug researchers were suddenly locked out of quarantined data centers and forced to work remotely.

Tasked with modeling billions of molecule combinations against the key proteins that COVID-19 needs to reproduce, computational chemist Andy Jennings used X-ray diffraction and 3D molecular modeling software to build a crystal structure of the virus that causes COVID-19.

These enormous calculations require more computing power than even some of the largest pharmaceutical companies can accommodate. It's difficult for most companies to justify buying an on-premises server cluster big enough to speed through a few bursts because, for the better part of the year, that cluster sits idle.

Cloud bursting solves this dilemma by allowing a scientist to fit his computational resources to the needs of his current research challenge. Coupled with a user-friendly platform to deploy CADD simulations on elastic cloud infrastructure without a long learning curve, Jennings was able to start running his simulations in less than 24 hours. The simulations ran on demand. Without a need to queue requests or schedule renderings (Figure 25.3).

Jennings then simulated the reactions between different molecules and proteins, tested multiple compounds until he found one that completely bound to the

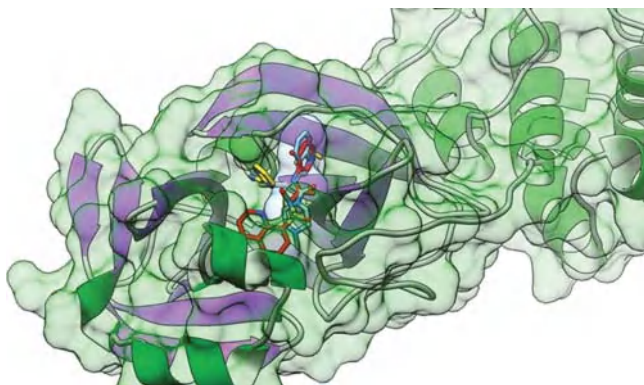


Figure 25.3 Drug molecules bound to an active site of a protease. The solid green and solid purple regions represent the protein backbone of the COVID-19 main protease, with the protein surface shown partially transparent. The multi-colored sticks are different drug molecules bound to the active site of the protease.

structure's surface, inhibiting the protease and stopping the COVID-19 virus from replicating.

Whether simulating a drug molecule of 20 atoms with quantum mechanics to learn how electrons behave or assessing multiple molecules made up of 2 million atoms, these tasks could take weeks using a cluster of traditional on-premises high-performance computers.

Simulating a drug molecule's reaction to different proteins could take 1000 seconds of CPU hours to complete. These enormous calculations require more computing power than even some of the largest pharmaceutical companies can accommodate. It's difficult for most companies to justify buying an on-premises server cluster big enough to speed through a few bursts because, for the better part of the year, that cluster sits idle. Cloud bursting solves this dilemma by allowing a scientist to fit his computational resources to the needs of his current research challenge.

Coupled with a user-friendly platform to deploy CADD simulations on elastic cloud infrastructure without a long learning curve, Jennings was able to start running his simulations in less than 24 hours. **The** simulations ran on demand. Without a need to queue requests or schedule renderings.

25.4 Faster Free Energy

To better understand intermolecular interactions between ligands, drug candidates, and proteins, life sciences software company, QuantumBio is using MovableType (MT) methods as an alternative to conventional protein-ligand (PL) docking and scoring.

To validate the efficacy of the MT protocol across a broad range of protein classes, researchers at QuantumBio recently used the industry-standard Comparative

Assessment of Scoring Functions, which contains 57 protein targets and 285 ligands. They then used 10 protein targets with a total of 248 ligands selected from the PDBBind database to further explore MT performance in virtual screening tasks targeting large-ligand structural diversity for individual receptors.

A MT module using protein–ligand poses were used in conjunction with QuantumBio’s DivCon software suite, which provides modules for quantum mechanics, molecular mechanics, and mixed-QM/MM (ONIOM) support for protein–ligand treatment and characterization. The MT poses were generated through X-ray, Cryo-EM, and NMR through docking and molecular dynamics for cases in which docking alone was less predictive. QuantumBio’s method consisted of four primary submodules:

1. **MT_{ScoreES}(EndState)**: Single protein–ligand structure bound to a single ligand to form a protein–ligand conformation, which quickly and easily scores the conformation.
2. **MT_{ConfSearch}(ligand)**: Simplified Molecular-Input Line-Entry System (SMILES) string, representing a drug-like compound to generate a set of ligand conformers using the MT database
3. **MT_{ScoreE}(Ensemble)**: In contrast to MT_{ScoreES}, MT_{ScoreE} utilizes multiple three-dimensional models, including an enzyme structure along with a ligand to create a protein–ligand complex from this ligand using a third-party docking package. This helps computational chemists explore an ensemble of bound-ligand and free-ligand conformations and produce a single value for the binding affinity.
4. **MT_{Dock}**: MT_{ScoreE} Multiple, reasonable placements (poses) for the conformers generated by MT_{ConfSearch}(ligand). This can be accomplished using third-party “pose generators” (called dockers, such as those provided by MOE or Glide). MT_{Dock} is a method built on MT_{Score} that accomplishes that goal independently of third-party dockers.
5. **MT_{Flex}**: Finally, the above-noted MT_{ConfSearch}(ligand) method has been expanded to include protein sidechains/loops as well as ligands.

To run its MovableType simulations, QuantumBio used the pharma.GridMarkets.com platform, allowing users to run as many ligands as they want to analyze, simulating large-scale free energy calculations over hundreds or even thousands of compounds per day.

QuantumBio’s free energy analysis, using induced-fit (flexible-target/flexible-ligand) docking with a conventional docking program plus MT, was performed in 15–30 minutes per ligand. Since ligands were treated separately, each ligand was characterized in a completely independent manner.

Through the GridMarkets “Envoy” plugin, QuantumBio can submit 10’s, 100’s, or even 1000’s of ligands, all of which could be run in parallel within a day, depending upon the number of machines requested and the number available.

In previous experiments [7, 8], the focus on the MT method was pairing the technology with docking software such as external docking methods supplied by the CCG in the MOE) or internal heatmap-based docking methods.

QuantumBio has since expanded this work, using conventional molecular dynamics (cMD) snapshots determined with the AMBER18 package.

Table 25.1 This table shows the impact of increased sampling on the predictability of MT (Pearson-R) of the Merck KGaA set [9]. Time is reported as the average time (in CPU minutes) among the set of protein–ligand complexes treated.

Protein	#Ligands	dock+MT _{ScoreE}			pure-cMD+MT _{ScoreE}		
		AMBER	GARF	Time (m)	AMBER	GARF	Time (m)
HIF-2 α	42	0.107	0.237	4	0.271	0.409	742
PFKFB3	40	0.010	0.056	9	0.676	0.642	1452
SHP-2	26	−0.065	0.156	23	0.232	0.350	1684
TNKS2	27	0.591	0.541	6	0.680	0.447	860

The bold font is for the highest Pearson-R (best score of the row).

The results, summarized in Table 25.1, show that in cases in which MOEDock+MT_{ScoreE} are less predictive, the additional global sampling in cMD+MT_{ScoreE} increases the predictive capabilities of the method. While these results were with the AMBER MD engine coupled with MT, similar results would likely be observed with alternative MD engines such as NAMD and GROMACS.

25.5 Vision for the Future

Each new generation of GPU hardware continues to outperform the previous one by larger factors than seen in the CPU market. This shows the clearest impact in the AI/ML space itself, where each consecutive generation (say, H100 over A100 over V100, using NVIDIA nomenclature) enables qualitatively more complex and powerful models to be trained.

This phenomenon can be witnessed by rapidly increasing model parameter counts (now in the trillions) and the ability of natural language applications to mimic human speech and writing. Cloud computing also plays a role in this evolution, as access to the latest GPU hardware becomes pervasive. Demand for HPC resources will only increase as GPU and ML penetration grows in CADD and other science domains.

A current area of focus for cloud providers is building extensive “GPU super clusters,” consisting of dozens, hundreds, or even thousands of servers, each hosting multiple GPU accelerators, and connected with similar low-latency, high-bandwidth networks that have long been used in HPC. GPU-specific communication technologies, such as GPU Direct, GPU Storage Direct, and NVLink, help effectively distribute the largest ML and HPC problems across super clusters.

With low-latency, high-bandwidth all-to-all GPU communication, even a GPU super cluster of just a few nodes can crunch through complex molecular modeling simulations in minutes, moving the goal posts on the most innovative new problems that can be tackled at scale.

25.6 Concluding Remarks

The premise of this chapter is that improvements in computational resources and CADD methods are driving ever more progress in the science and discovery of drugs, particularly when exploited through a powerful orchestration tool, such as pharma.GridMarkets.com. Similarly, it has been estimated that most of the improvements in the study of protein folding (as measured by the error in protein structure prediction relative to experimental observations in a common benchmark) are tied to the influence of an exponential increase in computational power over time [10].

As stated in the introduction, (i) GPUs, (ii) AI/ML, and (iii) cloud computing are the current drivers of this virtuous correlation between increases in computing power and scientific accuracy. These three trends will continue to drive progress for years to come (with perhaps longer-term advances, such as domain-specific architectures [DSAs] and quantum computing).

CADD platforms like pharma.GridMarkets.com will continue to be a driving force in democratizing that power and putting it in the hands of drug researchers worldwide to push their field forward.

Disclaimer

The views expressed in this chapter are solely those of the authors and do not necessarily reflect the views of affiliated institutions.

References

- 1 Osakwe, O. (2016). The significance of discovery screening and structure optimization studies. In: *Social Aspects of Drug Discovery, Development and Commercialization*. Academic Press via Elsevier, Inc.
- 2 *Growth Released Structures*. <https://www.rcsb.org/stats/growth/growth-released-structures>.
- 3 Phillips, J. C. et al. Scalable molecular dynamics on CPU and GPU architectures with NAMD. *J. Chem. Phys.* 153, 44130 (2020). doi: <https://doi.org/10.1063/5.0014475>
- 4 *50-Year-Old Grand Challenge in Biology*, <https://www.deepmind.com/blog/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology>.
- 5 Zhou, H. et al. (2019). Structure-based discovery of SD-36 as a potent, selective, and efficacious PROTAC degrader of STAT3 protein. *J. Med. Chem.* 62: 11280–11300.
- 6 Drummond, M.L., Henry, A., Li, H., and Williams, C.I. (2020). Improved accuracy for modeling PROTAC-mediated ternary complex formation and targeted protein degradation via new in silico methodologies. *J. Chem. Inf. Model.* 60: 5234–5254.

- 7 Zheng, Z., Zheng, O.Y., Borbulevych, H.L. et al. (2020). MovableType software for fast free energy-based virtual screening: protocol development, deployment, validation, and assessment. *J. Chem. Inf. Model.* 60 (11): 5437–5456. <https://doi.org/10.1021/acs.jcim.0c00618>.
- 8 Westerhoff, L.M. and Zheng, Z. (2021). Fast, routine free energy of binding estimation using MovableType. In: *ACS Symposium Series*, vol. 1397 (ed. K.A. Armacost and D.C. Thompson), 247–265. Washington, DC: American Chemical Society <https://doi.org/10.1021/bk-2021-1397.ch010>.
- 9 Liu, W., Liu, Z., Liu, H. et al. (2022). Free energy calculations using the movable type method with molecular dynamics driven protein–ligand sampling. *J. Chem. Inf. Model.* 62 (22): 5645–5665. <https://doi.org/10.1021/acs.jcim.2c00278>.
- 10 Neil C. Thompson, Shuning Ge, Gabriel F. Manso, 2022. *The Importance of (Exponentially More) Computing Power*. arXiv:2206.14007v1

26

The Quantum Computing Paradigm

Thomas Ehmer¹, Gopal Karemore², and Hans Melo³

¹Merck KGaA, IT Healthcare R&D Digital Innovation, Frankfurter Str. 250, Darmstadt 64293, Germany

²Novo Nordisk A/S, Digital & Research Intelligence, Novo Allé 1, Bagsværd 2880, Denmark

³Menten AI, Inc., 1160 Battery Street East, Suite 100, San Francisco, CA 94111, USA

Nature isn't classical, dammit, and if you want to make a simulation of nature, you'd better make it quantum mechanical, and by golly it's a wonderful problem, because it doesn't look so easy.

Richard Feynman [1]

26.1 What to Expect

26.1.1 Motivation

Molecular biology and biochemistry involve the study of the structures and interactions of molecules in living organisms, and these processes are governed by the laws of quantum mechanics. As we see in some other chapters of this book, this means that the behavior of these molecules can be described using the principles of quantum mechanics, such as wave-particle duality and uncertainty. We also see that, while the theoretical foundations of these quantum mechanical processes are well understood, it can be challenging to compute the solutions to the relevant quantum mechanical equations, particularly for larger and more complex systems. Therefore, researchers often rely on classical mechanical models or approximations that simplify the calculations, but these models may not always accurately capture the full complexity of the quantum mechanical interactions. Quantum computation may offer a way to more accurately simulate these processes and better understand the underlying molecular mechanisms in biology and biochemistry.

Also, the maturity of quantum devices – such as quantum computers – increased rapidly during the last decade. With a larger number of quantum bits and decreasing amount of noise in quantum operations, the current hardware is becoming more and more competitive with classical devices. This leads to a growing list of possible applications and research communities, both in industry and university. Even though the

current quantum hardware is not able to compete with classical devices for realistic use cases, the potential becomes clearer.

The tricky question is whether the emerging branches of quantum computation may eventually deliver a significant advance over traditional approaches, and if so, how would algorithms look like and what else is needed?

26.1.2 Structure of the Chapter

We will start in Section 26.2 from the basics to see what is meant by “a new paradigm” in relation to quantum computing, the fundamental difference to classical computing, and introduce concepts (“specific features”) of quantum mechanics and quantum information theory like entanglement, superposition, interference. Followed by a short overview of its possible architectures (*annealing* and *gate*). Before we finally dive into Sections 26.3.6, 26.4, and 26.5 into the main categories where quantum computing can play a role and could provide an advantage over other tools, specifically in the context of computational drug discovery, we like to highlight in Section 26.2.3 some more general challenges, both conceptual and technological. Only then can we look at some famous algorithms and show some examples per category – and at the same time discuss the specific challenges. Here in Section 26.3.11, we will present and reference work that is relevant also for material research, chemistry, drug discovery and pharma in general. At the end of Section 26.7 we encourage the reader to contribute and join the fantastic quantum computing community by hinting at existing communities and also proposing some topics for future research and/or development directions.

A brief disclaimer: We have celebrated 40 years of quantum computing [2], and it is a very active field, which by its nature comes with quantum leaps. By the time of this writing we see every week several publications on the progress of quantum computing for novel algorithms, error corrections, technical maturity, and applications in chemistry and machine learning (ML), so by the publication of the book, we might see quantum leaps implemented, which were not yet expected when writing about them. A recent report from RAND gives a focused overview of potential applications of quantum computing and simulation in the life sciences [3].

There are other interesting quantum technology applications that are also of relevance to drug discovery that we however can not consider in this article, mainly further quantum simulators (i.e. building and controlling multi-particle systems) and quantum sensing applications. An associated emerging topic of interest that we only like to mention briefly but cannot elaborate on in this chapter is the upcoming field of quantum biology, where in principle the same technology used to control computing qubits could be used to control chemical reactions and downstream biological results [4].

26.2 Another New Paradigm

We use computers and mobile devices everywhere in our daily life, and chips are embedded nearly everywhere. Indeed, it is meanwhile so natural that we can ignore

(and tend to forget) how digital information processing is done. While the exact date and origin of the introduction of the binary system to the western world by Gottfried Wilhelm Leibniz is controversial among historians [5–7], his mathematical work, *Essay d’une nouvelle science des nombres* submitted in the year 1701 was eventually the basis for the digital revolution three centuries later. Interestingly, Leibniz at his time did stick himself to the base 10 for his early computation machines [8]. Without overstressing the definition from Thomas Kuhn [9] of what a paradigm is, we could say that this “digital” already qualifies as a “new paradigm,” because looking back only 60 years, we can say that it changed our world and our interaction with it fundamentally, and if we look at Jeremy Rifkins *Third Industrial Revolution* [10], it will continue to do so. Now we see in the press that yet again another new “quantum” paradigm is appearing. Let us explore why this is so and what the new “quantum” paradigm is about, and if it really has the hyped potential to be as fundamental as the “digital” paradigm was. Spoiler, it has to do with the fact that we deal with a so-called non-Von-Neumann architecture.

26.2.1 Digital

To understand what is different in quantum computing, let us have first a brief refresher of the “classical” digital world and look at the specific features of this tool, namely that information is encoded and stored in bits. These bits have specific features, namely:

1. a bit is ALWAYS in one (and ONLY one) of two states, often called 0 and 1
2. we can apply any function (or operator) to it
3. it can be freely read WITHOUT disturbing the original state in the memory

So, bits can be imagined as miniaturized switches, with the position (or state) “ON” or “OFF,” technically realized in principle as a switch that either lets a current (or light) pass or not. Long-term storage (Hard disk) is typically magnetized $\uparrow$ upward or $\downarrow$ downward, or encoded optically (reflective or transparent). Communication is either high-/low frequency or light on/off. All digital information is thus encoded in a sequence of zeros and ones. Using the smiley example here in the following, two different two-character strings and their binary representation, we see that *one bit string encodes one piece of information*.

$$\begin{aligned} 8) &\rightarrow 0011100000101001 \\ ;) &\rightarrow 0011101100101001 \end{aligned} \tag{26.1}$$

Computation with (or manipulation of) bits in the digital world is done by applying specific logic gates using Boolean logic, sequentially. These operators are based on Boolean algebra, where the two (only) values are *TRUE* or *FALSE*, usually denoted by either 1 or 0. The prime operators are two-bit manipulations like conjunction (AND, $\wedge$), disjunction (OR, $\vee$), and the one-bit operation negation (NOT, $\neg$) – describing logical operations.

We need to remember that – while we do have seamless applications, powerful high-level programming languages and graphical operating systems, cell phones, internet, 3D printers, and control electronics of nearly every machine imaginable – under the hood, all programs are a long sequence of these fundamental

Table 26.1 Boolean logic table, with **irreversible** two-bit operations and the irreversible one-bit initialization operation (=0).

Input		Output			Input		Output
(2-bit)		(2-bit)			(1-bit)	(1-bit)	
x	y	x AND y	x OR y	x XOR y	x	NOT x	x = 0
0	0	0	0	0	0	1	0
0	1	0	1	1	–	–	–
1	0	0	1	1	1	0	0
1	1	1	1	0	–	–	–

The one-bit NOT operation is reversible.

logic operators acting on bits, managed in integrated micro-electronics that can operate at ambient conditions with more-or-less protection from the environment.

Digital information can be freely read and copied without destroying the original information state in the memory. And – given good error correction codes – it can be easily transmitted and stored. Even if you have DRM,¹ the digital copy is *identical* to the original.²

Worth noticing: Looking at the two-bit gates AND, OR (and XOR³), which are processing two logic inputs and yielding one logic output, we see that the sole knowledge of the output logic value is *not* sufficient to infer the logic value of the two inputs. So they are often addressed as *irreversible* logic gates.

This is the foundation of the “digital world,” from using your dish washer, doing online banking, having Zoom calls, programming machine learning algorithms to detect novel drug candidates...

26.2.2 Quantum

26.2.2.1 Refresher – Quantum Mechanics and Its Features

The features of quantum mechanics allow us to explain phenomena observed on atomic scale and below – which caused a *scientific* paradigm shift 120 years ago, and is still penetrating more and more disciplines. Reading the history of quantum physics is very enlightening, and there are wonderful books and talks, which we mention at the end in Section 26.7.

What are these features? Sadly, exactly those that are counter-intuitive to “classical” logic (i.e. local realism and determinism), see further below, and thus bring some novel challenges.

1 Digital Rights Management was introduced not to prevent the copying of information itself, but to control the capability to display the information on a specific device.
2 We will later see that copying *quantum* information is impossible [11].
3 The XOR can be written as a combination of AND, OR, and NOT.

26.2.2.2 |WTFUQC> – The $\hbar(|y\rangle + |o\rangle)$ pe

As we will see further below, the TRL⁴ of quantum computers is in a state where one can do interesting fundamental experiments, get familiar with concepts, and explore the physics of the novel devices (in real and simulated). On the other hand, they are not yet in a state to be useful for most (or any) applications. The community is in a STATE OF superposition between Waiting Till and Working Toward a Fully Useful Quantum Computer is available,⁵ and thus also in a superposition of hype and hope – so that with the novel computing paradigm we can address (in a novel way) problems that are so far (NP-) hard, if not impossible to compute. Therefore, the claim is to be a new paradigm. Use-cases and applications range from quantum chemistry on a quantum computer, material science or optimization and logistics problems, better Machine learning, and completely new approaches for information management-like shifting away from the statistic-driven approach of NLP⁶ to a conceptual understanding of whole sentences based on compositional category theory, see e.g. [13]. Consulting companies like bcg [14] and Accenture [15] provide reports on the potential value, and industry consortia like QUTAC [16] analyze the potential to use the new paradigm for use cases, which are relevant to their purpose [17] and develop solutions in partnership with subject matter experts.

26.2.3 Challenges

If we want to understand the **(Business) Value** of quantum computers we need to analyze not only their (business) *potentials* but also identify *needed capabilities*. This was done in a nice thesis by Riccardo Silvestri for the healthcare industry, based on open interviews with subject matter experts in healthcare research and development, as well as C-level executives, who shared their views [18]. As with every novel technology, its introduction into daily practice is dependent on (i) the understanding when and where to use it, (ii) its reliability or maturity, and (iii) acceptance, specifically in larger organizations where it needs to overcome (or tunnel through) the barrier of doubt.

26.2.3.1 Cognitive Challenge – Quantum Literacy

Quantum mechanics is (sometimes) counter-intuitive, and most of us are not “quantum native” and have not developed a playful instinct (or quantum literacy) [19] from early childhood on to understand what these “quantum features” are and where and how they can be used to formulate a problem “QC-compatible.” What do we mean by that? And why is that a problem?

Looking at the following picture (Figure 26.1) triggers our usual instinct to look if the picture [20] shows a rabbit **OR** a duck, but the interesting fact is that it can also be seen as a quantum *feature*, called superposition, namely it is a **mix of BOTH at the same time**.

⁴ Technology Readiness Level.

⁵ We propose to use the acronym WTFUQC.

⁶ Natural Language Processing, currently big statistical approaches exist, the largest known as LLMs (GPT-4, Claude 2, Llama 2, Orca, Cohere etc.) have $\mathcal{O}(10^{11})$ (=100 billion) parameters [12].

Cognitive Challenge – Quantum computing requires quantum literacy being able to think differently and formulate problems “quantum-like”



So far we are conditioned to ask:

Either
“Duck” OR “Rabbit”

Quantum Mechanics teaches us:

Superposition of
“Duck” and “Rabbit”

Figure 26.1 Cognitive challenge to identify the feature of a superposition. Source: Fliegende Blaetter Nr. 2465, page 147, 1892 [20]/CC-BY-SA 4.0.

The speedup for quantum computing comes only by using features of quantum information theory, formulating problems in a way that they can be solved with quantum algorithms is a fundamental new way of looking at these problems, to benefit, we need to learn how to use “quantum-ness” in order to start using quantum computing, and then get value out of it.

Useful quantum algorithms make use of the quantum specifics,⁷ just to mention them here

1. **Quantum = discrete energy levels:** There is no continuum, *nature is grainy*, we have a fundamental smallest amount of energy, denoted as h , Plancks *Wirkungsquantum*, often denoted as $\hbar = \frac{h}{2\pi}$, and in more advanced quantum theories also a minimal length (Planck length) and a shortest time unit (Planck time). The ground state of a system is often denoted as $|0\rangle$ in the so-called Dirac notation, higher excitation as $|1\rangle, |2\rangle, \dots, |n\rangle$.
2. **Superposition, “Schrödingers cat”:** Quantum systems have quantized Eigenstates (e.g. energy levels) and can be in a superposition of these, e.g. $|\Psi\rangle = \alpha|0\rangle + \beta|1\rangle$ – the most often used picture is that of a cat “half dead and half alive” – which was introduced by Erwin Schrödinger as a Gedankenexperiment in 1935 to illustrate the weakness of the Copenhagen interpretation of quantum mechanics if you apply quantum features of atomic particles to the macroscopic world 1 : 1 [21].
3. **Non local realism, entanglement:** What Einstein called *spukhafte Fernwirkung*, (spooky action at a distance) – describes that in quantum mechanics we can have an entangled pair of particles – which are independent of a classical physical connection, i.e. they can be separated over time and distance and still behave as one ensemble. In quantum computing, we will later hear of Bell states.

⁷ Today there are some algorithms that make use of these, we mention them further below in Section 26.3.6.

We can also generate entanglement of more than two particles, a feature often used in quantum computing.

4. **In-determinism, Born rule** [22]: The wave function gives a probability distribution, and repeated single observations produce as a statistical ensemble according to the distribution. Famous experiment is the refraction at a double slit, which causes an interference pattern, while the single electron is producing a single dot, but overall the distribution follows a wave-interference pattern [23]. The interested reader can check the Wikipedia article and find there interesting references to the original work as well as interpretations [24].
5. **Noncommutativity, the sequence of observations matter**: If you apply an operator (observation/measurement) to a quantum system, you collapse the system into one state and into the eigenvectors of that state (or Basis states), and when applying several operators, depending on the sequence you can get different results. Historically observed first by the Stern–Gerlach Experiment, a very famous Experiment from 1922 [25] where you measure the spin of a particle using an in-homogeneous magnetic field along a specific axis and the particle beam is showing not a homogeneous line but single dots, hinting that the magnetic moment/spin is not continuously distributed but indeed discrete, (i.e. quantized) along an (arbitrary) magnetic axis, and applying the Z-operator (=measuring along the z-axis) *determines* the state (say we measure $\uparrow$ along the z-axis we will always measure with 100% certainty again $\uparrow$ if we repeat the measurement at that beam). Further examining the separated beams shows that even if a particle is in a prepared pure Eigenstate along one axis (here $\uparrow$ at z-axis), while it will stay in that state during repeated measurement against the same Z-operator (axis), it is *undefined against another axis (say x-axis)*, so, an application of the X-operator (i.e. a measurement along the x-axis) gives in-deterministic results in this case (say $\leftarrow$ OR $\rightarrow$ on x-axis) as per Born-rule. So when measuring ZX (or XZ), the sequence of measurements (first x or first z) matters and thus also determines the final state. A beautiful description is given by Feynman in his famous quantum lectures [26]. This noncommutativity also leads to Heisenberg’s uncertainty principle [27], stating that you cannot measure all parameters describing a quantum state with infinite precision, but that the product of the measurements is always greater than the *Wirkungsquantum* $\hbar$.

I Don’t Understand How Quantum Works? Interestingly, after over 100 years since the first discoveries of quantum mechanics, there is still ongoing debate and controversy about the philosophical interpretation of these quantum features. Wikipedia shows us 13 different interpretations [28] – all have their strengths and weaknesses and some can be better applied than others depending on the context. For trying to explain the speedup of quantum computing, very often the “many-worlds” interpretation [29] is taken to explain the “parallel computation,” maybe to only place where this interpretation makes sense.

Bypassing the hugely interesting and enlightening philosophical discussions about the interpretations of quantum mechanics, the theory gives us a mathematical framework initially developed in 1932 by John von Neumann in his book

Mathematische Grundlagen der Quantenmechanik [30], later in 1939 expanded by Paul Dirac introducing *A new notation for quantum mechanics*, where a Bra-Ket notation is used to describe quantum mechanical states of a system as a vector in Hilbert Space [31].

We refrain from making too many interpretation statements that go back to the classical world (like “superposition means the system is in many states at the same time”) because they are eventually wrong or at least might be misleading, and in the end we need to use mathematics to describe a quantum system.

Independent from the chosen mathematical formalism, a state in quantum mechanics $|\psi\rangle$ can be described by an equation using complex numbers.

In (classical) physics, we are measuring and analyzing quantities like, e.g. position, momentum, or energy, a so called observable, which can be retrieved from the current state of a system. In *classical* physics, we can say this observation is a function F applied to a state S and returns a real value x – which corresponds to the measured quantity, e.g.

$$F(S) = x \quad (26.2)$$

If we measure the observable with is e.g. heat of a gas and want to have the Energy E of the system, the state (heat) S is characterized by the temperature T of the system and there exists a function F that allows us to calculate the energy, so that $F(T) = E$.

As introduced in 1926 by Max Born [22], in **quantum physics** an observable is not a function, but an **operator**, represented by a matrix O , which is acting on a vector (quantum state) $|\psi\rangle$. The matrix operator allows us to retrieve quantity from the system, whereas here the result is discrete (grainy). The eigenvalues λ_i of the matrix O are **the only** possible values the observable can take after being measured. The eigenvectors $|a_i\rangle$ can then be interpreted as states in which the system is left after measuring the associated eigenvalue λ_i . We write

$$O|\psi\rangle \rightarrow \lambda_i|a_i\rangle \quad (26.3)$$

where the arrow indicates the “collapse” of a state $|\psi\rangle$, which is in superposition to the state $|a_i\rangle$. The values reflect the probability distribution of the state.

Schrödingers Cat Applying this to our famous quantum feline, where we put (as a thought experiment) a cat in an nontransparent box together with a clever mechanism that kills the cat depending on the state of a radioactive atom, the state of the cat is acting as a drastic version of a Geiger counter, a measurement device whose output is dependent on the state of the atom.

This means: Because of the statistical (random) process of radioactive decay of a single atom, we cannot precisely know the state of the decayed atom, only the statistics over time if we observe many of them. The cat in the box is either dead (atom is decayed) or alive (atom is not decayed) – NOW because we cannot know the state of the atom unless we observe it, also we cannot know the state of the cat unless we open the box and observe it. That’s it.

It does NOT mean: That the cat *itself* is actually in a “zombie” state of half-alive-half-dead as often portrayed (ask the cat if in doubt).

26.2.3.2 Cultural Challenge

As we know from Kuhn, old habits die hard. Of course also sometimes for good reasons. In nearly all areas we have existing classical computing methods, software, hardware, data, knowledge, service providers, workforce... which are doing the job, and do so very powerfully.

“OK, But Does It Work At All?” When planning to introduce a novel tool (quantum computing), which is (i) based on a weird concept, (ii) is less mature conceptually and (iii) has a low TRL, the motivation and drive to test the technology AND to learn novel programming skills WHILE not having ready software NOR having mature hardware, is relatively low, and we have high internal “cultural” resistance. Additionally, the lack of clear benchmarks, which indicates when to use what is often prohibitive to find budget. This is why nearly all bigger companies have dedicated (“protected”) teams that can explore proof-of-concepts independent of the day-to-day struggle over projects, and do so in partnerships with peers, specific consultancies, start-ups and academia.

“...and Does It Scale ?” Another aspect that makes it difficult is that quantum computing has a complex road-map for its components with longer timelines, paired with uncertainties for each milestone. If we combine this with the drug development process, we quickly see that we get into an area where we have too much risk-parameters so that no classical return on investment (ROI) can be calculated.

Overall, all these are criteria that could qualify “Quantum Computing as a novel paradigm.” – The interpretation if it is so or not, is definitely observer dependent.

26.3 Quantum Computing Overview

While Quantum computing is totally different, we still have **some similarities to classical computing**:

- we have an information carrier (a qubit instead of bit),
- from a well-defined status A, we want to reach with computation a status B, which can be interpreted as result of the computation,
- you can get more reliability on the status B by error correction.

26.3.1 Quantum Simulators

Before investigating quantum *computers*, we like to mention that there is a more general approach, called quantum *simulators*. These quantum simulators are what Feynman initially meant by using quantum systems to emulate quantum systems. These simulators cover a wide range on the spectrum of quantum devices from specialized quantum experiments to universal quantum computers. These quantum devices utilize quantum *features* like entanglement and many-particle behaviors to explore and solve hard scientific, engineering, and computational problems [32]. In other words, quantum simulators provide an alternative route to understanding the properties of complex quantum mechanical systems, and via

the simulation, they create clean realizations of specific systems of interest, which allow precise realizations of their properties. A recent overview [33] mentions different applications of quantum simulators and computers for molecular biology and explores how quantum computation may improve the practical applications of the quantum foundations of molecular biology by providing computational benefits for simulations of biomolecules. They show how quantum computation can be used to solve both traditional quantum mechanical problems related to the electronic structure of biomolecules, as well as classical problems such as protein folding and drug design as well as consider how data-driven approaches in bioinformatics may be enhanced by quantum simulation and quantum computation.

26.3.2 Embedding Quantum into Computing

A quantum computer will not stand alone, we think of it as a co-processor for specific problems, like a GPU⁸ or TPU⁹ does for gaming or artificial intelligence/classical Machine Learning, a QPU¹⁰ will accelerate the computation of dedicated problems. The challenge is not only in formulating algorithms but also the interface to data loading and the encoding of data in qubits (Figure 26.2).

26.3.3 What's the New Idea?

We need to mention here that we have two different flavors of quantum “computing” that use different aspects of quantum mechanics, namely *gate-based* quantum computing and *quantum annealing*. There is controversy if the latter is correctly called “computing” – but we do not need to concern ourselves with that. We will look at them later in a bit more detail. In quantum (gate) computing, we are using the *specific features* of quantum mechanics to encode information, perform calculations using specific unitary (reversible) operations, and then measure the system (multiple times) and interpret the output. We deal here with a completely different kind of logic and architecture – means we have a non-Von-Neumann architecture.

From looking back at quantum mechanics, we are in a lucky position that in quantum *computing* and introducing the quantum-bit (*qubit*) in quantum information theory, we deal with the easiest system, namely one with only two states, so the vector describing its state $|\psi\rangle$ has only two complex variables α and β (or sometimes also used c_1 and c_2).¹¹

26.3.4 Introducing the Concept of the Qubit

The definition of a qubit lies in the center of quantum computation theory. To understand the fundamental power of the new concept, let us first briefly recall

⁸ Graphics Processing Unit.

⁹ Tensor Processing Unit.

¹⁰ Quantum Processing Unit.

¹¹ A brief overview of possible physical realization of such system is described later in Section 26.3.9, where we will also see the specific features and technical challenges of each.

The journey of a QUANTUM ALGORITHM

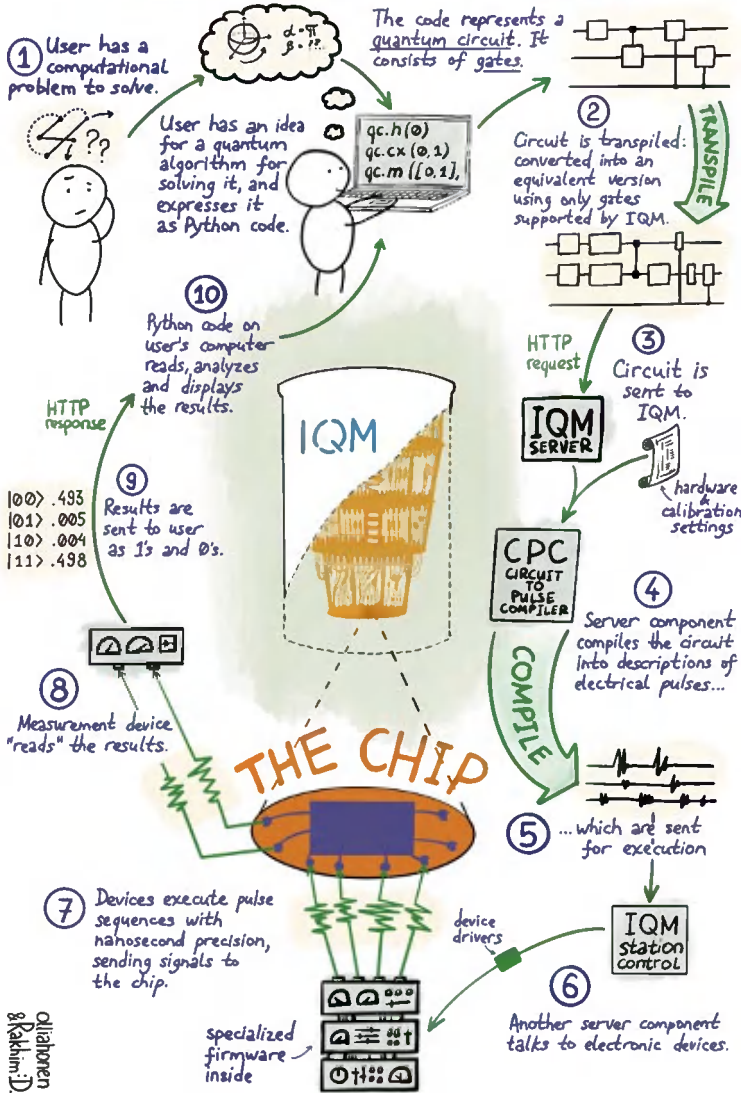


Figure 26.2 The journey of a quantum algorithm. Source: Reproduced with permission from [34]/IQM Quantum Computers.

the definition of classical bit from above. A **classical** bit is an unit of information, which describes a two-dimensional classical system. Thus, the classical system could be either in the state:

$$|0\rangle = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \text{ or in state } |1\rangle = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad (26.4)$$



Figure 26.3 A quantum smiley.
Source: Adapted from [35].

The physical representation of a bit is two *flip-flop* states representation, for instance, two distinct voltages of electric circuit or two distinct levels of light intensity. The state is defined, and measuring it does not change the state. This is sufficient for classical physics, and this is how the classical computer works.

The quantum computer uses the effects of quantum mechanics, such as a superposition of states. A qubit is an unit of information, which describes a two-dimensional **quantum** system, and the general state of a qubit is represented by a pair of complex numbers (α, β) and a set of basis vectors – typically, one chooses the vectors $|0\rangle, |1\rangle$ from above.

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle = \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \quad (26.5)$$

26.3.4.1 Superposition

The difference between a bit and a qubit is now that, while a classical bit can only be in one state, the qubit can be in any combination (or superposition) of the two states. If we recall the bit-strings from Eq. (26.1), where one 16-bit string encoded one type of information;) or 8), a 16-qubit string can encode 2^{16} “characters”, one of which is shown in Figure 26.3 where we can see that special character showing the superposition as an overlay of both corresponding classical strings.¹²

The powerful specific features of quantum mechanics now come into play because the state vector of our qubit can be a superposition of these two states, where the factors α and β are complex numbers, and the sum of the squares equals to one.

26.3.4.2 The Bloch Sphere

We can also describe the 2D-vector quantum state encoded by a qubit as an arrow pointing to the surface of a sphere with radius 1 – the so-called Bloch sphere and its polar coordinates ϕ and ϑ (see Figure 26.4)

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle, \text{ with } \|\alpha\|^2 + \|\beta\|^2 = 1 \quad (26.6)$$

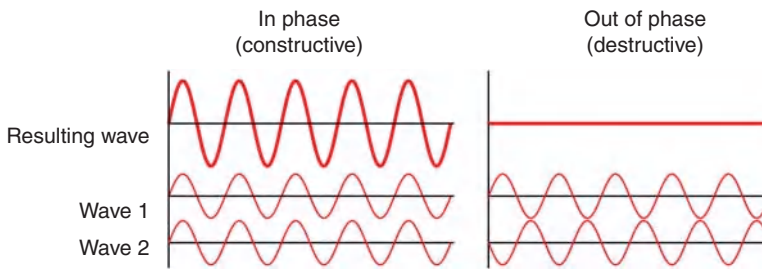
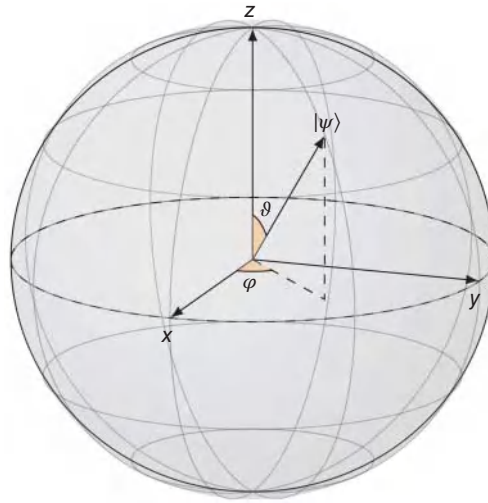
$$\alpha = \cos \frac{\vartheta}{2}, \text{ and } \beta = e^{i\phi} \sin \frac{\vartheta}{2} \quad (26.7)$$

$$\text{where } \phi \in [0, 2\pi] \text{ and } \vartheta \in [0, \pi] \quad (26.8)$$

26.3.4.3 Interference

Because the state is described by *complex* amplitudes and the square of the amplitudes represents the probability of the state, we have the possibility of *negative* probabilities, and therefore can use this to reduce the probability of unwanted states and enhance the probability of wanted states. This is the fundamental principle of

¹² In the smiley example, we bring the qubit q_8 in a 50 : 50 superposition of $|0\rangle$ and $|1\rangle$, and assure that it has the same value as the qubit at position q_9 by entangling both.

Figure 26.4 Bloch sphere.**Figure 26.5** Interference illustrated. Source: Adapted from [36].

gate quantum computing, resulting from the wave-like description. The wave-like interference as illustrated in Figure 26.5 is fully explained by the presence of complex numbers in probability amplitudes.

Example The probabilities in real numbers when added are always greater or equal: $p_1 + p_2 \geq p_1$ and $p_1 + p_2 \geq p_2$. The complex amplitudes when squared are also real, but now the addition of complex numbers $|c_1 + c_2|^2$ can increase or decrease the probability. The probability amplitude $c_1 = \frac{i}{2}$ when squared is equal to probability $|c_1|^2 = \frac{1}{4}$. The probability amplitude $c_2 = \frac{-i}{2}$ when squared is also equal to probability $|c_2|^2 = \frac{1}{4}$, however, the sum of probability amplitudes $c_1 + c_2$ yields probability $|c_1 + c_2|^2 = |\frac{-i+i}{2}|^2 = 0$, which is certainly lower. The complex numbers can cancel or overlay each other, which has a physical meaning of interference. This is the core of quantum mechanics, allowing to explain wave-like behavior of particles.

26.3.4.4 Nondeterminism

Remember, the quantum state is a probability distribution, generally a qubit is not in a defined Eigenstate (where either α or β are zero), and we can only know its

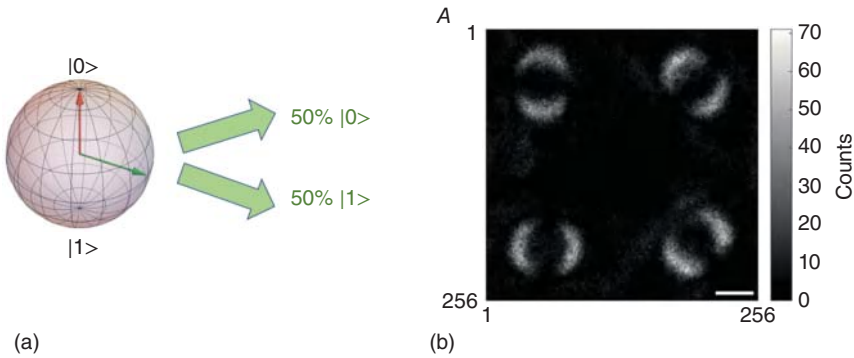


Figure 26.6 Illustration of (a) 50/50 superposition (green vector) between state $|0\rangle$ and $|1\rangle$ and (b) Imaging entanglement as a bell-type nonlocal behavior. Source: [11] PAUL-ANTOINE MOREAU 2019/American Association for the Advancement of Science/CC BY 4.0.

state after a measurement. We measure in a set of basis vectors,¹³ the measurement results in a random(!) value depending on the probability distribution for this basis, and once we measure, the feature of superposition is gone.

26.3.4.5 Entanglement

We will soon see in Section 26.3.7 how entanglement is generated in a quantum computer. It is a phenomenon of two “particles,” where the value of q_0 is dependent on (opposite of) q_1 , independent from the time or space where the measurement is done. This is mathematically expressed via a tensor product. Important again, the values are random, but correlated. Example is an entangled pair of spin-particles where the overall spin is zero, and one is always spinning up, the other down.

26.3.4.6 Multi-particle Registers

Assume we want to analyze multi-particle states instead of only one particle states. The machinery that should be used to accomplish that is called tensor product of state spaces, and the procedure is called assembling of quantum states. Having k independent particle states: $\{|\psi_1\rangle, |\psi_2\rangle, \dots, |\psi_k\rangle\}$ we can describe them by general state Ψ :

$$|\Psi\rangle = |\psi_1\rangle \otimes |\psi_2\rangle \otimes \dots \otimes |\psi_k\rangle = |\psi_1 \psi_2 \dots \psi_k\rangle \quad (26.9)$$

Thus, if $|\psi_i\rangle$ are n -dimensional vectors for each i , then the state $|\Psi\rangle$ will have n^k dimensions. Remember: Because the state is described by *complex* amplitudes and the square of the amplitudes represents the probability of the state, we have the possibility of *negative* probabilities, and therefore can use this to reduce the probability of unwanted states and enhance the probability of wanted states. This is the fundamental principle of gate quantum computing, resulting of the wave-like description. The wave-like interference is fully explained by the presence of complex numbers in probability amplitudes.

¹³ For more mathematical details please refer to standard textbooks.

Box 26.1 Qubit features

Qubits are the building blocks of quantum computing and have the following features.

- **Indeterminism:** A qubit is a quantum system and thus can only be measured in its Eigenstates of the selected Basis. This means if a qubit is not in an Eigenstate, we need to perform many measurements to get the statistical distribution.
- **Superposition:** Shown here in Figure 26.6a is the application of a Hadamard gate (see Eq. (26.10) in the Basis $|0\rangle, |1\rangle$, to bring the initial $|0\rangle$ (red vector) into a 50/50 superposition (green vector) $H|0\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$. After measurement we expect 50/50 distribution of qubits measured in either state.
- **Entanglement:** e.g. in Bell State $|q_1q_0\rangle = |\Phi_+\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$ we see that it is a superposition of two qubits states with the same value for q_1 and q_0 , i.e. $|00\rangle$ and $|11\rangle$, which means as both qubits have the same value, by measuring one qubit q_0 , e.g. in $|0\rangle$ we know that the other q_1 must be in the same state. In a Bell state we measure either both in $|0\rangle$ or both in $|1\rangle$ – and in this case with a probability of 50%.

26.3.5 Quantum Computing Stack

A full quantum computing stack requires several components, and each part addresses different needs and requires different expertise, as illustrated in Figures 26.2 and 26.7 and also later in the hardware Section 26.3.9.

- **User level:** The user typically uses a high level language like Python and modules to encode the software/application, including the quantum algorithm and data-readouts – often together with classical algorithm parts in a so-called hybrid mode.
- **Logical layer:** The quantum algorithm is encoded in logical gates, and in the next step has to be *transpiled* according to the instruction set that can be understood by the processor's limited basic gate set, data preparation and readout are done. Also, circuit simplification is done at this stage, often a complex-looking circuit can be reduced to fewer gates using, e.g. ZX calculus¹⁴ [37].
- **Compilation:** Here the final circuit, together with machine specific instruction sets is compiled to a series of control signals, which manipulate the quantum chip computation. We have vast differences depending on the underlying chosen approach to build a qubit physically – see later in Section 26.3.9. Error correction to translate the logical qubits into redundant physical qubits as well as error mitigation (where possible) is done here. Also here we see drastic differences depending on the chosen topology of the connectivity of the qubits on the hardware.

¹⁴ A powerful diagrammatic tool to simplify multi-qubit gates.

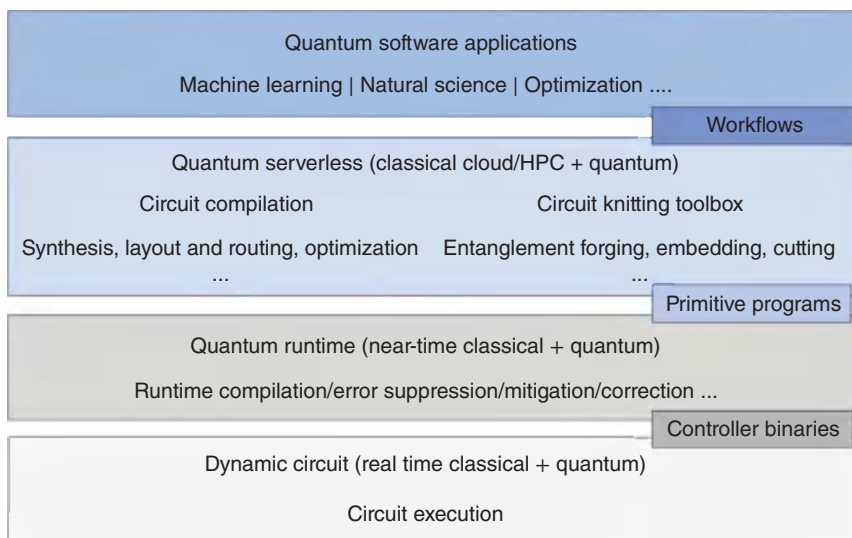


Figure 26.7 Overview of a quantum stack. Source: Reproduced from [38]/with permission of AIP Publishing.

- **Physical control:** The circuit execution is done on the level of physical qubits, which will be manipulated using analog control signals (e.g. electro magnetic [EM] waves in the GHz range) to manipulate the qubits states (e.g. rotating the spin-vector).
- **Physical encoding:** Interaction of the chip with control signals (modification and measurement).

We will see that overall, quantum computing is often close to orchestrating a choreography of (hidden)¹⁵ dancers or can be compared to throwing stones into a pond in a very sophisticated way and interpreting the resulting wave pattern.

26.3.6 Major Applications of Quantum Computing

We have seen that the new tool “Quantum computing” promises to overcome computational limitations with better and faster solutions. From an application standpoint several interest groups or consortia are emerging and working together to either address domain-specific questions like the Pharma-specific Quantum Computing Community of Interest (QuPharm, Pistoia Alliance, QED-C, and QPARC) [39] or, as mentioned earlier, form industry consortia that address similar problem classes, but independent from the respective industry, like QUTAC [16], [17]. Before we dive into the specifics of relevance for drug discovery and development, let us have a brief look at some famous algorithms (on a gate computer) – which are also the reason why quantum gets so much attention, and can be educational for

¹⁵ During operations one must not “observe,” the system must not interact with its environment in order to avoid decoherence.

exploring the needed steps in terms of hardware maturity. Then we look into the domains of optimization, simulation, and machine learning problems.

26.3.6.1 Classes of Applications

In Quantum Computing we see mainly three different classes of potential applications, and we will explore them in the following.

- **Simulations – e.g. Doing quantum chemistry with quantum computers:** Because we can encode 2^n states of information with n qubits, we are able to explore system sizes that are otherwise (classically) not realistically computable.
- **Optimization:** There is experimental evidence that optimization problems can be well implemented on quantum annealing machines and that for very complex models, the hybrid/quantum approach to solve is at least as good as implementing classical solvers on classical machines. The problem class that works well is QUBO (Quadratic Unconstrained Binary Optimization), which is very similar to formulating the problem as Ising model.¹⁶ This is a problem known to be *NP*-hard. A very nice overview of what problems could qualify to be implemented on adiabatic quantum optimization (annealing) can be found e.g. here [40]. On a gate computer, you can in principle do the same, but then use quantum approximate optimization algorithm (QAOA) algorithms.
- **Quantum machine learning:** A promising area is quantum machine learning (QML), and we have several domains here and will explore them later in Section 26.4.

26.3.6.2 Famous Quantum Algorithms (Gate)

1. **Shor:** e.g. for prime number factorization One of the most famous algorithms is the one from Peter Shor [41], from 2006, which is using the quantum features (calculating with a superposition of states) for period-finding in a hybrid classical/quantum algorithm. Why is it famous? Because it gives a *mathematical proof* that one can factor prime numbers exponentially faster (in theory) than using classical algorithms, and prime number factorization is the fundamental mechanism for RSA encryption. A nice explanation can be found in a recent blog here [42].
2. **Grover:** e.g. reverse telephone book search Another famous algorithm was developed by Lov Grover already 1996 [43] – A quantum algorithm solving such a problem is Grover’s algorithm, which finds an element in an unordered set faster than any classical search algorithm, in his words “Imagine a phone directory containing N names arranged in completely random order. In order to find someone’s phone number with a 50% probability, any classical algorithm (whether deterministic or probabilistic) will need to look at a minimum of $N/2$ names. Quantum mechanical systems can be in a superposition of states and simultaneously examine multiple names. By properly adjusting the phases of various operations,

¹⁶ The Ising Model is a mathematical model that doesn’t correspond to an actual physical system. It’s a huge (square) lattice of sites, where each site can be in one of two states.

successful computations reinforce each other while others interfere randomly. As a result, the desired phone number can be obtained in only $\mathcal{O}(\sqrt{N})$ steps. The algorithm is within a small constant factor of the fastest possible quantum mechanical algorithm.”

Advantage/Supremacy Another aspect that caused lot of momentum in the development of quantum computing was the race to what was initially called “supremacy,” an outdated term now replaced with “advantage”, i.e. defining a problem where a quantum computer would outperform any form of classical computers. Maybe of most relevance to this audience is the hope that one finds exponential speedup for solving chemical problems, most often in finding the ground state of a molecule. In a very recent paper it has been shown that so far there is no clear evidence that this might be the case [44].

There is even a class of specific experiments designed to proof the superiority of quantum algorithms over classical computing, most of them are huge fantastic efforts of engineering, with results of academic value, but no real practical application. Typically they are designed to work where quantum is good and classical is bad – namely in the generating and evaluation of random numbers. The focus of the community is meanwhile to identify proof of concepts that can add value to specific sub-processes and could at least in theory scale – under the condition of existing hardware.

26.3.7 Gate Quantum Computing

The idea of quantum computers appeared in the 1980s. One approach, known as gate model, expresses the interactions between qubits and quantum gates, specific operations/control sequences to modify the state of the qubits. As we have seen, classical computers can be described in terms of Boolean gates – for example, AND, OR, NOT gates. This means that (generally electric) signals arrive at the gates, and the signals coming out of the gates go into different gates, at the design of the processor architect and dependent on the algorithm/program you wrote.

Quantum gates need to be different because while in quantum mechanics we can *teleport* information, it *can not be copied* – we have the so-called “no-clone” theorem [45], going back to the uncertainty principle from Heisenberg [27]. Therefore, operations need to be *reversible*, and so there is no AND gate¹⁷ in a gate model quantum computer. We achieve this reversibility by unitary transformation, which acts as gates.

Instead, there are gates with names such as Hadamard gates and Toffoli gates – which are unitary transformations to the vector, and unlike many classical logic gates, quantum logic gates are (have to be) reversible. Mathematically you multiply a unitary matrix with a state vector.

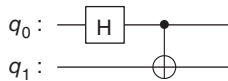
A very nice “explorer” of gate quantum computing and its notation is, e.g. quirk[46].

17 Recall from Table 26.1 that the AND is irreversible.

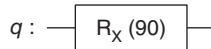
A gate that acts on n qubits is logically represented by a $2^n \times 2^n$ unitary matrix and can geometrically be interpreted as higher dimensional rotation of the vector.¹⁸

This means we have 2×2 matrix one-qubit gates (X, H, other rotation around either axis of the Bloch sphere), as well as 4×4 matrix two-qubit gates (e.g. CNOT) – see further below Eq. (26.12), or an 8×8 matrix for three-qubit gates (e.g. Toffoli gate, or also the controlled-NOT-NOT and Fredkin gate) – the latter are relevant for efficient error-correction codes.

The magic of quantum computing happens as soon as you introduce *multi*-qubit gates, by bringing one (or more) qubit into a superposition and then use this as control for another gate – e.g. taking the one-qubit gate Hadamard on qubit zero ($|q_0\rangle$), and use that as control for the two-qubit gate CNOT.



The other building block of quantum computation is phase shifts, namely rotation of the vector with a certain angle around one axis, e.g. a 90° rotation around the x -axis is denoted by R_x , also known as $\sqrt{X}$ gate. In variational (or parametrized) algorithms one works with a set of parametrized angles θ .



Box 26.2 Gate Computing – The Principle

- The general approach to use *gate*- quantum computing is: you create a quantum-data (e.g. via superposition), you entangle and compute in superposition (amplitude modification, phase modification) and then readout the data in a statistical measurement process along one axis (typically z -Axis), “collapsing” the quantum data to a **classical** bit.
- Quantum gates are rotations of the state vector and thus reversible.
- Quantum algorithm look like a musical score, where each line represents the time for a qubit and specific gates for this qubit are noted, if they span multiple qubits the gates span multiple lines. There are several programming languages, illustrated in Figure 26.8 is the one from IBM (qiskit)

26.3.7.1 Example: The Hadamard Gate

Is the gate for creating a 50 : 50 superposition of $|0\rangle$ and $|1\rangle$,

$$\hat{H} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \quad (26.10)$$

¹⁸ Rotations are reversible.

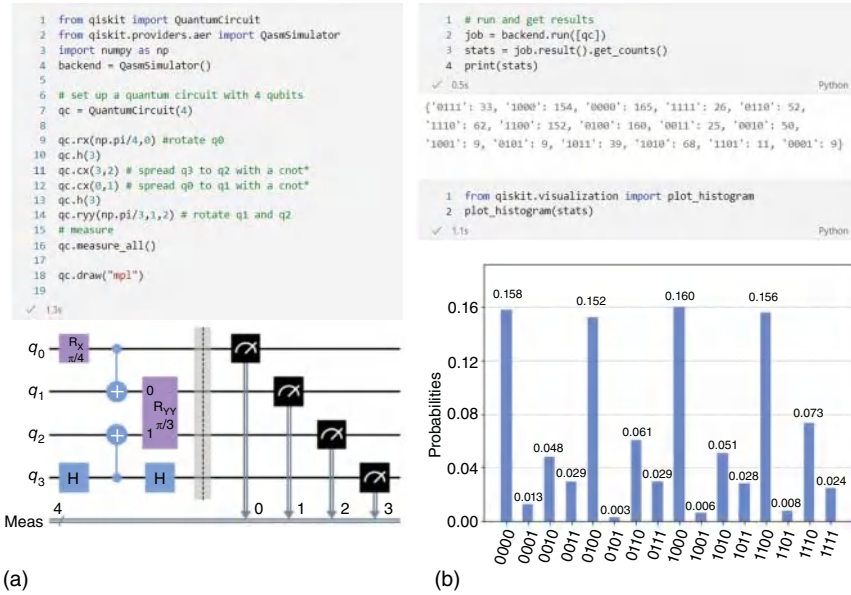


Figure 26.8 Example for Gate computing algorithm, we leave the interpretation of the results to the reader. (a) A four qubit circuit and (b) resulting statistics.

The Hadamard gate transforms the standard basis (or z basis) to the computational basis (or x basis) and back, $\hat{H}|0\rangle = |+\rangle$, $\hat{H}|+\rangle = |0\rangle$, and $\hat{H}|1\rangle = |-\rangle$, $\hat{H}|-\rangle = |1\rangle$. where $|+\rangle = \frac{|0\rangle + |1\rangle}{\sqrt{2}}$

26.3.7.2 Example: The CNOT Gate

We have seen that in quantum computing the operations must be reversible, this means a gate must have the same number of input qubits as output qubits. For two qubits, the only nontrivial two-qubit gate is the CNOT gate, for flipping a target qubit q_t in dependency of the control qubit q_c . In qiskit as example this is written as `circuit.cx(c, t)` which applies the CNOT gate with qubit q_c as the control qubit and qubit q_t as the target qubit in the quantum circuit “circuit”. More generally, the inputs are allowed to be a linear superposition

$$a|00\rangle + b|01\rangle + c|10\rangle + d|11\rangle \Rightarrow a|00\rangle + b|01\rangle + c|11\rangle + d|10\rangle \quad (26.11)$$

The action of the CNOT gate can be represented by a 4×4 matrix

$$CNOT = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \quad (26.12)$$

Constructing the Bell state $|q_1 q_0\rangle = |\Phi_+\rangle = CNOT|0\rangle \otimes H|0\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$.

If we use two control qubits to construct a controlled-controlled-NOT or CCNOT gate, this is then called Toffoli gate, for more examples and information please refer to the quantum information introduction textbooks like e.g. [47].

26.3.7.3 Development Kits

Several companies like IBM (qiskit) [48], Google (cirq) [49], Microsoft (Q#) [50], Amazon (Braket) [51], Rigetti (Forest) [52], Xanadu (PennyLane and Strawberry Fields) [53] or Quantinuum (tket) [54] have developed a framework to do quantum computing on different back-ends (real quantum machines or also simulated on your notebook). Meanwhile, other specific programming packages and full-stack libraries are available mainly for python 3.8 and higher [55]. Additional approaches include tools to also emulate the interaction between algorithms and specific hardware, like used e.g. by quEra (bloqade) [56].

26.3.8 Adiabatic/Annealing

A alternative model using other quantum features is the so-called quantum annealing, or adiabatic quantum computing, a form of analog computing. While annealing is a technology known from metallurgy describing the heating of metal and then a slow cooling, it is also used as terminology for a similar approach (“heat” and “cool”) translated to the quantum realm. Maybe the best reference to learn about quantum annealing is directly at D-Wave – the company that has a long-standing track record developing quantum annealers and using them to solve mainly optimization problems like knapsack or traveling salesperson, or generally problems that can be formulated as a QUBO¹⁹ [57]. D-Waves development kit is called *Ocean* [58]. Usecases of relevance to the audience here include also exploratory work of designing peptides on a quantum computer [59] and will be shown in more detail in the Section 26.5. It has been mathematically shown that the adiabatic model is equivalent to the gate model, meaning it can do anything the gate model can do, speeds may differ. Recent work describes how to use tensor network algorithms to optimize quantum circuits for adiabatic quantum computing [60].

26.3.8.1 Intuitive Explanation

For the power of adiabatic computing the potential mechanism is that because all the qubits are coupled/entangled at the beginning of the optimization process, all possible states are explored at once, recall that the number of states that can be explored with N qubits at once is 2^N (where a traditional computer has $2 \cdot N$ bits), which brings us back to the potential for exponential speed-up of quantum computers.

The physical principle behind quantum annealing is that if a quantum state is initiated into the ground state (lowest energy) of a system, then any small enough changes to that system will keep it in the ground state. The process is as follows

- Initialize the state into ground state of an easy-to-implement initial system
- Design a target system where the ground state encodes the solution to some interesting problem
- Transition from the initial system to the target system slowly enough that all changes are small, and end up in the ground state of the target system
- Read out the final state, which was the ground state that was constructed for the solution of your problem

¹⁹ Quadratic unconstrained binary optimization.

26.3.8.2 A Brief Explanation

As anneal begins, the system starts in the lowest energy state (well separated) from any other energy level. As the problem Hamiltonian is introduced, other energy levels may get closer to the ground state. The closer they get, the higher the probability that the system will jump from the lowest energy state into one of the excited states. There is a point during the anneal where the first excited state – that with the lowest energy apart from the ground state – approaches the ground state closely and then diverges away again. The minimum distance between the ground state and the first excited state throughout any point in the anneal is called the minimum gap.

26.3.8.3 In Summary

The systems start with a set of qubits, each in a superposition state of $|0\rangle$ and $|1\rangle$, when they undergo quantum annealing, the couplers and biases are introduced and the qubits become entangled. At this point, the system is in an entangled state of many possible answers. By the end of the anneal, each qubit is in a classical state that represents the minimum energy state of the problem, or one very close to it.

26.3.8.4 Digital Annealer

Fujitsu is offering a “quantum inspired” approach where classical digital processors are used to emulate the annealing process. Using a digital circuit design inspired by quantum phenomena, the digital annealer has a high $N-N$ connectivity and focuses on rapidly solving complex combinatorial optimization problems without the added complications typically associated with quantum computing methods. There are some interesting papers of relevance to the exploration of a large chemical space, e.g. [61] and for *de-novo* drug design [62].

26.3.9 Hardware

Building a quantum computer does not only entail finding the right physical implementation for qubits but also includes building the entire stack around it as illustrated in Figure 26.9. Efforts to improve semiconductor materials and processing steps for qubit manufacturing are ongoing. A major research area is the development of new quantum error correction methods, both using software and hardware approaches. Further system approaches include the development of better control systems (lasers, photonics systems microwave technologies, detectors) and cryogenic (cooling) systems. Lastly, the development of hardware-specific software and algorithm is required to complete the full stack quantum computer.

There is a wide range of different physical implementations for a qubit, using either natural systems or artificial systems. For our purpose, it suffices to know an overview and potential advantages for each, also due to the different potential topologies for implementing a resulting QPU and its control electronics.

Typically, most of the qubits need to be operated at extreme cold temperatures (20 mK, while some can operate at 3 K), and the challenge is the control technology in these environments, as you must keep the system cool and at the same time radiate heat via the control electronics.

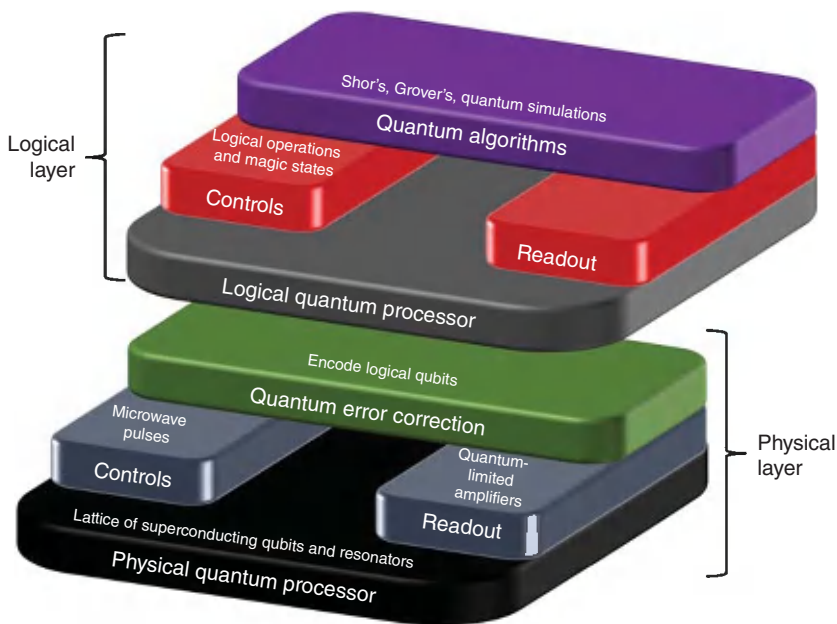


Figure 26.9 Building blocks of a quantum computer. Source: Reproduced from [63]/Springer Nature/Public Domain CC BY 4.0.

- natural qubits** are physical systems occurring in nature and thus are intrinsically equal in their properties, like energy levels, which are used to represent the ground state and excited states.
 - Quantum dots/electron spin** makes use of the natural $s = 1/2$ spin particles, either by using quantum dots and specific dotting of the Si material (most with Phosphor atoms) or a modified CMOS technology (SiMOS) where only one electron is in the area between source and drain. The modification of qubits is done via electromagnetic control pulse waves in the MHz range. The operating temperature can be in the lower Kelvin (3 K) [64].
 - Vacancy in diamond:** Here an artificial diamond is produced via vapor deposition, at certain layers doted with Nitrogen, Silicium, or Germanium, which creates an electron vacancy with spin $m = 1$. Excitation is done via laser light, readout is optical, and control is via microwaves and magnetic fields. The electron-vacancy spin can couple with the nuclear spin of the ^{13}C diamonds in the diamond and stabilizes the spin. Several institutes are in the progress of building demonstrators [65].
 - Ions:** Here single atoms (typically Barium, Ytterbium, but also Calcium) are cooled, ionized, and trapped electronically, and excited with Lasers of different frequencies. The ions are nonstationary, this means they can be transported and brought to interact with other ions far away with swap operations. Players are IonQ, AQT, and Quantinuum as well as some Universities. Also, the system can operate at room temperature but requires ultrahigh vacuum.

- **Neutral atoms:** Here neutral atoms, typically Rubidium are trapped optically and excited to a Rydberg state. Also, the natural atoms are freely movable and allow for N -to- M coupling. Players are QuEra, PASQAL, AtomComputing and ColdQuanta. Except for the handling of the atoms (via Laser trap instead of electronic fields) they have same operations as the ions [66].
- **Photons:** There are two different applications, either to use single photons and bring them to interference via classical mirrors (Universities) and/or Mach-Zehnder devices (PsiQuantum), or squeezed states of multiple photons with delay loops (some kind of optical Si wave guides) to generate superposition in time (Xanadu) – or a mix of both.
- **artificial qubits** are engineered and have thus resulting imperfections or a *personality* that varies from each individual qubit to qubit and making them more flexible but also more fragile and difficult to control
 - **Superconducting artificial atoms/transmons:** The most commonly used type of qubit used by IBM, Google, and Rigetti. These are resonant circuits with a tunnel junction (Al/AlO_x), excitation is with microwave and resonators. Typically, here we have also only a next-neighbor connectivity, while ideas to overcome this are in the roadmap of the providers.
 - **Quasi-particle/topological:** An approach to use quasi-particles²⁰ for the encoding of quantum information. One idea is to use quasi-particles called Anyons with fractal spin. So far technical realization has not been proven.

26.3.10 Technological Challenges

Most algorithms require perfect logical (fully error-corrected qubits), which then require several hundred physical qubits for one logical qubit. Currently, the technology is still in the so-called NISQ state [67], which stands for Noisy Intermediate Scale Quantum Systems, which means we have approximately 100 physical qubits. A recent record was presented in November 2022 by IBM with their *System Two* and a 433 qubit architecture. Additional challenge for the artificial qubits, specifically the transmons, is the need for precise tuning of the individual qubit frequency. The transmon system is a nonharmonic oscillator, and the qubits are ground state and first excited state. Due to the deformation of the harmonic oscillator potential one can avoid the excitation energy raises the qubits from first to second excitation, the control of these MHz signal pulses for many qubits needs very precise and fast, tune-able microwave generators.

26.3.10.1 Errors

As we have seen in Sections 26.2.2 and 26.2.3.1, quantum systems must not be disturbed. We have several sources for errors in the quantum computer, namely (i) interaction of the qubit with its environment, and (ii) the fidelity and precision of the gates (and its control electronics) that are doing the operations on the state-vector. So quantum error correction is a must, and also a very active research field [68].

²⁰ Quasi particles are macro states like, e.g. the movement of “La Ola” in a mass of people in a stadium.

For the first case, we have two possibilities, namely a “bit-flip,” i.e. a change from $|0\rangle \leftrightarrow |1\rangle$, and on the other hand also a phase flip, where the vector rotates around one of the axes. The most crucial efforts these days are to elaborate error correction algorithms [69] – with the challenge of the no-cloning theorem, one needs to work with several auxiliary qubits that are entangled with the information-carrying qubit. When measuring the ancilla qubits it is possible to correct the information of the qubit without reading the qubit itself; however, the overhead is very large and consumes up to 90% of gates. It is assumed to have a ratio of 1 : 13 (nine physical plus four ancilla) [70] or more for error-corrected logical qubits.

Knowing that the gates themselves have also errors in the order of 1% for two-qubit gates (like SWAP) we see the needed effort to improve control electronics. There are some topologies that make error correction easier because they allow to entangle more qubits, but on the other hand, these are slower in the execution.

26.3.10.2 Scalability

Another difficulty is the scalability of the NISQ demonstrators to several billion qubits, here the sheer size of cryostats needed is a significant challenge for most architectures that operate at ultralow temperatures. There are some approaches that recently came out of Stealth and promise to scale to billions of qubits, (photonic [both PsiQuantum²¹ as well as Xanadu²²]) and SiMOS (DiraQ)²³ which can operate above Millikelvin, and there is still hope of topological qubits, which would be insensitive to temperatures. Other approaches are to integrate the classical readout, control, error correction, and data processing functions within a quantum processor in the cooled zone (Seeqc²⁴). We have also seen a new “unimon” presented by IQM, which comes with fidelities up to 99.99% [71]. A recent interesting approach is to combine silicon photonic and silicon spin (Photonic) [72]. The race is open and likely the final word is not yet spoken.

26.3.10.3 Connctivity

Finally, the quantum algorithms draw their power from entanglement and superposition between many qubits. Most of the current architectures have only very few “cheap” possibilities to connect one dedicated qubit q_i to any and many q_j other qubits, which result in very costly SWAP operations.

Equipped with this knowledge we can now dive into the application of quantum computers for computational drug discovery, while a recent work from Riverlane gives perspective on potential advantage [73].

26.3.11 Quantum Computer for Molecular Biology

Molecular biology is concerned with the structures and functions of molecules in living organisms, and these processes can be understood in terms of the principles of chemistry and molecular physics. These principles are governed by quantum

²¹ www.psiquantum.com.

²² <https://www.xanadu.ai>.

²³ <https://diraq.com>.

²⁴ <https://seeqc.com>.

electrodynamics, which describes the electromagnetic interactions of electrons and atomic nuclei in and between molecules.

Quantum many-particle theory and computational solution methods have been developed to predict the energies and structures of molecules with high accuracy, but these methods become less accurate as the size of the molecular system increases. As a result, it can be challenging to accurately simulate and understand the behavior of larger bio-molecules, such as proteins and nucleic acids, using traditional quantum chemical approaches. To address this limitation, as we see throughout this book, researchers have developed various approximations and classical mechanical models to study the behavior of these larger systems. However, these models may not always accurately capture the full complexity of the quantum mechanical interactions, and there is ongoing research into developing quantum algorithms that can be used to simulate and understand the behavior of larger bio-molecules.

A significant body of work has emerged in recent years optimizing quantum algorithms for chemistry, both for fault-tolerant quantum computers and current NISQ (Noisy Intermediate-Scale Quantum) devices [67], including various experimental implementations [44, 74–79] and [80–82].

Current developments in optimizing quantum algorithms for chemistry have a focus on estimating the energies of the ground states of electronic structures. These ground-state energies are a fundamental property that can be extracted from a quantum chemistry simulation, but they are not typically measured in the laboratory. Therefore, there is a need for quantum algorithms that can estimate other properties of interest, such as those relevant to industry. The goal of the quantum computing community is to make progress toward larger NISQ devices or future fault-tolerant quantum computers that can perform more complex quantum chemistry simulations.

In a recent work by Google and Boehringer Ingelheim [83] the authors investigate the efficient quantum computation of molecular forces and other energy gradients and introduce new quantum algorithms for computing molecular energy derivatives with significantly lower complexity than prior methods as also illustrated in Figure 26.10.

One common challenge in developing these algorithms is the need to map the representation of electrons onto the qubit structure of the quantum computer, which is often done using the Jordan–Wigner or Bravyi–Kitaev transform. There is a drive to design algorithms that have favorable scaling with the size of the system being simulated, in order to make quantum computations feasible.

One class of algorithms that has been developed for NISQ hardware are hybrid quantum-classical algorithms, which divide the computational workload between a classical and a quantum computer. The variational quantum eigensolver (VQE) is an example of such an algorithm, which uses a parametrized quantum circuit to represent a wave function ansatz and estimates the energy of this ansatz through term-by-term measurement. Other examples of hybrid classical-quantum algorithms include the variational quantum time evolution for quantum dynamics simulations and the variational imaginary-time evolution for ground state energy optimization. On fault-tolerant quantum computers, the quantum phase estimation

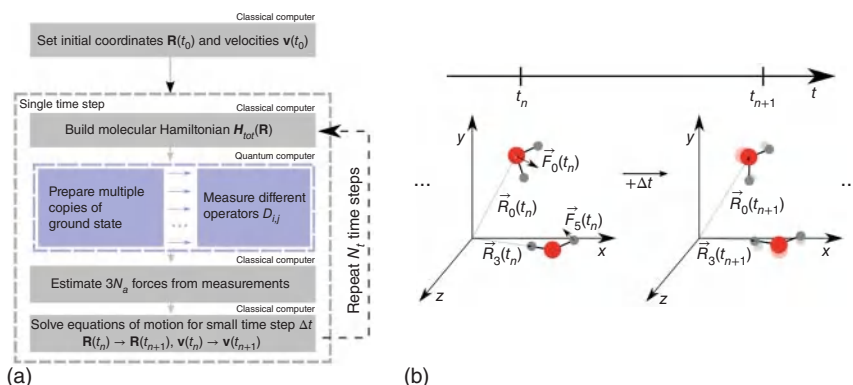


Figure 26.10 Schematic representation of molecular dynamics enhanced by a NISQ device. (a) Flowchart highlighting the hybrid setup in which a NISQ device is used to calculate the forces, while a classical computer updates the nuclei coordinates $\mathbf{R}$, velocities $\mathbf{v}$, and the molecular Hamiltonian. A typical MD simulation requires $N_t = O(10^6 - 10^9)$ time steps with a step size of $\Delta t = O(10^{-15})$ seconds. (b) Example of a single update of the nuclei coordinates $\mathbf{R}$ of two water molecules with red balls and gray balls denoting oxygen and hydrogen atoms, respectively. $\vec{F}_i$ denotes the three-dimensional force vector on the i^{th} atom. Source: Reproduced from [83]/American Physical Society/Public Domain CC BY 4.0.

(QPE) algorithm can be used to measure the full-CI (configuration Interaction) energy of a system by projecting a prepared state into an energy eigenstate of the Hamiltonian.

Generally, there are several classes where quantum effects play a role and quantum computing is used (see also [33] for deeper investigation)

- Local quantum effects in biochemical processes
- Global quantum effects in functional bio-molecules
- Quantum computing for structural biology
- Quantum computing for data-driven approaches to molecular biology

26.3.11.1 Local Quantum Effects in Biochemical Processes

In recent years, there has been growing interest in using quantum computation to study chemical reactions in bio-molecular systems. This is because chemical reactions in bio-molecules often involve a small number of atoms that are involved in bond-breaking and bond-formation processes, while the rest of the bio-molecule can be treated as a spectator environment. This means that the complex environment can be efficiently modeled, leaving the local quantum description of the reaction event as the main challenge.

To accurately describe chemical reactions using quantum mechanics, it is necessary to calculate the electronic energy of a given arrangement of atomic nuclei in bio-molecules. However, this calculation suffers from the *curse of dimensionality* due to the large number of orbitals required for accurate results and the exponential increase in computational complexity with the number of particles. As we see throughout this book, a variety of practical solution methods have been developed

for traditional computing hardware, but each method has its own limitations and trade-offs. Still, these traditional approaches have reached a remarkable degree of sophistication, accuracy, usefulness, and acceptance [84–86].

As mentioned above, the idea is to use quantum computers to significantly reduce the complexity of these calculations and enable more accurate predictions of chemical reactions in bio-molecules.

26.3.11.2 Global Quantum Effects in Functional Bio-Molecules

We will only briefly mention other bio-molecular processes influenced by quantum effects acting on a larger spatial scale. Two prominent examples are *light harvesting*, the process exploited by plants to convert light into chemical energy and *magnetoreception*, the process that enables birds to sense weak magnetic fields. It should be noted that we have controversy about the extent of quantum effects, whether they play a key role or not. In contrast to the chemical reactions described in the previous paragraph, quantum effects span, in this case, the whole molecule. For both phenomena, quantum mechanical effects affect the interaction between molecules – the chromophores, for light-harvesting, and the unpaired electron, for magnetoreception. Simulating quantum effects happening on such a large spatial scale is much harder than for chemical reactions.

The electronic wave functions of these systems depend on thousands of electrons, and this makes calculations extremely challenging for both classical and quantum computations. As a way out, effective Hamiltonians (usually referred to as *excitonic Hamiltonians*) have been designed that largely reduce the problem complexity but still encode the key quantum mechanical effects that drive these biological phenomena.

Quantum computing might have the potential to overcome these limits, provided that either a sufficiently large number of logical qubits can be provided in huge fault-tolerant quantum computers or a large analog quantum simulator can be built that can represent Hamiltonians with so many quasiparticles.

26.3.11.3 Quantum Computing for Structural Biology

Molecular biology relies heavily on reversible, non-covalent interactions between bio-molecules, which can be described using quantum approaches. However, full atomistic quantum mechanical simulations of large bio-molecules may still be beyond the reach of any simulation method, due to the curse of dimensionality. To overcome this challenge, classical surrogate potentials, known as force fields, can be used to approximate the potential energy surface (PES) of a molecule. These force fields describe the interactions between atoms in a molecule using a classical parametric interaction potential and can be used in classical atomistic simulations based on the Newton equation of motion of the atoms.

Force fields have been used extensively in molecular biology to simulate the dynamics of bio-molecules, including in the protein folding problem, where the goal is to predict the three-dimensional structure of a protein based on its amino acid sequence. Protein folding can be solved by simulating the dynamics of the protein under thermal equilibrium conditions, and the three-dimensional structure

of a protein is largely determined by non-covalent interactions between its amino acid residues.

Quantum computing has the potential to enhance classical simulations in molecular biology in several ways. One possibility is to use quantum computation to parametrize or refine force fields based on exact quantum chemical simulations of small polypeptides, which are too large to be accurately calculated using classical methods. Another approach is to use quantum computation to solve optimization problems that are classically complex, such as the protein folding problem or molecular docking, in which the goal is to predict the most stable binding configuration of a ligand to a target. This could be done using quantum algorithms for optimization or quantum-assisted classical simulations.

We will touch this in Section 26.5.

26.3.11.4 Quantum Computing for Data-Driven Approaches to Molecular Biology

A topic that is widely discussed is the potential added value that quantum computing brings to the table in order to boost machine-learning algorithms in molecular biology [87, 88]. We will touch this question in Section 26.4 and also like to point the reader to [89, 90].

26.4 Quantum Machine Learning

26.4.1 Introduction

The field of machine learning is a subdivision of artificial intelligence that has found numerous applications in various scientific and engineering domains, including pattern recognition, natural language processing, computer vision, biomedical and life sciences data analysis, and others. Employing machine learning (ML) techniques provides effective tools for bolstering the processes of drug discovery and development. Notably, these techniques can help with target discovery, target validation, detection of digital biomarkers, and analysis of data generated within both nonclinical and clinical phases. However, the continually expanding scale and inherent complexity of biological data present serious challenges for effectively developing informative and predictive models of underlying biological processes using machine learning.

In spite of the vast computational capacity that high-performance computing in the cloud offers, machine learning algorithms persistently face difficulties due to insufficient computing power. The emergence of the quantum computer represents a significant step forward in addressing issues tackled by classical computers, including exponential computing power, computing speed, and solving NP-type problems. As discussed previously, quantum computing relies on the fundamentals of quantum mechanics, like superposition, entanglement, and interference. These concepts enable massive parallelism, vast correlation, and the ability to find the solution to the problem Hamiltonian. Due to the tremendous computational ability offered by quantum computing, researchers have explored the possibility of combining quantum computing with classical machine learning. There have been some

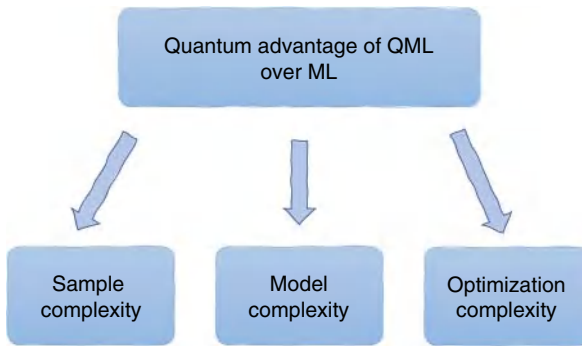


Figure 26.11 Advantage of quantum machine learning over classical machine learning models.

survey papers that mainly overview general ideas of different machine learning algorithms in the quantum version putting a spotlight on quantum technology and introducing a challenge to determine if QML will provide an advantage over classical machine learning or not (see Figure 26.11) [91].

26.4.2 The Shortcoming of Classical ML Models

The goal of machine learning (ML) is to make accurate predictions of unseen data. This is known as a generalization and is often referred to as a **model capacity** or **model complexity** to express different relationships between variables. Significant effort has been made to understand the generalization capabilities of classical ML models. The second aspect is **Sample Complexity** also called PAC Learning [92] from examples, i.e. Sample complexity refers to the number of samples needed to generalize from data to make an accurate prediction or inference. QML methods promise to resolve both model complexity and sample complexity. Abbas et al. [93] demonstrate that well-designed quantum neural networks offer an advantage over classical neural networks through a higher effective dimension and faster training ability thereby achieving more capacity of the ML model, which authors verify on real quantum hardware. Caro et al. [94] demonstrate a comprehensive study in QML using parameterized quantum circuits after training on a limited number of data samples and still achieving a good generalization. The third aspect is **computational/optimization complexity**, i.e. many tasks in machine learning such as maximum likelihood estimation using hidden variables, principal component analysis, training of neural networks, etc., require optimization of a non-convex objective function. It is a computationally complex problem to optimize non-convex functions (NP-hard problem). Classical optimization methods, such as gradient descent, often get stuck at local minima or saddle points without discovering the global minimum. Adiabatic quantum computers use quantum annealing to solve non-convex optimization problems. By prospecting low-energy configurations of an appropriate energy function and exploiting tunnel effects to evade local minima, adiabatic quantum computers present opportunities to derive benefits that go beyond just speeding up machine learning execution [92–95].

26.4.3 Types of QML

Generally, one differentiates between the involvement of quantum and classical computing resources, and whether the data comes in the form of quantum information or classical information. This leads to the four different approaches combining the disciplines of quantum computing and machine learning. The first letter refers to whether the system under study is classical or quantum, while the second letter defines whether a classical or quantum information processing device is used (Figure 26.12) [96].

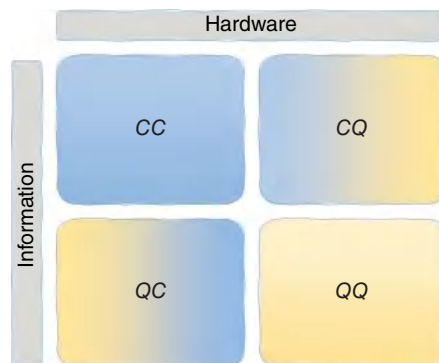
CC – Classical information on classical hardware: This describes the conventional case of classical machine learning, where input data is encoded classically, for example, in the form of vectors (e.g. texts, images, videos), and is used to train a model using exclusively classical hardware. Most research and development in the field of Machine Learning is focused on this case [96]. This category is also called quantum-inspired machine learning as classical data processing on a classical computer involves quantum mechanical concepts. Examples of a more generous interpretation of quantum-inspired algorithms are tensor networks and (digital) annealers [98, 99].

Tensor Networks are mathematical models originally developed for the study of many-body quantum systems in condensed matter physics, tensor networks allow for addressing large-scale problems and are at the heart of generalized regression and classification techniques, support tensor machines, higher-order canonical correlation analysis, higher-order partial least squares, and generalized eigenvalue decomposition. Furthermore, Tensor Networks enable the optimization of deep neural networks.

(Digital) annealing: Adiabatic quantum computers such as those produced by D-Wave or Fujitsu realize an energy minimization process called quantum annealing and they are specifically tailored toward solving quadratic unconstrained binary optimization (QUBO) problems. Applications of QUBOs in machine learning are many, e.g. Hopfield networks, binary classification, etc.

QC – Quantum information on classical hardware: Here, the input data is either quantum or quantum-related information, which is fed to classical

Figure 26.12 Quantum machine learning topology table. Source: Adapted from E. Aïmeur, G. Brassard, S. Gambs, “Machine Learning in a Quantum World” [97].



Machine Learning models that run on conventional hardware, see, e.g. [100] (cit. on p. 27).

QQ – Quantum information on quantum hardware: In this case input is either quantum or quantum-related datasets that use quantum hardware for learning, e.g. placing a QML procedure that directly receives inputs from physical experiments in a superposition state.

CQ – Classical information on quantum hardware: In this case input is classical but learning is done on quantum hardware. Amongst all scenarios, this is the most interesting one as all conventional Machine Learning tasks can also be accomplished with the involvement of quantum hardware. The expectation is that quantum speedup will make it possible to considerably accelerate learning processes or even tackle problems that are still beyond the reach even of current supercomputers. Indeed many common learning tasks involve linear algebra routines on very large systems of equations or optimization or search problems for which it seems likely that quantum advantages can be realized [99].

26.4.3.1 CQ Algorithm Types

HHL algorithm is a quantum algorithm published in 2008 for solving linear systems and named after the developers Aram Harrow, Avinatan Hassidim and Seth Lloyd [101]. The algorithm estimates the result of a scalar measurement on the solution vector to a given linear system of equations. HHL has a runtime of $\mathcal{O}(\log n \cdot k^2)$. This constitutes an exponential speedup over the fastest classical algorithms, which generally run in $\mathcal{O}(n \cdot k)$. In June 2018, Zhao et al. developed an algorithm for performing Bayesian training of deep neural networks in quantum computers with an exponential speedup over classical training due to the use of the quantum algorithm for linear systems of equations, providing also the first general-purpose implementation of the algorithm [102].

26.4.3.2 Quantum Regression Model or Quantum Algorithm for the Method of Least Squares

The least squares regression approach is commonly used for fitting a regression model to a given data. One of the limitations of these methods is the availability of QRAM. QRAM is proposed to allow retrieval of stored quantum information, as well as the updating of stored information after quantum computation. However, Date and Potok recently showed how to perform linear regression on adiabatic quantum computers. In order to accomplish this, they cast the regression problem as a QUBO problem, which they then solve on a D-Wave 2000Q adiabatic quantum computer [103].

26.4.3.3 Quantum Clustering Model

For $k = 2$, k -means clustering becomes a bipartition problem that can be cast as a QUBO problem and thus be solved via adiabatic quantum computing, e.g. D-Wave system [104]. Recently, Tomesh et al. apply the QAOA algorithm and run numerical simulations to compare their approach against classical k -means clustering [105]. Their results indicate that there exist data sets where QAOA might outperform standard k -means clustering on a core set.

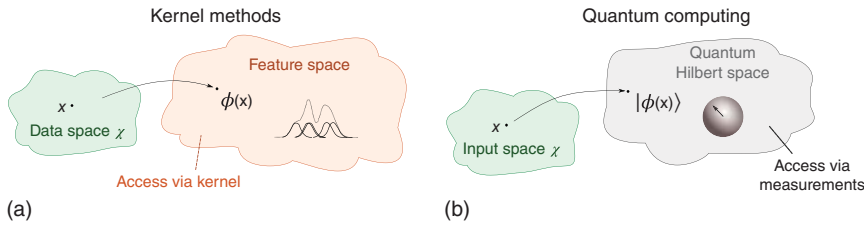


Figure 26.13 Calculating feature maps for SVM (a) on quantum computer. In quantum support vector machine (QSVM), data is mapped from low dimensional to high dimensional space by computing a feature map, which is computationally expensive. (b) Through quantum computer data is projected from input space to quantum Hilbert space to aid the calculation of kernel. Source: Picture is taken from Maria Schuld 2021, [108].

26.4.3.4 Quantum Kernel Model or Quantum Support Vector Machines

In machine learning, support vector machines (SVM) are supervised learning models with associated learning algorithms that analyze data for classification analysis. SVMs are one of the most robust prediction methods, being based on statistical learning frameworks. A key concept in classification methods is that of a kernel. Data cannot typically be separated by a hyperplane in its original space. A common technique used to find such a hyperplane consists of applying a nonlinear transformation function to the data. These nonlinear transformations are called feature maps, which are computationally very expensive to calculate on a classical computer and scale exponentially with the size of the problem. Recently there has been a proposal on how these feature maps can be computed in quantum Hilbert space on a quantum processor (see also Figure 26.13) [106, 108]. One of its implementation is given by IBM's Qiskit library [107].

26.4.3.5 Quantum Neural Networks

A neural network is a series of algorithms that endeavours to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates. It is composed of small interconnected computational units called neurons, which receive weighted input from other neurons. The weighted input received by a neuron is then typically summed up and subjected to a nonlinear activation function. Over the past decade, they have been in the limelight and are currently the most common tool in machine learning. Training large neural networks are computationally burdensome, and quantum neural networks have become a popular topic among quantum computing researchers. There have been many proposals on the quantum version of neural networks, but they all require their input/output to be read from and written to a QRAM, which is not possible to implement on NISQ-era quantum hardware. Mitarai et al. propose a hybrid quantum-classical algorithm for quantum circuit learning classical-quantum hybrid algorithm for machine learning on near-term quantum processors, which authors call quantum circuit learning, which learns a given task by tuning parameters implemented on it through iterative optimization of the parameters [109]. Theoretical investigation shows that a quantum circuit can approximate nonlinear functions, which is further confirmed by numerical simulations. Current quantum devices have serious limitations, including limited numbers of qubits and circuit depth. Variational Quantum Algorithms

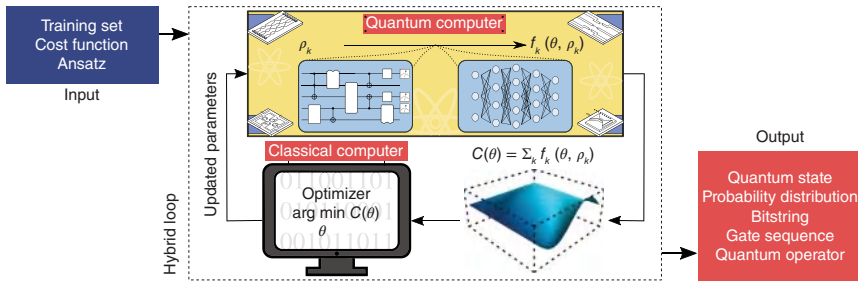


Figure 26.14 Schematic diagram of a Variational Quantum Algorithm (VQA). The inputs to a VQA are a cost function $C(\theta)$, with θ a set of parameters that encodes the solution to the problem, an ansatz whose parameters are trained to minimize the cost, and (possibly) a set of training data $\{\rho_k\}$ used during the optimization. Here, the cost can often be expressed as some set of functions $\{f_k\}$. Also, the ansatz is shown as a parameterized quantum circuit (on the left), which is analogous to a neural network (also shown schematically on the right). At each iteration of the loop, one uses a quantum computer to efficiently estimate the cost (or its gradients). This information is fed into a classical computer that leverages the power of optimizers to navigate the cost landscape $C(\theta)$ and solve the optimization problem. Once a termination condition is met, the VQA outputs an estimate of the solution to the problem. The form of the output depends on the precise task at hand. The red box indicates some of the most common types of outputs, pictures, and legends. Source: Adapted from [25].

(VQAs), which use a classical optimizer to train a parametrized quantum circuit, have emerged as a leading strategy to address these limitations, see Figure 26.14. VQAs have now been proposed for essentially all applications that researchers have envisioned for quantum computers, and they appear to be the best hope for obtaining a near-term quantum advantage.

26.4.3.6 Quantum Generative Models

In classical machine learning, auto-encoders are often used as generative models and have been extended toward more capable architectures such as Generative Adversarial Networks (GANs). Generative models can be expected to benefit considerably from potential quantum speedup through quantum sampling, which provides lower computational complexity than sampling in corresponding classical implementations.

Quantum Helmholtz machines are a class of neural networks that specifically allow for generative modeling. They consist of two sub-networks, a bottom-up recognition network that maps input data to a distribution over hidden variables and a top-down generative network that generates novel data from an instantiation of the values of the hidden variables. The training of a Helmholtz machine usually happens in an unsupervised manner. Its implementation on quantum hardware can be found here in [111, 112].

Born machines are purely quantum mechanical interpretations of Boltzmann machines, which are yet another type of neural network, in particular, a probabilistic extension of the Hopfield network. These Born machines are yet another kind of generative model and can be used to represent distributions of classical

Table 26.2 Overview of some quantum machine learning algorithms and their time complexities compared to their classical counterpart.

Algorithm	Classical	Quantum	QRAM
Linear regression	$\mathcal{O}(N)$	$\mathcal{O}(\log N)$	Yes
Gaussian process regression	$\mathcal{O}(N^3)$	$\mathcal{O}(\log N)$	Yes
Decision trees	$\mathcal{O}(N \log N)$	Unclear	No
Ensemble methods	$\mathcal{O}(N)$	$\mathcal{O}(\sqrt{N})$	No
Support vector machines	$\approx \mathcal{O}(N^2) - \mathcal{O}(N^3)$	$\mathcal{O}(\log N)$	Yes
Hidden Markov models	$\mathcal{O}(N)$	Unclear	No
Bayesian networks	$\mathcal{O}(N)$	$\mathcal{O}(\sqrt{N})$	No
Graphical models	$\mathcal{O}(N)$	Unclear	No
k -Means clustering	$\mathcal{O}(kN)$	$\mathcal{O}(\log kN)$	Yes
Principal component analysis	$\mathcal{O}(N)$	$\mathcal{O}(\log N)$	No
Persistent homology	$\mathcal{O}(\exp N)$	$\mathcal{O}(N^5)$	No
Gaussian mixture models	$\mathcal{O}(\log N)$	$\mathcal{O}(\text{polylog } N)$	Yes
Variational autoencoder	$\mathcal{O}(\exp N)$	Unclear	No
Multilayer perceptrons	$\mathcal{O}(N)$	Unclear	No
Convolutional neural networks	$\mathcal{O}(N)$	$\mathcal{O}(\log N)$	No
Bayesian deep learning	$\mathcal{O}(N)$	$\mathcal{O}(\sqrt{N})$	No
Generative adversarial networks	$\mathcal{O}(N)$	$\mathcal{O}(\text{polylog } N)$	No
Boltzmann machines	$\mathcal{O}(N)$	$\mathcal{O}(\sqrt{N})$	No
Reinforcement learning	$\mathcal{O}(N)$	$\mathcal{O}(\sqrt{N})$	No

Source: Reproduced from [116]/John Wiley & Sons/Public Domain CC BY 4.0.

data in terms of quantum states in superposition. More information about the Born Machines can be found in [113].

Utilizing quantum computers in Generative modeling tasks to potentially enhance conventional machine learning algorithms has emerged as a promising application but poses big challenges due to the limited number of qubits and the level of gate noise in available devices. However, there are first performance characterizations of quantum generative models [114] and some authors provided the practical and experimental implementation of a quantum-classical generative algorithm for various applications [115]. Table 26.2 gives an overview of some Quantum machine learning algorithms and their time complexities compared to their classical counterpart.

26.4.4 Limitations of QML

The limitations of quantum machine learning (QML) can be attributed to the limitations of NISQ-era quantum hardware, specifically, the number of qubits, coherence

time, and gate depth. It is important to note that QML has not yet convincingly demonstrated significant advantages compared to classical machine learning approaches. Thus far, only certain instances have exhibited incremental advantages through the use of quantum-inspired techniques, and a few cases involving hybrid quantum computing experiments show promise for consideration in the near future.

26.5 Designing Peptides and Proteins

26.5.1 Background

Modern drug discovery relies on identifying disease-modifying molecules that bind to a biological target (typically a protein), which is believed to play a role in disease biology. Once a hit molecule has been identified, it is then optimized to improve drug properties of interest such as binding affinity, selectivity, stability, oral bioavailability, etc. If a compound with the desired properties is found, it is then put forward through drug development, clinical development, and eventually commercialization if successful. This entire process is long and capital-intensive, taking 10–15 years at a cost of US\$ 2 billion, and unfortunately with a dismally low success rate.

In recent decades, however, computer-aided drug design methods have advanced as an alternative method to accelerate hit discovery and optimization of lead drug candidates, many of these methods are explained in throughout this book. More recently, quantum computing has emerged as a potential tool to complement classical approaches with many potential applications from target assessment to hit and lead identification and optimization [117]. In this section, we present peptide and protein design as a near-term application of quantum computing for drug discovery.

26.5.2 *De Novo* Rational Design

Of the many challenges that need to be overcome in the drug discovery process, hit identification – the process of identifying biologically active molecules – is a particularly daunting one where scientists are tasked with finding the right molecule amongst an almost infinite number of possibilities.

The process involves screening a large number of molecules against a target, and it has become common practice to rely on trial-and-error high throughput screening, where large libraries of millions of compounds are tested in the wet lab against an isolated target.

In contrast to trial-and-error screening, *de novo* rational design aims to “design” the right molecule computationally from first principles. This approach takes advantage of the structure of a drug target to create new ligand compounds that are complementary and are capable of binding to the target protein activating or inhibiting its function.

Specifically, scientists leverage three-dimensional structural data of the drug target, such as X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy, to characterize the binding pocket or binding surface and then

apply physics-based simulations to design new molecules with specific binding interactions against the target. One key advantage of this approach is that completely new chemical matter, not available in any database, can be generated.

26.5.3 Peptide and Protein Design

While these methods were traditionally focused on designing small molecules, in recent years scientists have developed methods for designing larger molecules including peptides and even large proteins. For example, the Rosetta software, which is one of the leading software packages for protein design and structure prediction [118], has been applied to design new protein topologies [119, 120], large macromolecular assemblies [121–124], proteins with the ability to bind to toxic small molecules [125, 126] or to other proteins of therapeutic interest [127, 128], and enzymes able to catalyze reactions that no known natural enzyme can catalyze [129, 130].

More recently, Rosetta has been generalized to the design of diverse synthetic heteropolymers that are built from noncanonical side-chains and backbone chemistries [131–135].

Peptides, in particular, are a class of molecules that sit in between the size of small molecules and biologics, making them ideal drug molecules for undruggable targets. The first peptide therapy was insulin, which has been a huge success for pharmaceutical companies with steady growth. Peptides have been shown to be key biologic mediators with increased selectivity, low toxicity, and high potency, but designing the right type of peptide has been challenging [136, 137].

26.5.4 Protein Design as an Optimization Problem

Computational design of peptide and protein ligands involves astronomically large search problems. The goal here is to identify a sequence that predictably folds into a specific binding conformation that is complementary to the drug target of interest – this problem is known as sequence design.

Computationally, the problem can be formulated as a combinatorial optimization problem. Given N designable sequence positions and D discrete side-chain identity and conformation possibilities (aka. rotamers) at each position, there are $D \cdot N$ possible solutions to the problem of finding the optimal selection of one rotamer per position.

Exhaustive search for these problems rapidly exceeds the capabilities of even the largest supercomputers, as typical design tasks can involve hundreds of positions, with hundreds or even thousands of possibilities for the rotamer at each position. Designing large proteins or peptides with noncanonical amino acids can therefore rapidly become intractable. For this reason, heuristic methods are commonly used. The Rosetta software suite, for example, relies on simulated annealing-based heuristics for protein design tasks.

Rosetta's primary design heuristic, called the Packer, attempts to solve the sequence design problem using simulated annealing-based searches of rotamer

space, which, although not guaranteed to converge to the global optimum, tend to find high-quality solutions near the optimum very rapidly [118, 136]. Unfortunately, as the designable positions (N) and number of rotamers increase (D), the rotamer space quickly grows too large for simulated annealing approaches to work effectively if at all. The Packer's simulated annealing approach is also very sensitive to the shape of the energy landscape, relying on broad energy wells for which a downhill path to the lowest-energy state exists, and sometimes failing to find solutions in narrow energy wells. Alternative approaches, such as dead-end elimination or branch-and-bound searches, have also been used [138–142], though these are typically too slow for most design tasks.

26.5.5 Quantum Optimization for Peptide and Protein Design

Quantum computing provides an attractive alternative. Whereas classical computers typically solve difficult combinatorial problems by iterating through many possibilities, either exhaustively or using stochastic or deterministic search heuristics, quantum computers can leverage the superposition of quantum states to represent all possible solutions to a posed problem as a superposition of quantum states [1, 143].

A quantum algorithm can then sample efficiently from states that have been programmed to correspond to the global optimum (or near optimum) upon measurement [144].

The major advantage of a quantum computing approach is the massive parallelism that can be achieved by modeling many solutions simultaneously. In fact, the number of solutions that can be modeled simultaneously doubles with each additional qubit added to the system, allowing exponential scaling far beyond anything achievable with classical computers for certain classes of search problems. It is thus not surprising that this problem has gathered significant interest as an early application of quantum computing for both quantum annealing [59] and gate-based [145] hardware.

26.5.6 Quantum Annealing Approach

Mulligan et al. 2020 [59] showed that the rotamer optimization problem—the central problem that must be solved when designing a protein—could be formulated as a QUBO problem with direct mapping to the D-Wave quantum annealer. Furthermore, they demonstrated that this mapping can be made without simplifying the design task or sacrificing the accuracy of the existing classical methods, ensuring that the quantum approach will be at least as useful as current classical approaches. The algorithm they developed, called qPacker, allows them to not only map the protein design problem but to also design scientifically valid peptide designs (Figure 26.15).

Using classical folding simulations, they also showed that the output from the quantum design algorithm had scientific validity comparable to the output from Rosetta's Packer. Because the QPacker is a direct mapping of the Rosetta Packer to the quantum architecture, it can be applied to any design task to which Rosetta

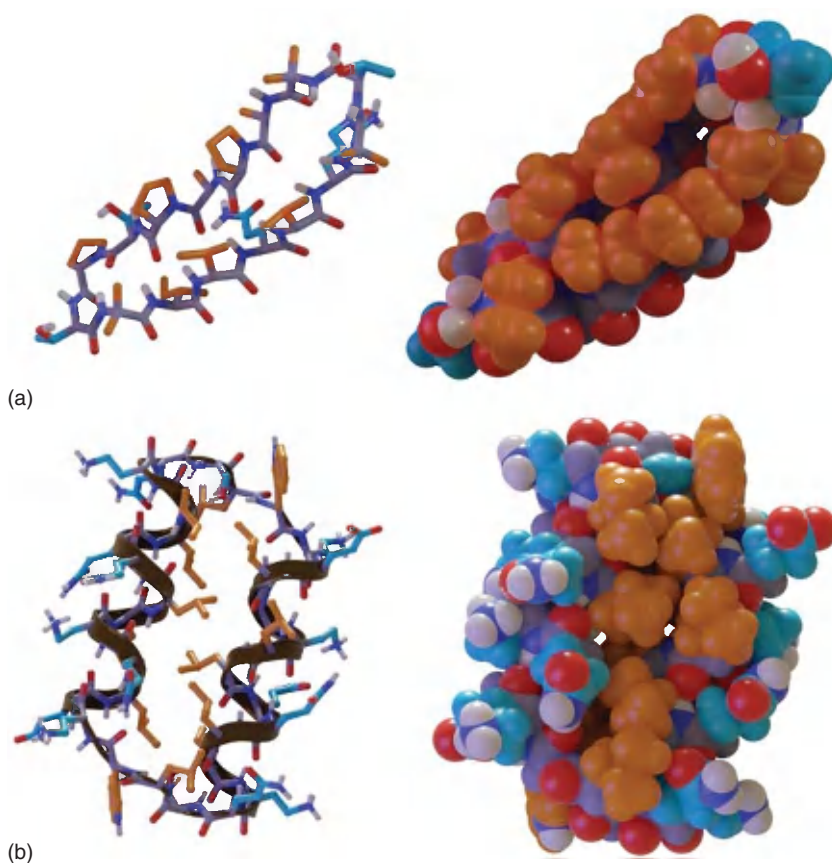


Figure 26.15 Representative designs produced by the QPacker. (a) Sticks (left) and space-filling (right) models of a representative 16-residue α -sheet design. Apolar side-chains are shown in orange, polar side-chains, in cyan, and backbone atoms in gray. Nitrogen, oxygen, and polar hydrogen atoms are shown in blue, red, and white, respectively. The QPacker consistently found solutions with good side-chain packing, particularly between apolar groups. (b) Ribbon and sticks (left) and space-filling (right) models of a representative 32-residue S₂-symmetric coiled-coil design. Colors are as in the previous panel. The excellent packing of the hydrophobic core is evident. Other features important for folding, such as salt bridges and side-chain hydrogen-bonding interactions, are also in evidence. Source: Figure and caption are taken from [59] Mulligan 2019 / with permission from BioRxiv.

can currently be applied, but with potentially better scaling than their classical counterparts. As larger quantum computers are introduced, the authors argued that the QPacker may allow larger design tasks than will ever be possible on classical hardware.

26.5.7 The qPacker Algorithm

The qPacker algorithm effectively replaces Rosetta's simulated annealing algorithm with a quantum annealing approach. It takes advantage of the fact that the problem

is already rotamer-level pair-wise decomposable to express the problem as a one-body two-body problem, which is equivalent to the Ising model and thus directly mappable to the D-wave quantum annealer hardware.

Given N designable positions with rotamers at the N th position indexed as $r_{N,1}$ through r_{N,D_N} , a particular solution is given by a vector $\vec{S} = (S_1, S_2, \dots, S_N)$, where S_1 through S_N are the indices of the chosen rotamers at each position.

The energy of the solution is given by,

$$E(\vec{S}) = \sum_{i=1}^N O(S_i) + \sum_{j=1}^{N-1} \sum_{k=j+1}^N T(S_j, S_k) \quad (26.13)$$

where $O(S_i)$ represents a one-body energy of rotamer S_i , and $T(S_j, S_k)$ represents a two-body interaction energy between rotamers S_j and S_k . The energy terms are pre-computed using the Rosetta energy function.

The problem is then to find the set of rotamers, one at each position, that minimizes the energy function.

Finding the global minimum of the optimization problem can be formulated as adiabatically evolving the ground state of an initial Hamiltonian, H_S (with a known and easy-to-prepare ground state), to the ground state of the target Hamiltonian, H_T (for which the quantum-mechanical ground state corresponds to the minimum of the classical objective function being optimized). Samples can be drawn from H_T to find the minimum of the classical objective function.

Formulated as a transverse Ising model, the transition between states is controlled by slowly changing the transverse field (29). The overall Hamiltonian, parameterized with a value τ that varies from 0 at the start of the annealing phase to 1 at the end, is given by:

$$H(\tau) = (1 - \tau)H_S + \tau H_T, \quad \text{for } 0 \leq \tau \leq 1 \quad (26.14)$$

When $\tau = 0$, the lowest-energy state (typically an equal superposition of all states) is easily prepared and gives all classical configurations equal probability. When $\tau = 1$, the system corresponds to the ground-state Ising problem that one wishes to solve. If the system transitions from $\tau = 0$ to $\tau = 1$ sufficiently slowly, then the solution is guaranteed with high probability; however, the temperature of the system, the influence of external sources of thermal or electrical noise, and the nature of H_T all influence the trade-off between speed and accuracy. Quantum annealing is particularly useful for finding solutions to posed optimization problems in which the search space is discrete, with many local minima, as is the case for rotamer optimization tasks.

The D-Wave quantum annealer allows posed problems to be expressed as QUBO tasks, which involve an objective function f with a functional form very similar to that of the Rosetta energy function. Given n qubits that, on measurement, can take values of 1 or 0, yielding a state $s(q_1, q_2, \dots, q_n)$, the functional form of the objective function for a measured state is:

$$f(\vec{S}) = \sum_{i=1}^n o_i q_i + \sum_{j=1}^{n-1} \sum_{k=j+1}^n t_{j,k} q_j q_k \quad (26.15)$$

where o_i represents the one-qubit penalty for q_i being 1, and $t_{j,k}$ represents the two-qubit penalty if q_j and q_k are both 1. The D-Wave system is programmed by setting the values of o_i and $t_{j,k}$ coefficients. The quantum annealing algorithm then seeks the state $\vec{s}_{min}$, which minimizes the function $f(\vec{s})$.

Given this, the QPacker algorithm can be developed simply by assigning each rotamer under consideration to a different qubit, and then applying three simple rules. First, each o_i must be set to be the classically pre-computed one-body Rosetta energy for that rotamer. Second, each $t_{j,k}$ for qubits j and k representing rotamers at different sequence positions must be set to be the classically pre-computed two-body Rosetta energy for that pair of rotamers.

And third, each $t_{j,k}$ for qubits j and k representing different rotamers at the same position must be assigned a large positive value (effectively prohibiting solutions in which more than one rotamer is selected at a given position). When the D-Wave quantum processing unit is programmed in this way, a Rosetta design task is translated without distortion or simplification into a quantum annealing problem.

26.5.8 Gate-Based Approaches

QAOA [146] offers a gate-based alternative to quantum annealing for solving combinatorial optimization problems. As such, it is conceivable that the qPacker algorithm proposed by [59] could be reformulated as a QAOA-type circuit. This is an area of future development, with early work starting to show potential implementation on trapped ion devices [145].

26.6 Conclusion

A big part of computational drug discovery is “chemical simulation” in the wider sense. In order to identify where quantum computers could play a role, an accurate assessment of how quantum computers can be used for chemical simulation, especially their potential computational advantages, needs to be done.

It is important to note that the existing quantum simulation algorithms are not straightforward or universal, and their effectiveness depends on the specific problem being solved. As a result, it will require a combination of expertise from various fields, including mathematics, physics, computer science, chemistry, and biology, to continue developing and improving these algorithms. This will require insight and creativity to design algorithms that are tailored to the specific problem at hand and that can take advantage of the unique capabilities of quantum computers. It will also require a deep understanding of the underlying physical and chemical processes being simulated, as well as the specific applications and goals of the simulation. As quantum simulation algorithms continue to evolve, they have the potential to revolutionize a wide range of fields by providing more accurate and efficient ways to simulate and understand complex quantum systems.

While currently the technical maturity of quantum computing is in an early phase and is not yet ready for industrial use, we still see an emerging ecosystem and

big funding from government, industry, private sector, and academia. However, it should be noted that the journey is on for a long-term, multi-year exploration, not a quick-win-sprint.

Several components are needed to bring forward a fully useful quantum computing environment (FUQC).

There are opportunities in the domains such as quantum chemistry, drug design and discovery, biomolecular processes, biological optimization, and genetics and genomics. Further developments in quantum hardware are needed, as well as the development of specialized quantum algorithms and investment in key enabling materials. On the cognitive side, there are also challenges related to the availability and types of skills and quantum literacy.

Finally, because this is a rapidly evolving field with a great multi-stakeholder community and constructive team spirit, the authors can only recommend to get networked and also check out events of conference organizers, e.g. IQT, QT Quantum.Tech, QuantumBusiness, Q2B just to name a few.

The time to be in is now.

26.7 Further Reading

If you are new to quantum mechanics or want to refresh – a really entertaining course is that from Leonard Susskind and his Theoretical Minimum Series [147].

There are really good introductions into quantum computing and quantum information science, e.g. from Nielsen and Chuang [47] or from Sutor [148]. Readers coming from the programming side might give Johnston, Harrigan, and Gimeno-Segovia [149] a look, it's a book with essential algorithms and code samples. A different perspective on quantum processes as well as general considerations about a novel graphical notation is provided by Coecke and Kissinger [150], or with Gogioso for a younger audience in an inclusive style [151]. Whurley addresses the audience in a “for dummies” style [152].

Meanwhile, there are excellent free introductory books on quantum technologies, e.g. by Oliver Ezratty [153] and courses available on the web via Youtube, e.g. the courses from the QuTech Academy at TUDelft [154], as well as commercial education courses, e.g. MITx [155], and more available on all educational platforms, Udemy, edX, LinkedIn, Coursera, just to name a few.

An educational resource that is often neglected is gaming – there are several quantum games, one interesting resource is the *Quantum Odyssey* from QuarksInteractive [156]. An overview of further resources can be found here [157].

Finally, regarding the application of quantum computing in the Pharmaceutical Industry, we like to point the reader to excellent papers, for e.g. simulation of large-scale protein–ligand interactions [158, 159], mRNA codon optimization [160], protein folding [161], network medicine [162], and very recent on drug design on quantum computer [163] or the work like “Quantum computing in pharma: A multilayer embedding approach for near future applications” from Riverlane [164] just to name a few.

References

- 1 Feynman, R.P. (1982). Simulating physics with computers. *International Journal of Theoretical Physics* 21: 467–488.
- 2 Editorial (2022). 40 years of quantum computing. *Nature Reviews Physics* 4 (1): 1. <https://doi.org/10.1038/s42254-021-00410-6>.
- 3 Gunashekar, S., Flanagan, I., d'Angelo, C. et al. (2022). *Using Quantum Computers and Simulators in the Life Sciences: Current Trends and Future Prospects*. RAND Corporation.
- 4 Arye and Aiello, C.D. (2022). The future of biology is quantum. <https://arye.substack.com/p/the-future-of-biology-is-quantum?s=r> (accessed 29 August 2023).
- 5 Grötschel, M., Knobloch, E., Schiffrers, J. et al. (ed.) (2016). *Vision als Aufgabe: Das Leibniz-Universum im 21. Jahrhundert*. Berlin-Brandenburgische Akademie der Wissenschaften. ISBN: 978-3-939818-67-0. <https://www.zib.de/groetschel/pubnew/Vision.pdf>.
- 6 Jean-pascal, A. (2015). “Cette caractéristique secrete et sacree” : Leibniz et Bouvet lecteurs du Yijing. <https://www.academia.edu/13978537/Cettecaract%C3%A9ristiquesecr%C3%A8teetsacr%C3%A9eLeibnizetBouvetlecteursduYijing> (accessed 29 August 2023).
- 7 Scholz, E. (2022). *G. W. Leibniz als Mathematiker*. Vorweg <http://www2.math.uni-wuppertal.de/~scholz/preprints/Leibniz.pdf> (accessed 9 December 2023).
- 8 Wolfram, S. (2013). Dropping in on Gottfried Leibniz— Stephen Wolfram writings. <https://writings.stephenwolfram.com/2013/05/dropping-in-on-gottfried-leibniz/> (accessed 4 September 2023).
- 9 Kuhn, T.S. (2012). *The Structure of Scientific Revolutions*, 50e. Chicago, IL: University of Chicago Press. ISBN: 9780226458120
- 10 Rifkin, J. (2013). *The Third Industrial Revolution: How Lateral Power is Transforming Energy, the Economy, and the World*. Basingstoke, England: Palgrave Macmillan. ISBN: 9780230341975
- 11 Moreau, P.-A., Toninelli, E., Gregory, T. et al. (2019). Imaging bell-type non-local behavior. *Science Advances* 5 (7): eaaw2563. <https://doi.org/10.1126/sciadv.aaw2563>.
- 12 Brown, T.B., Mann, B., Ryder, N. et al. (2020). Language models are few-shot learners. <http://arxiv.org/abs/2005.14165>.
- 13 Coecke, B., de Felice, G., Meichanetzidis, K., and Toumi, A. (2020). Foundations for near-term quantum natural language processing. <http://arxiv.org/abs/2012.03755>.
- 14 Bobier, J.-F., Langione, M., Tao, E., and Gourévitch, A. (2021). What happens when ‘if’ turns to ‘when’ in quantum computing? <https://www.bcg.com/de-de/publications/2021/building-quantum-advantage> (accessed 29 August 2023).
- 15 Accenture (2021). <https://www.accenture.com/us-en/insights/technology/quantum-impact> (accessed 29 August 2023).

- 16 QUTAC Consortium (2021). QUTAC - quantum technology and application consortium. <https://www.qutac.de/?lang=en> (accessed 4 September 2023).
- 17 Bayerstadler, A., Becquin, G., Binder, J. et al. (2021). Quantum technology, and application consortium QUTAC. Industry quantum computing applications. *EPJ Quantum Technology* 8 (1): 25. <https://doi.org/10.1140/epjqt/s40507-021-00114-x>.
- 18 Silvestri, R. (2020). Business value of quantum computers: analyzing its business potentials and identifying needed capabilities for the healthcare industry. <https://www.researchgate.net/publication/343683519BusinessValueofQuantumComputersanalyzingitsbusinesspotentialsandidentifyingneededcapabilitiesforthehealthcareindustry> (accessed 4 September 2023).
- 19 Nita, L., Smith, L.M., Chancellor, N., and Cramman, H. (2021). The challenge and opportunities of quantum literacy for future education and transdisciplinary problem-solving. *Research in Science & Technological Education* 41 (2): 564–580. <https://doi.org/10.1080/02635143.2021.1920905>.
- 20 NN. 97, 1892. <https://digi.ub.uni-heidelberg.de/diglit/fb97> (accessed 29 August 2023).
- 21 Schrödinger, E. (1935). Die gegenwärtige situation in der quantenmechanik. *Naturwissenschaften* 23: 807–812.
- 22 Born, M. (1926). Zur Quantenmechanik der Stossvorgänge. *The European Physical Journal A* 37 (12): 863–867. <https://doi.org/10.1007/bf01397477>.
- 23 Joensson, C. (1961). Elektroneninterferenzen an mehreren künstlich hergestellten Feinspalten. *The European Physical Journal A* 161 (4): 454–474. <https://doi.org/10.1007/bf01342460>.
- 24 Wikipedia (2022). *Born Rule*. <https://en.wikipedia.org/wiki/Bornrule> (accessed 4 September 2023).
- 25 Gerlach, W. and Stern, O. (1922). Der experimentelle nachweis der richtungsquantelung im magnetfeld. *Zeitschrift für Physik* 9: 349–352. <https://doi.org/10.1007/BF01326983>.
- 26 Feynman, R.P. (1963). *The Feynman Lectures on Physics*, Chapter 5, vol. 3. Spin One. https://www.feynmanlectures.caltech.edu/III_05.html (accessed 9 December 2023).
- 27 Heisenberg, W. (1927). Ueber den anschaulichen inhalt der quantentheoretischen kinematik und mechanik. *The European Physical Journal A* 43 (3–4): 172–198. <https://doi.org/10.1007/bf01397280>.
- 28 Wikipedia Interpretation of QM (2022). *Interpretations of Quantum Mechanics*. <https://en.wikipedia.org/wiki/Interpretationsofquantummechanics> (accessed 4 September 2023).
- 29 Graham, N. and Dewitt, B.S. (ed.) (1973). *The Many Worlds Interpretation of Quantum Mechanics*. Princeton, NJ: Princeton University Press. ISBN: 9780691081311.
- 30 von Neumann, J. (1996). *Mathematische Grundlagen der Quantenmechanik*. Berlin, Heidelberg: Springer-Verlag. ISBN: 978-3-540-59207-5. <https://doi.org/10.1007/978-3-642-61409-5>.

- 31 Dirac, P.A.M. (1939). A new notation for quantum mechanics. *Mathematical Proceedings of the Cambridge Philosophical Society* 35 (3): 416–418. <https://doi.org/10.1017/S0305004100021162>.
- 32 Altman, E., Brown, K.R., Carleo, G. et al. (2019). Quantum simulators: architectures and opportunities. *arXiv [quant-ph]*. <http://arxiv.org/abs/1912.06938>.
- 33 Baiardi, A., Christandl, M., and Reiher, M. (2022). Quantum computing for molecular biology. *arXiv [quant-ph]*. <http://arxiv.org/abs/2212.12220>.
- 34 Davletkaliyev, R. (2022). https://www.linkedin.com/posts/iqm-quantumcomputers_iqm-quantumcomputing-algorithm-activity-6963814546411008000-IU7A/ (accessed 9 December 2023).
- 35 Wootton, J. (2017). Making a quantum computer smile. <https://medium.com/qiskit/making-a-quantum-computer-smilecee86a6fc1de> (accessed 9 December 2023).
- 36 Quibik Wjh31 (2010). <https://commons.wikimedia.org/w/index.php?curid=10073387> (accessed 4 September 2023).
- 37 Coecke, B. and Duncan, R. (2011). Interacting quantum observables: categorical algebra and diagrammatics. *New Journal of Physics* 13: 043016. <https://doi.org/10.1088/1367-2630/13/4/043016>.
- 38 Bravyi, S., Dial, O., Gambetta, J.M. et al. (2022). The future of quantum computing with superconducting qubits. *Journal of Applied Physics* 132 (16): 160902. <https://doi.org/10.1063/5.0082975>.
- 39 Pistoia Alliance and Associated Organizations (2021). Quantum computing community of interest, Pistoia Alliance, QuPharm, QED-C, QPARC. <https://www.pistoiaalliance.org/quantum-computing/quantum-computing/> (accessed 30 August 2023).
- 40 Lucas, A. (2014). Ising formulations of many NP problems. *Frontiers in Physics* 2: 5. <https://doi.org/10.3389/fphy.2014.00005>.
- 41 Shor, P.W. (1997). Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM Journal on Computing* 26 (5): 1484–1509. <https://doi.org/10.1137/S0097539795293172>.
- 42 Classiq Technologies (2022). <https://www.classiq.io/insights/quantum-algorithms-shors-algorithm> (accessed 4 September 2023).
- 43 Grover, L.K. (1996). A fast quantum mechanical algorithm for database search. *arXiv*.
- 44 Lee, S., Lee, J., Zhai, H. et al. (2022). Is there evidence for exponential quantum advantage in quantum chemistry? <http://arxiv.org/abs/2208.02199>.
- 45 Wootters, W.K. and Zurek, W.H. (1982). A single quantum cannot be cloned. *Nature* 299 (5886): 802–803. <https://doi.org/10.1038/299802a0>.
- 46 Quirk (2022). Quirk, A drag-and-drop quantum circuit simulator. <https://algassert.com/quirk> (accessed 4 September 2023).
- 47 Nielsen, M.A. and Chuang, I.L. (2010). *Quantum Computation and Quantum Information: 10th Anniversary Edition*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511976667>.
- 48 IBM (2022). <https://docs.quantum-computing.ibm.com> (accessed 9 December 2023).

- 49 Google Quantum AI (2022). <https://quantumai.google/cirq> (accessed 29 August 2023).
- 50 Bradben (2019). What are the Q# programming language and QDK - Azure Quantum. <https://docs.microsoft.com/en-us/azure/quantum/overview-what-is-qsharp-and-qdk> (accessed 29 August 2023).
- 51 Amazon (2022). <https://aws.amazon.com/braket/?nc1=hls> (accessed 29 August 2023).
- 52 Rigetti (2022). <https://qcs.rigetti.com/sdk-downloads> (accessed 4 September 2023).
- 53 Xanadu Inc. (2022). <https://www.xanadu.ai/products/pennylane/> (accessed 29 August 2023).
- 54 Quantinuum (2022). <https://www.quantinuum.com/developers/tket> (accessed 4 September 2023).
- 55 QOSF (2022). <https://github.com/qosf/awesomequantum-software> (accessed 9 December 2023).
- 56 Quera (2023). <https://www.quera.com/bloqade> (accessed 9 December 2023).
- 57 D-Wave (2020). What is quantum annealing, D-Wave. https://docs.dwavesys.com/docs/latest/c_gs_2.html (accessed 9 December 2023).
- 58 D-Wave (2022). <https://www.dwavesys.com/solutions-and-products/ocean/> (accessed 29 August 2023).
- 59 Mulligan, V.K., Melo, H., Merritt, H.I. et al. (2020). Designing peptides on a quantum computer. *BioRxiv*, 752485
- 60 Keever, C.M. and Lubasch M. (2023). Towards adiabatic quantum computing using compressed quantum circuits. <https://doi.org/10.48550/arXiv.2311.05544>.
- 61 Fujitsu Quantum (2020). <https://www.fujitsu.com/global/imagesgig5/Quantum-Inspired-Optimization-Services-Pharmaceutical.pdf> (accessed 4 September 2023).
- 62 Snelling, D., Shahane, G., Shipman, W. et al. (2020). A quantum-inspired approach to de-novo drug design. <https://www.fujitsu.com/fi/imagesgig5/Healthcare-Assets-Whitepaper.pdf> (accessed 4 September 2023).
- 63 Gambetta, J.M., Chow, J.M., and Steffen, M. (2017). Building logical qubits in a superconducting quantum computing system. *npj Quantum Information* 3 (1): 1–7. <https://doi.org/10.1038/s41534-016-0004-0>.
- 64 Yang, C.H., Leon, R.C.C., Hwang, J.C.C. et al. (2020). Operation of a silicon quantum processor unit cell above one kelvin. *Nature* 580 (7803): 350–354. <https://doi.org/10.1038/s41586-020-2171-6>.
- 65 Gulka, M., Wirtitsch, D., Ivády, V. et al. (2021). Room-temperature control and electrical readout of individual nitrogen-vacancy nuclear spins. *Nature Communications* 12 (1): 4421. <https://doi.org/10.1038/s41467-021-24494-x>.
- 66 Wintersperger, K., Dommert, F., Ehmer, T. et al. (2023). Neutral atom quantum computing hardware: performance and end-user perspective. *EPJ Quantum Technology*, 10 (1). <https://doi.org/10.1140/epjqt/s40507-023-00190-1>.
- 67 Preskill, J. (2018). Quantum Computing in the NISQ era and beyond. *Quantum* 2: 79. <https://doi.org/10.22331/q-2018-08-06-79>.

- 68 Djordjevic, I. (2012). *Quantum Information Processing and Quantum Error Correction: An Engineering Approach*. San Diego, CA: Academic Press. ISBN: 9780123854919.
- 69 Wikipedia Contributors (2022). Quantum error correction. <https://en.wikipedia.org/w/index.php?title=Quantumerrorcorrection&oldid=1101107832> (accessed 29 August 2023).
- 70 McCormick, K. (2021). How quantum computers will correct their errors. <https://www.quantamagazine.org/how-quantum-computers-will-correct-their-errors-20211116/> (accessed 29 August 2023).
- 71 Hyppä, E., Kundu, S., Chan, C.F. et al. (2022). Unimon qubit. *Nature Communications* 13 (1): 6895. <https://doi.org/10.1038/s41467-022-34614-w>.
- 72 Photonic (2023). <https://photonic.com/technology/> (accessed 9 December 2023).
- 73 Blunt, N.S., Camps, J., Crawford, O. et al. (2022). Perspective on the current state-of-the-art of quantum computing for drug discovery applications. *Journal of Chemical Theory and Computation*. <https://doi.org/10.1021/acs.jctc.2c00574>.
- 74 Babbush, R., Wiebe, N., McClean, J. et al. (2018). Low-depth quantum simulation of materials. *Physical Review X* 8 (1): 011044. <https://doi.org/10.1103/physrevx.8.011044>.
- 75 Huggins, W.J., McClean, J.R., Rubin, N.C. et al. (2021). Efficient and noise resilient measurements for quantum chemistry on near-term quantum computers. *npj Quantum Information* 7 (1): 1–9. <https://doi.org/10.1038/s41534-020-00341-7>.
- 76 Lee, J., Berry, D.W., Gidney, C. et al. (2021). Even more efficient quantum computations of chemistry through tensor hypercontraction. *PRX Quantum* 2: 030305. <https://doi.org/10.1103/PRXQuantum.2.030305>.
- 77 McClean, J.R., Romero, J., Babbush, R., and Aspuru-Guzik, A. (2016). The theory of variational hybrid quantum-classical algorithms. *New Journal of Physics* 18 (2): 023023. <https://doi.org/10.1088/1367-2630/18/2/023023>.
- 78 Reiher, M., Wiebe, N., Svore, K.M. et al. (2017). Elucidating reaction mechanisms on quantum computers. *Proceedings of the National Academy of Sciences of the United States of America* 114 (29): 7555–7560. <https://doi.org/10.1073/pnas.1619152114>.
- 79 von Burg, V., Low, G.H., Häner, T. et al. (2021). Quantum computing enhanced computational catalysis. *Physical Review Research* 3: 033055. <https://doi.org/10.1103/PhysRevResearch.3.033055>.
- 80 Bharti, K., Cervera-Lierta, A., Kyaw, T.H. et al. (2022). Noisy intermediate-scale quantum algorithms. *Reviews of Modern Physics* 94: 015004. <https://doi.org/10.1103/RevModPhys.94.015004>.
- 81 Hempel, C., Maier, C., Romero, J. et al. (2018). Quantum chemistry calculations on a trapped-ion quantum simulator. *Physical Review X* 8: 031022. <https://doi.org/10.1103/PhysRevX.8.031022>.
- 82 O'Malley, P.J.J., Babbush, R., Kivlichan, I.D. et al. (2016). Scalable quantum simulation of molecular energies. *Physical Review X* 6: 031007. <https://doi.org/10.1103/PhysRevX.6.031007>.

- 83 O'Brien, T.E., Streif, M., Rubin, N.C. et al. (2022). Efficient quantum computation of molecular forces and other energy gradients. *Physical Review Research* 4: 043210. <https://doi.org/10.1103/PhysRevResearch.4.043210>.
- 84 Dykstra, C., Frenking, G., Kim, K., and Scuseria, G. (ed.) (2011). *Theory and Applications of Computational Chemistry: The First Forty Years*. Elsevier Science and Technology. ISBN: 9780080456249
- 85 Kohn, W. (1999). Nobel lecture: electronic structure of matter—wave functions and density functionals. *Reviews of Modern Physics* 71 (5): 1253–1266. <https://doi.org/10.1103/revmodphys.71.1253>.
- 86 Pople, J.A. (1999). Nobel lecture: quantum chemical models. *Reviews of Modern Physics* 71 (5): 1267–1274. <https://doi.org/10.1103/revmodphys.71.1267>.
- 87 Ching, T., Himmelstein, D.S., Beaulieu-Jones, B.K. et al. (2018). Opportunities and obstacles for deep learning in biology and medicine. *Journal of the Royal Society Interface* 15 (141): 20170387. <https://doi.org/10.1098/rsif.2017.0387>.
- 88 Greener, J.G., Kandathil, S.M., Moffat, L., and Jones, D.T. (2022). A guide to machine learning for biologists. *Nature Reviews Molecular Cell Biology* 23 (1): 40–55. <https://doi.org/10.1038/s41580-021-00407-0>.
- 89 Biamonte, J., Wittek, P., Pancotti, N. et al. (2017). Quantum machine learning. *Nature* 549 (7671): 195–202. <https://doi.org/10.1038/nature23474>.
- 90 Cerezo, M., Verdon, G., Huang, H.-Y. et al. (2022). Challenges and opportunities in quantum machine learning. *Nature Computational Science* 2 (9): 567–576. <https://doi.org/10.1038/s43588-022-00311-3>.
- 91 Sajjan, M., Li, J., Selvarajan, R. et al. (2022). Quantum machine learning for chemistry and physics. *Chemical Society Reviews* 51 (15): 6475–6573. <https://doi.org/10.1039/d2cs00203e>.
- 92 Sweke, R., Seifert, J.-P., Hangleiter, D., and Eisert, J. (2020). On the quantum versus classical learnability of discrete distributions. <http://arxiv.org/abs/2007.14451>.
- 93 Abbas, A., Sutter, D., Zoufal, C. et al. (2021). The power of quantum neural networks. *Nature Computational Science* 1 (6): 403–409. <https://doi.org/10.1038/s43588-021-00084-1>.
- 94 Caro, M.C., Huang, H.-Y., Cerezo, M. et al. (2022). Generalization in quantum machine learning from few training data. *Nature Communications* 13 (1): 4919. <https://doi.org/10.1038/s41467-022-32550-3>.
- 95 Kulkarni, V., Kulkarni, M., and Pant, A. (2020). Quantum computing methods for supervised learning. *arXiv:2006.12025 [quant-ph]*. <https://doi.org/10.48550/ARXIV.2006.12025>. <http://dx.doi.org/10.48550/ARXIV.2006.12025>.
- 96 Radic, M. (2019). Quantum-enhanced Machine Learning in the NISQ era. <https://elib.uni-stuttgart.de/handle/11682/10642> (accessed 4 September 2023).
- 97 Aïmeur, E., Brassard, G., and Gambs, S. (2006). *Machine Learning in a Quantum World*, 431–442. Berlin, Heidelberg: Springer-Verlag. ISBN: 9783540220046
- 98 Arrazola, J.M., Delgado, A., Bardhan, B.R., and Lloyd, S. (2020). Quantum-inspired algorithms in practice. *Quantum* 4 (307): 307. <https://doi.org/10.22331/q-2020-08-13-307>.

- 99 Federal Office for Information Security (2022). Quantum Machine Learning –State of the Art and Future Directions. <https://www.bsi.bund.de/EN/Service-Navi/Publikationen/Studien/QML/QML.html> (accessed 29 August 2023).
- 100 Melnikov, A.A., Nautrup, H.P., Krenn, M. et al. (2018). Active learning machine learns to create new quantum experiments. *Proceedings of the National Academy of Sciences of the United States of America* 115 (6): 1221–1226. <https://doi.org/10.1073/pnas.1714936115>.
- 101 Harrow, A.W., Hassidim, A., and Lloyd, S. (2008). Quantum algorithm for solving linear systems of equations. <http://arxiv.org/abs/0811.3171>.
- 102 Zhao, Z., Pozas-Kerstjens, A., Rebentrost, P., and Wittek, P. (2019). Bayesian deep learning on a quantum computer. *Quantum Machine Intelligence* 1 (1–2): 41–51. <https://doi.org/10.1007/s42484-019-00004-7>.
- 103 Date, P. and Potok, T. (2021). Adiabatic quantum linear regression. *Scientific Reports* 11 (1): 21905. <https://doi.org/10.1038/s41598-021-01445-6>.
- 104 Arthur, D. and Date, P. (2021). Balanced k-means clustering on an adiabatic quantum computer. *Quantum Information Processing* 20, 294 (9): <https://doi.org/10.1007/s11128-021-03240-8>.
- 105 Tomesh, T., Gokhale, P., Anschuetz, E.R., and Chong, F.T. (2021). Coreset clustering on small quantum computers. *Electronics* 10 (14): 1690. <https://doi.org/10.3390/electronics10141690>.
- 106 Havlicek, V., Corcoles, A.D., Temme, K. et al. (2018). Supervised learning with quantum enhanced feature spaces. <http://arxiv.org/abs/1804.11326>.
- 107 Qiskit (2017). <https://qiskit.org/documentation/stable/0.24/tutorials/machinelearning/01qsvmlclassification.html> (accessed 4 September 2023).
- 108 Schuld, M. (2021). Supervised quantum machine learning models are kernel methods. <http://arxiv.org/abs/2101.11020>.
- 109 Mitarai, K., Negoro, M., Kitagawa, M., and Fujii, K. (2018). Quantum circuit learning. *Physical Review A* 98 (3): 032309. <https://doi.org/10.1103/physreva.98.032309>.
- 110 Cerezo, M., Arrasmith, A., Babbush, R. et al. (2020). Variational quantum algorithms. <http://arxiv.org/abs/2012.09265>.
- 111 Dayan, P., Hinton, G.E., Neal, R.M., and Zemel, R.S. (1995). The Helmholtz machine. *Neural Computation* 7 (5): 889–904. <https://doi.org/10.1162/neco.1995.7.5.889>.
- 112 van Dam, T.J., Neumann, N.M.P., Phillipson, F., and van den Berg, H. (2020). Hybrid Helmholtz machines: a gate-based quantum circuit implementation. *Quantum Information Processing* 19 (6): <https://doi.org/10.1007/s11128-020-02660-2>.
- 113 Coyle, B., Mills, D., Danos, V., and Kashefi, E. (2020). The Born supremacy: quantum advantage and training of an Ising Born machine. *npj Quantum Information* 6 (1): <https://doi.org/10.1038/s41534-020-00288-9>.
- 114 Riofrio, C., Doetsch, J., Ehmer, T. et al. (2023). A performance characterization of quantum generative models. <https://arxiv.org/abs/2301.09363>.
- 115 Gili, K., Hibat-Allah, M., Mauri, M. et al. (2022). Do quantum circuit born machines generalize? <http://arxiv.org/abs/2207.13645>.

- 116** Outeiral, C., Strahm, M., Shi, J. et al. (2021). The prospects of quantum computing in computational molecular biology. *Wiley Interdisciplinary Reviews: Computational Molecular Science* 11 (1): e1481. <https://doi.org/10.1002/wcms.1481>.
- 117** Langione, M., Bobier, J.-F., Meier, C. et al. (2019). *Will Quantum Computing Transform Biopharma R&D?* Boston Consulting Group.
- 118** Alford, R.F., Leaver-Fay, A., Jeliaskov, J.R. et al. (2017). The Rosetta all-atom energy function for macromolecular modeling and design. *Journal of Chemical Theory and Computation* 13 (6): 3031–3048.
- 119** Koga, N., Tatsumi-Koga, R., Liu, G. et al. (2012). Principles for designing ideal protein structures. *Nature* 491 (7423): 222–227.
- 120** Kuhlman, B., Dantas, G., Ireton, G.C. et al. (2003). Design of a novel globular protein fold with atomic-level accuracy. *Science* 302 (5649): 1364–1368.
- 121** Gonen, S., DiMaio, F., Gonen, T., and Baker, D. (2015). Design of ordered two-dimensional arrays mediated by noncovalent protein-protein interfaces. *Science* 348 (6241): 1365–1368.
- 122** Hsia, Y., Bale, J.B., Gonen, S. et al. (2016). Design of a hyperstable 60-subunit protein icosahedron. *Nature* 535 (7610): 136–139.
- 123** King, N.P., Sheffler, W., Sawaya, M.R. et al. (2012). Computational design of self-assembling protein nanomaterials with atomic level accuracy. *Science* 336 (6085): 1171–1174.
- 124** King, N.P., Bale, J.B., Sheffler, W. et al. (2014). Accurate design of co-assembling multi-component protein nanomaterials. *Nature* 510 (7503): 103–108.
- 125** Tinberg, C.E. and Khare, S.D. (2017). Computational design of ligand binding proteins. In: *Computational Protein Design*, Methods in Molecular Biology, vol. 1529 (ed. I. Samish), 363–373. New York: Humana Press.
- 126** Tinberg, C.E., Khare, S.D., Dou, J. et al. (2013). Computational design of ligand-binding proteins with high affinity and selectivity. *Nature* 501 (7466): 212–216.
- 127** Fleishman, S.J., Whitehead, T.A., Ekiert, D.C. et al. (2011). Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science* 332 (6031): 816–821.
- 128** Strauch, E.-M., Bernard, S.M., La, D. et al. (2017). Computational design of trimeric influenza-neutralizing proteins targeting the hemagglutinin receptor binding site. *Nature Biotechnology* 35 (7): 667–671.
- 129** Gordon, S.R., Stanley, E.J., Wolf, S. et al. (2012). Computational design of an α -gliadin peptidase. *Journal of the American Chemical Society* 134 (50): 20513–20520.
- 130** Siegel, J.B., Zanghellini, A., Lovick, H.M. et al. (2010). Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. *Science* 329 (5989): 309–313.
- 131** Bhardwaj, G., Mulligan, V.K., Bahl, C.D. et al. (2016). Accurate de novo design of hyperstable constrained peptides. *Nature* 538 (7625): 329–335.

- 132 Dang, B., Wu, H., Mulligan, V.K. et al. (2017). De novo design of covalently constrained mesosize protein scaffolds with unique tertiary structures. *Proceedings of the National Academy of Sciences of the United States of America* 114 (41): 10852–10857.
- 133 Drew, K., Renfrew, P.D., Craven, T.W. et al. (2013). Adding diverse non-canonical backbones to Rosetta: enabling peptidomimetic design. *PLoS One* 8 (7): e67051
- 134 Hosseinzadeh, P., Bhardwaj, G., Mulligan, V.K. et al. (2017). Comprehensive computational design of ordered peptide macrocycles. *Science* 358 (6369): 1461–1466.
- 135 Renfrew, P.D., Choi, E.J., Bonneau, R., and Kuhlman, B. (2012). Incorporation of noncanonical amino acids into Rosetta and use in computational protein-peptide interface design. *PLoS One* 7 (3): e32637
- 136 Kuhlman, B. and Baker, D. (2000). Native protein sequences are close to optimal for their structures. *Proceedings of the National Academy of Sciences of the United States of America* 97 (19): 10383–10388.
- 137 Lao, B.B., Drew, K., Guarracino, D.A. et al. (2014). Rational design of topographical helix mimics as potent inhibitors of protein–protein interactions. *Journal of the American Chemical Society* 136 (22): 7877–7888.
- 138 Charpentier, A., Mignon, D., Barbe, S. et al. (2018). Variable neighborhood search with cost function networks to solve large computational protein design problems. *Journal of Chemical Information and Modeling* 59 (1): 127–136.
- 139 Donald, B.R. (2011). *Algorithms in Structural Molecular Biology*. MIT Press.
- 140 Gordon, D.B. and Mayo, S.L. (1999). Branch-and-terminate: a combinatorial optimization algorithm for protein design. *Structure* 7 (9): 1089–1098.
- 141 Leach, A.R. and Lemon, A.P. (1998). Exploring the conformational space of protein side chains using dead-end elimination and the A* algorithm. *Proteins: Structure, Function, and Bioinformatics* 33 (2): 227–239.
- 142 Traoré, S., Allouche, D., André, I. et al. (2013). A new framework for computational protein design through cost function network optimization. *Bioinformatics* 29 (17): 2129–2136.
- 143 Feynman, R.P. (1986). Quantum mechanical computers. *Foundations of Physics* 16 (6): 507–532.
- 144 Kadowaki, T. and Nishimori, H. (1998). Quantum annealing in the transverse Ising model. *Physical Review E* 58 (5): 5355.
- 145 Galda, A., Mulligan, V., MacCormack, I. et al. (2022). Peptide design with quantum approximate optimization algorithm. *Bulletin of the American Physical Society*.
- 146 Farhi, E., Goldstone, J., and Gutmann, S. (2014). A quantum approximate optimization algorithm. *arXiv preprint arXiv:1411.4028*.
- 147 Susskind, L. (2011). <https://theoreticalminimum.com/courses/quantum-mechanics/2012/winter> (accessed 4 September 2023).
- 148 Sutor, R. (2019). *Dancing with Qubits - How quantum computing works and how it can change the world*. Packt.

- 149 Johnston, E.R., Harrigan, N., and Gimeno-Segovia, M. (2019). *Programming Quantum Computers*. O’Rilly. ISBN: 978-1-492-03968-6
- 150 Coecke, B. and Kissinger, A. (2017). *Picturing Quantum Processes: A First Course in Quantum Theory and Diagrammatic Reasoning*. Cambridge University Press. <https://doi.org/10.1017/9781316219317>.
- 151 Coecke, B. and Gogioso, S. (2023). *Quantum in Pictures*. Cambridge Quantum. ISBN: 978-1739214715
- 152 Hurley, W. and Smith, F. (2023). *Quantum Computing for Dummies*. Wiley. ISBN 9781119933908.
- 153 Ezratty, O. (2021). Understanding quantum technologies. <https://doi.org/10.48550/arXiv.2111.15352>.
- 154 TU Delft (2015). <https://qutech.nl/> (accessed 29 August 2023).
- 155 Massachusetts Institute of Technology (2016–2022). Quantum computing. <https://learn-xpro.mit.edu/quantum-computing> (accessed 29 August 2023).
- 156 Quarks Interactive (2021). Quantum Odyssey. <https://www.quarksinteractive.com/> (accessed 9 December 2023).
- 157 Seskir, Z.C., Migdał, P., Weidner, C. et al. (2022). Quantum games and interactive tools for quantum technologies outreach and education: a review and experiences from the field. *arXiv*.
- 158 Malone, F.D., Parrish, R.M., Welden, A.R. et al. (2022). Towards the simulation of large scale protein–ligand interactions on NISQ-era quantum computers. *Chemical Science* 13: 3094–3108. <https://doi.org/10.1039/D1SC05691C>.
- 159 Kirsopp, J.J.M., Di Paola, C., Manrique, D.Z. et al. (2021). Quantum computational quantification of protein–ligand interactions. <http://arxiv.org/abs/2110.08163>.
- 160 Fox, D.M., Branson, K.M., and Walker, R.C. (2021). mRNA codon optimization with quantum computers. *PLoS One* 16 (10): e0259101. <https://doi.org/10.1371/journal.pone.0259101>.
- 161 Robert, A., Barkoutsos, P.K., Woerner, S., and Tavernelli, I. (2021). Resource-efficient quantum algorithm for protein folding. *npj Quantum Information* 7 (1): 1–5. <https://doi.org/10.1038/s41534-021-00368-4>.
- 162 Maniscalco, S., Borrelli, E.-M., Cavalcanti, D. et al. (2022). Quantum network medicine: rethinking medicine with network science and quantum algorithms. *arXiv [quant-ph]*. <http://arxiv.org/abs/2206.12405>.
- 163 Santagati, R., Aspuru-Guzik, A., Babbush, R. et al. (2023). Drug design on quantum computers. <http://arxiv.org/abs/2301.04114v1>.
- 164 Izsak, R., Riplinger, C., Blunt, N.S. et al. (2023). Quantum computing in pharma: a multilayer embedding approach for near future applications. *Journal of Computational Chemistry* 44 (3): 406–421. <https://doi.org/10.1002/jcc.26958>.

Index

a

- AAFAA polypeptide 184
- ab initio calculations 163
- absorption, distribution, metabolism, excretion, and toxicity (ADMET)
 - Orion 600
 - QSAR 498–499
 - tox prediction 220
- acceleration methods 44, 460
- acetazolamide (AZM) 169
- active pharmaceutical ingredient (API) 432
- adaptive multi-splitting approach 55, 57
- adenosine receptors (ARs) 28, 29, 31–33, 53
- ADME IQ Consortium 510
- AF₃₀ 265
- AI-based methods 228, 236–243, 277–278, 290, 369
- AI-based protein models, in structural biology
 - challenges and opportunities 243–246
 - with computational approaches 236–243
 - with Cryo-EM and X-ray crystallography 229–232
 - deep learning models 235, 236
 - with mass spectrometry 234, 235
 - with NMR structures 232–234
- alchemical methods
 - Bennet's Acceptance Ratio 10
 - challenges 13–15
 - multiple compounds 11, 12
 - non-equilibrium methods 11
 - one-step perturbation approach 12, 13
 - thermodynamic integration 10
- allosteric modulation, of human A₁ adenosine receptor 29–32
- AlphaFill 244
- AlphaFold2 model 227, 228, 230, 231, 233–235, 238–242, 245, 265, 590
- AlphaFold2 modification
 - for accurate free energy prediction 265
 - multiple sequence alignment 265–266
- AlphaFold2 neural network 233, 365
- AlphaFold2 prediction confidence score (pLDDT) 230
- AlphaFold 175, 213, 227, 228, 230, 233, 238, 241–243, 245, 246, 369, 444, 451, 543, 590
- AlphaFold-Multimer 243, 265, 266
- alvaBuilder 378–380
- alvaDesc 380
- AM1/d-phot QM Hamiltonian 571
- Amazon Web Services (AWS) 581
- amino acid-polyamine-organocation transporter (GadC) 235
- ANSURR 233
- antimicrobial resistance (AMR) evaluation 133–138
- apolar molecules 86
- Apple Silicon chips 581
- applicability domains 279, 300–308, 370, 409, 499, 507–509, 516, 519, 522
- application programming interfaces (APIs) 177, 266, 399
- approximate k-nearest neighbor (ANN) 348
- ARM chip technologies 581
- artificial intelligence (AI)
 - bioactivity data
 - databases annotated 368
 - ligand-based drug design 368–369
 - structure-based drug design 369
 - and machine learning-based tools 290
 - models 498–499
 - web-based tools 369
- AstraZeneca (AZ) 326, 420, 501, 502, 510, 516, 520
- AtomNet PoseRanker (ANPR) 264

- augmented intelligence 368, 383
- AutoDesigner 329–331
- AutoDock Vina 263, 455, 476
- AutodockGPU 267
- automated QM refinement technique 158
- AutoQSAR/DeepChem 460
- auxiliary polarization (AP) method 187
- b**
- Bennett Acceptance Ratio (BAR) 9, 10
- BERT-style language model 266
- β -lactamases (BLs) 133, 134, 137, 138, 452, 459, 460, 572
- biased docking 431
- biased MD method
 - bias force-based methods 50
 - bias potential-based methods 49, 50
 - coarse-graining and master equation 51
 - knowledge-biased methods 50, 51
 - temperature-and barrier-scaling 49
- bias force-based methods 50
- bias potential-based methods 49, 50
- bile salt export pump (BSEP) inhibition 502
- binding database (BindingDB) 397, 400
- binding energy 32, 69, 71, 75, 77, 124, 125, 141, 186, 197, 198, 200, 472, 482
- binding free energy calculations 3
 - alchemical methods
 - Bennett's Acceptance Ratio 10
 - challenges 13–15
 - free energy perturbation 9
 - multiple compounds 11, 12
 - nonequilibrium methods 11
 - one-step perturbation approach 12, 13
 - thermodynamic integration 10
 - endpoint methods
 - linear response approximation 7–9
 - MM/GBSA 5–7
 - MM/PBSA 5–7
 - pathway methods 3, 15–17
- binding free energy prediction approach 255
- binding kinetics
 - prediction from GaMD simulations 37–39
 - reweighting of GaMD simulations 26–27
- BintScore 597, 599, 600
- bioactive compounds
 - antituberculosis chemical dataset 376–377
- de novo design
 - DNMT-focused libraries 379, 380
 - libraries 377–379
 - structure-based or ligand-based approaches 377
 - virtual screening 373
 - VISAS 373–380
- BioAssay Ontology (BAO) 402
- bioisostere replacement 320, 321
- bit collision 346
- Bloch sphere 638, 639, 645
- blood brain barrier (BBB) applicability
 - domain 302, 303
- Boltzmann constant 73, 566
- Boltzmann-enhanced discrimination of receiver operating characteristic (BEDROC) score 478, 479
- bond detached atom (BDA) 186
- Born Haber cycle 67, 68
- Born machines 660, 661
- bovine serum albumin (BSA) 97–106
- Bruton's tyrosine kinase (BTK) 129, 131
- c**
- caffeine binding and unbinding, in human
 - adenosine A_{2A} receptor 28, 29
- calcineurin-cyclosporin A-cyclophilin ternary complex 539
- Cambridge Structural Database (CSD)
 - drug discovery
 - hit identification 425–428
 - hit-to-lead 428, 429
 - lead optimisation 430–432
 - target identification and target validation 422–425
 - drug products 432
 - HBP method 433
 - impact drug discovery 433, 434
 - tools 420
- carbapenemase 135, 136, 572
- cardiac troponin C, small-molecule calcium sensitizers for 37
- Car-Parinello MD (CP-MD) 127
- cartridges 341, 342, 355–357
- CavVis 422
- ceftazidime 136
- cell permeability 220
- CHARMM General Force Field (CGenFF)
 - program 97
- CHARMM/TIP3P force field 571
- Chemaxon 344
- ChEMBL database 323, 369, 379, 397, 399
- Chemical Abstract Service (CAS) 396
- Chemical Computing Group (CCG) 619

- chemical databases 318, 323, 328, 338, 341, 368, 370, 373, 396, 474, 479, 497
- chemical data search 344, 345
- chemical functional groups 349
- chemical hashed fingerprint (CFP) 346
- chemical libraries, public and commercial
 - available building blocks 328–329
 - reaction sources 327–328
 - vendor libraries 323–326
- chemical library networks (CLNs) 370, 382–383
- chemical multiverse 366, 369–372, 384
- chemical space 317, 338
 - bioisosteres 320, 321
 - chemical multiverse 370–372
 - compound registration systems 343
 - concept 370
 - constellation plots 370–372
 - corporate data warehouses 341, 342
 - definition 370
 - enumeration within de novo design 321–323
 - exploration approach 329, 331
 - matched molecular pairs 320, 321
 - novel compound design 342, 343
 - project data exploration 341
 - R-group enumeration 319, 320
 - SAR-by catalog 343, 344
 - scaffold enumeration 319, 320
 - size of 317, 320
 - simple SMILES enumeration 317–319
 - string operations 317–319
 - web servers and applications 370
- chemical space networks (CSNs) 382
- chemical vendors 324, 592
- chemoinformatics 56, 367, 368, 370, 396
- ChemSpider 399
- ciprofloxacin 340
- Clashscore 162, 166, 168
- class-conditional conformal predictors 307, 308
- classical information on classical hardware 657
- cloud computing 409, 516, 592, 605, 608, 609, 618, 619, 624, 625
- CNOT gate 646
- coarse-graining and master equation approaches 51
- ColAttn 266
- collective variable (CV) 47
- combined QM/MM energy calculations 123
- commercial libraries
 - CHEMriya 448
 - eXplore 449
 - Freedom Space 449
 - GalaXi 448
 - REAL Database and REAL Space 447, 448
 - VirtualFlow libraries 449, 450
 - ZINC library 449
- commercial SAR databases
 - CDDI 403
 - GOSTAR 401, 402
 - Reaxys Medicinal Chemistry (RMC) 402, 403
- Common Terminology Criteria for Adverse Events (CTCAE) format 220
- Community Structure Activity Resource (CSAR) data set 170
- complex molecular mechanics 619–621
- compound registration systems 343
- computer-aided drug design (CADD)
 - chemical structures of 365, 366
 - concepts and applications of 366
 - co-solvent technologies 85
 - identification and development 365
 - ligand-based drug design methods 365
 - machine learning (ML) algorithms 365
 - SBDD and LBDD 366
- ConfChangeMover (CCM) 235
- conformal predictors 307
- Conquest 420, 427, 428
- constellation plots 366, 370–372, 375, 377, 384
- Continuous Evaluation of Ligand Protein Predictions (CELPP) 485, 486
- conventional MD (cMD) 21, 623
- conventional QSAR model development
 - algorithms 504–506
 - applicability domain and reliability 507–509
 - chemical descriptors 503, 504
 - components 499
 - data curation
 - biological assay data 501–503
 - curating chemical structures 500
 - model deployment and accessibility 509
 - model interpretability vs. predictivity 509
 - model validation and evaluation metrics 506–507
 - performance 507
- conventional refinement methods 158, 159
- convolutional long-short term memory (ConvLSTM) 305
- convolutional neural network (CNN)
 - applications 259, 260

- convolutional neural network (CNN) (*contd.*)
 - descriptors 259
 - voxelized grid representation 258, 259
 - corehopping 320, 321
 - corporate data warehouses 341, 342
 - Cortellis Drug Discovery Intelligence (CDDI) 403
 - co-solvent MD methods
 - SILCS 90–106
 - solutes and their concentrations 87–88
 - strength of 89
 - 3D maps generated from 89
 - covalent inhibitors
 - advantages of 561
 - definition 561
 - irreversible
 - case studies of 571–574
 - dissociation constant of 566, 567
 - rate constant of the covalent step 567–569
 - reversible 569–571
 - SARS-CoV-2 138–143
 - two-step model for 128
 - covalent tyrosine kinase inhibitors 128–133
 - COVID-19 35, 76, 138, 143, 420, 621, 622
 - CRBN-containing ternary complexes 542, 547, 554, 555
 - Critical Assessment of Computational Hit-finding Experiments (CACHE) 258, 483–485
 - Crooks Gaussian intersection (CGI) 11
 - cross docking benchmark dataset 476
 - cryogenic electron microscopy (Cryo-EM) 176
 - AI-based protein model prediction 229–232, 424, 587
 - RS-DivCon refinement preliminary results 176
 - crystal structure prediction (CSP) 582, 605
 - Crystallographic Information File (CIF) 158, 421
 - CSD Conformer Generator 421, 427, 428
 - CSD-CrossMiner 424, 425, 428, 429, 431, 434
 - Cytochrome P450 409, 499
- d**
- database cartridges 342
 - database examples 339
 - data-driven drug discovery 367, 406, 408
 - DCAF15 539, 541, 546, 549, 550
 - DeepBindBC ResNet binary classifier 263
 - DeepChem 368, 460
 - deep docking 458–459
 - deep learning (DL)
 - approaches to molecular docking 461
 - architecture, of GLOW 26
 - deep learning methods 228, 230, 232, 236, 239, 243, 244, 257, 279
 - deep learning models
 - convolutional neural network 258–260
 - datasets 258
 - graph neural network 260, 261
 - molecular dynamics simulations
 - applications 263, 264
 - enhanced sampling 262
 - physics-inspired neural networks 262, 263
 - for protein–ligand complex 257–258
 - quality affecting issues
 - folding pathways 245
 - fold-switching proteins 245
 - glycosylation 245
 - intrinsically disordered proteins/regions 244, 245
 - mutations 245
 - structural biology 235, 236
 - deep neural network (DNN) models 227, 228, 236, 243, 278, 280, 281, 299, 302, 499, 517, 657, 658
 - DeepTox 369
 - de novo design 243, 321–323, 329, 365, 368, 370, 377–380, 384, 419, 520
 - de novo drug design 275–290, 369, 379, 648
 - de novo protein design 243
 - density functional theory (DFT) 123, 185, 266, 605
 - density-functional tight-binding (DFTB) 183
 - deoxy-D-arabino-heptulosonate-7-phosphate synthase (DAHPS) 427
 - Design-Make-Test-Analyze (DMTA) cycle 276, 277, 430
 - design unit (DU) 588–590
 - digital annealer 648, 657
 - Dirac notation 632
 - Directory of Useful Decoys (DUD) 476
 - dispersion (DI) interaction 187
 - dissipation-corrected targeted MD (dcTMD)
 - approach 55
 - distance metrics (D) 302, 345, 350
 - divanillin binding 101, 102

DivCon 158–160, 162, 176, 177, 623
 divide-and-conquer (D&C) quantum mechanics 159, 592
 DL-based binding affinity computation 267
 distributed training and inference
 data and model parallelism 267, 268
 federated learning 268
 single GPU optimizations 267
 DNA-dependent protein kinase catalytic subunit (DNA-PKcs) 424
 DNA methyltransferase (DNMT) 371
 docking-based approaches 215
 docking-based ultra-large virtual screenings
 Alon et al. (2021) 453
 Gorgulla (2018) 451, 452
 Gorgulla et al. (2020) 453
 Lyu et al. (2019) 452, 453
 Stein et al. (2020) 453
 USCF DOCK 454
 VirtualFlow 455
 docking-based virtual screening 371
 docking enrichment (DE) 32, 479
 docking programs 220, 444, 446, 452, 454, 455, 461, 593, 623
 DockQ scores 265
 double-strand breaks (DSBs) 424
 DrugBank 75, 76, 326, 399–401, 420
 drug design and discovery process 275, 276
 drug design data resource (D3R) 483
 drug-induced liver injury (DILI) 512
 DrugMAP 243
 drug metabolism and pharmacokinetics (DMPK) 397, 408, 523
 drug–target interactions (DTIs)
 graph-based methods 217, 218
 hit identification 213–215
 machine learning methods 215, 216
 dual-boost GaMD 26, 32, 33
 dual-boost LiGaMD 25
 DUD-E database 477

e

E3 ligase 539, 541, 549, 619
 electrostatic interaction screened in solution (ES+es) 187
 empirical valence bond (EVB) approach 122, 570
 energy decomposition analysis (EDA) 184
 5-enolpyruvylshikimate 3-phosphate synthase (EPSPS) 427
 enrichment factor 32, 277, 478, 591

ensemble-based virtual screening, of
 allosteric modulators of human A₁
 adenosine receptor 32, 33
 enumerated libraries of synthesizable compounds 338
 error-based estimation methods 301
 European Molecular Biology Laboratory (EMBL) 397
 Everett's multiverse 371
 exchange-repulsion (EX) 186, 187
 explainable ML methods 218
 explicit solvent 69, 84, 184, 185, 202, 238, 262
 extended connectivity fingerprint (ECFP) 219, 279, 346, 348, 349, 380
 extended similarity methods
 chemical diversity 382, 383
 chemical library networks 382, 383
 and clustering 382, 383
 diversity selection and medoid calculation 383
 framework 381, 382
 extended Tanimoto index 382

f

faster free energy 622–624
 fast Fourier Transforms (FFT) 96, 581
 FDA's Adverse Event Reporting System (FAERS) database 220
 federated learning (FL) 268, 279, 290
 figures of merit (FoM) 480
 File Transfer Protocol (FTP) 398
 findability, accessibility, interoperability, and reusability (FAIR) 398
 fingerprint generation methods
 features 346–349
 substructure-preserving fingerprints 345, 346
 flavin mononucleotide (FMN) 234
 Floes 585, 588, 593, 594, 596, 601–606
 flow-based programming 585
 fluctuation analysis (FA) 201
 FMO3/EDA 188
 FoldDock 243, 265
 4D Bioavailability (4DBA) metric 94
 F-pocket 590
 fragment-based drug discovery (FBDD) 157, 183, 232, 343
 fragment-based screening 214, 482
 fragment molecular orbital (FMO) method 184, 185
 free energy 4, 46
 alchemical 13

- free energy (*contd.*)
 - definition 4
 - GaMD, energetic reweighting of 25, 26
 - GLOW 26
 - modifying AlphaFold2 input protein database 265
 - perturbation 9
 - and thermodynamic cycles 4, 5
- free energy decomposition analysis (FEDA) 202
- free-energy perturbation (FEP) 9, 83, 84, 124, 132, 144, 240, 265, 266, 566, 569, 597, 598
- frozen domain (FD) formulation 203
- FTrees 349, 358
- functional class fingerprints (FCFP) 347
- g**
 - gate quantum computing 639, 640, 644, 645
 - Gaussian accelerated molecular dynamics (GaMD)
 - applications
 - binding kinetics prediction from 37–39
 - cardiac troponin C, small-molecule calcium sensitizers for 37
 - G-protein-coupled receptors 28–33
 - human angiotensin-converting enzyme 2 receptor 35–37
 - nucleic acids 33–35
 - binding kinetics from 26, 27
 - energetic reweighting for free energy calculations 25, 26
 - implementations and software 27, 28
 - ligand 24, 25
 - spontaneous binding of risdiplam 34
 - GBVI/wSA score 171–175, 543, 545, 546, 549, 551
 - Genentech 484, 499, 501, 502, 508, 510, 512, 513, 515, 516, 518, 519
 - generated databases (GDBs) 319, 450
 - generalized Born (GB) 5, 6, 69, 83, 262
 - generative adversarial networks (GANs) 660
 - generative AI
 - fragment growing 283, 284
 - in lead optimization 281–283
 - molecular design 281
 - novelty generation 283–285
 - synthetic scores 288, 289
 - geometric deep learning approach 261
 - Gestodene 605–608
 - GFE FragMaps 91, 95, 98–101
 - GHECOM 422
 - Gibbs free energy 4, 66–68, 70, 71, 196, 566
 - Gigadock Floe 593
 - Gigadock Warp 594
 - GlobalDistanceTest_TotalScore (GDT_TS) 227
 - global online structure–activity relationship (GOSTAR) 397, 401, 402
 - GLOW 22, 26, 29, 32
 - G-protein-coupled receptors (GPCRs)
 - allosteric modulation of human A₁ adenosine receptor 29–32
 - caffeine binding and unbinding, in human adenosine A_{2A} receptor 28, 29
 - classification 28
 - ensemble based virtual screening 32, 33
 - GPU-accelerated docking algorithms 267
 - GPU-accelerated free-energy perturbation simulations 266
 - graph-based methods for DTI prediction 217
 - grand challenge (GC) 483
 - graph attention (GAT) networks 261, 519
 - graph convolutional networks (GCNs) 219, 261, 263, 517
 - graphics processing units (GPUs) 80, 267, 409, 454, 461, 581
 - graph matching algorithm 354, 355
 - graph neural networks (GNNs)
 - applications 260, 261
 - descriptors 260
 - extension to attention based models 261
 - geometric deep learning approach 261
 - graph representation 260
 - grid inhomogeneous solvation theory (GIST) 70, 72, 74, 76–78
 - GridMarkets platform 619, 620, 623, 625
 - GRIMscore metric 477
 - Group-I p21-activated kinase 1 (PAK1) 75
 - GROW algorithm 322
 - Guide to Pharmacology 406, 589
- h**
 - Hadamard gate 641, 644–646
 - Hartree–Fock (HF) 186, 188
 - high-throughput screening (HTS) 128, 144, 157, 213, 275, 372, 398, 425, 443, 471, 593–594, 662
 - hit identification 83, 213–215, 255, 341, 373–380, 425–428, 434, 471, 473, 483, 489, 583, 591, 662

human A₁ adenosine receptor
 ensemble based virtual screening of
 allosteric modulators 32, 33
 modulation of 29–32
 human adenosine A_{2A} receptor, caffeine
 binding and unbinding in 28, 29
 human angiotensin-converting enzyme 2
 receptor 35–37
 human ether-a-go-go-related gene 502
 human liver microsomal (HLM) 499
 human liver microtissue (hLiMT) assays
 502
 Human Metabolome Database (HMDB)
 401
 hybrid DFT functionals 126
 hydrogen bond acceptors (HBA) 219, 317,
 324, 347, 423, 425, 477
 hydrogen bond donors (HBD) 52, 90, 100,
 219, 317, 324, 423
 hydrophobic effect 57, 65

i

idiosyncratic toxicity 561
 imatinib 46, 340
 IMiD-CRBN complexes 539
 immunomodulatory drugs (IMiDs) 539
 immunophilins 539
 implicit solvent 69, 262
 induced fit docking coupled with molecular
 dynamics (IFD-MD) 238
 induced-fit theory 472
 inhomogeneous solvation theory (IST) 70
 Inpharmatica 397, 399
 in silico ADME models
 ADMET QSAR models 513
 generative models 519, 520
 in silico toxicity models 512–515
 mechanism-based models 520, 521
 intermolecular energy force field (IEFF)
 606
 intramolecular graph transformer (IGT)
 261
 intramolecular hydrogen bonds (IHMB)
 219
 intrinsic reaction coordinate (IRC) 138,
 203
 in vitro-in vivo correlation (IVIVc) 498
 irreversible covalent inhibition
 case studies of 571–574
 dissociation constant of the noncovalent
 step 566, 567
 rate constant of the covalent step
 567–569

ISOLDE 230
 IsoStar 421, 424, 429, 430
 isothermal titration calorimetry (ITC) 68,
 77
 iterative learning cycle 511, 522

j

JChem PostgreSQL cartridge (JPC) 356,
 357

k

Kendall Rank correlation 481, 482
 Kendall's tau coefficient 481, 483
 kernel-based QSAR 301
 α -ketoamide derivatives 569, 570
 kinase knowledge base (KKB) 401, 403,
 404
 Kinase–Ligand Interaction Fingerprints and
 Structure (KLIFS) database 401
 KinomeX 368
 knowledge-based compound design 408
 knowledge-biased methods 50, 51
 KnowledgeSpace 327, 450–451
 Konstanz Information Miner (KNIME)
 399
 Kramers rate theory 27, 39, 48, 49, 56

l

language-based models, for binding affinity
 prediction 266
 lead optimization 4, 9, 11, 12, 17, 80, 83,
 124, 228, 255, 275, 281–283, 288, 299,
 331, 338, 340, 341, 375, 403, 422, 426,
 430–432, 471, 473, 510, 511, 514, 519,
 520, 582, 591, 592, 595, 596, 599
 lead optimization mapper (LOMAP) 11
 ligand-based binding free energy prediction
 approach 255
 ligand-based drug design (LBDD) 83, 365,
 366, 368–369
 ligand-based virtual screening (LBVS) 446,
 581, 591
 ligand Gaussian accelerated molecular
 dynamics (LiGaMD) 24, 25, 35, 39,
 57
 ligand grid free energy (LGFE) 91, 100, 102
 ligand libraries, preparation and 444, 446
 ligand strain energy 160, 161, 163, 166, 167
 LIGSITE 422, 423
 linear fingerprint techniques 349, 352
 linear interaction energy (LIE) method 8,
 69
 linear path-based hashed fingerprints 346

- linear response approximation (LRA)
 - framework 7–9
- linear-scaling QM semiempirical quantum mechanics (SE-QM) method 158
- linkerless PROTACs 541, 548–555
- Lipinski's Rule of Five 276, 317, 356, 448
- LIT-PCBA 477, 478
- local dynamics ensemble (LDE) 263
- lock-key model 472
- LogP model 379
- LogS aqueous solubility 379–380
- log-scaled fu (logfu) 503
- loratadine 432, 433
- m**
 - machine learning-based virtual screenings
 - AutoQSAR/DeepChem 460
 - Deep Docking 458, 459
 - MolPal 459, 460
 - machine learning (ML) methods
 - CADD 365
 - deep learning 216
 - explainable 218
 - supervised learning 216, 217
 - unsupervised methods 216, 217
 - MacroMolecular Data Service (MMDS) 587
 - MadFast Similarity Search 352
 - Madin-Darby canine kidney (MDCK) cells 501
 - Mahalanobis distance 301
 - Markov State Model (MSM) 51
 - Markush structure database 396
 - mass spectrometry 234, 235
 - matched molecular pair (MMP) 215, 281, 320–321, 330, 344, 374, 375, 402, 408, 497, 519
 - matched molecular pair analysis (MMPA) 215, 321, 374, 497
 - matched molecular series (MMS) 408, 497
 - maximum common substructure (MCS)
 - algorithm 550
 - maximum double cluster population 621
 - Maximum Unbiased Validation (MUV)
 - dataset 478
 - MD-refinement protocol 238
 - mechanical (ME) or electrostatic embedding (EE) schemes 568
 - Mercury software 432
 - metadynamics 21, 50, 54, 55, 127, 134, 262, 567, 568
 - metal-organic frameworks (MOFs) 420
 - MG-containing ternary complex crystal structures 541
 - MiniHashFingerprint (MHFP) 348
 - minimum energy pathway (MEP) 125
 - minimum information about a bioactive entity (MIABE) 407
 - MLN-4760 inhibitor, binding and dissociation of 38
 - MM-PB(GB)SA end-point method 124
 - Molecular ACCess System (MACCS) 219, 341, 346
 - molecular design 203, 281, 343, 497, 519, 520, 581–609
 - molecular docking
 - deep learning approaches 461
 - graphics processing units (GPUs) 461–462
 - molecular docking and scoring
 - benchmarking datasets for pose prediction 475, 476
 - benchmarking sets for pose prediction
 - Astex 476
 - cross docking 476
 - PDBbind 476
 - benchmarking sets for virtual screening
 - DUD and DUDE 477
 - LIT-PCBA 477, 478
 - MUV 478
 - community benchmarking exercises
 - CACHE 483–485
 - CELPP 485, 486
 - D3R 483
 - SAMPL challenge 482, 483
 - in drug discovery 473
 - evaluation metrics
 - BEDROC Score 479
 - docking enrichment (DE) 479
 - enrichment factor 478
 - Kendall Rank correlation 481, 482
 - receiver operating characteristic curve 480
 - RMSD 480
 - RMSE 481
 - RSR 480, 481
 - Spearman's rank-order correlation 481
 - lessons learned from benchmarking exercises
 - quality of crystal structures 486, 487
 - requirement 488, 489
 - statistics, error bars and confidence intervals (CI) 487, 488
 - sufficient metrics 487, 488

- multi-step process 472
 - protein–ligand interactions 472
 - software packages 472
 - workflow 473
 - molecular dynamics (MD)
 - simulations 45
 - time scale problem of 46, 47
 - molecular electrostatic potential (MEP) 185
 - molecular glue
 - CRBN-containing systems 546
 - linkerless PROTACs 541, 548–555
 - methodology 541, 542
 - mezigdomide 540
 - Per System basis 544
 - procedure 543
 - protein–protein docking 546, 547
 - protein score 545
 - serendipity 540
 - ternary complexes 543
 - molecular kinetics calculation, theory of
 - biased MD simulations 49–51
 - Kramers rate theory 48, 49
 - nonequilibrium statistical mechanics 47, 48
 - Molecular Libraries Program (MLP) 397
 - molecular mechanics generalized Born
 - surface area (MM/GBSA) 5–7, 69
 - molecular mechanics Poisson – Boltzmann surface area (MM/PBSA) 5–7, 69
 - molecular mechanics (MM) simulations 183
 - molecular operating environment (MOE)
 - LowModeMD method 552
 - Site Finder tool 552
 - Molecular Pool-based Active Learning (MolPal) 459
 - MolProbity score (MPScore) 162, 167
 - MoveableType (MT) methods 622, 623
 - MT_{ConfSearch}(ligand) 623
 - MT_{Dock} 623
 - MT_{Flex} 623
 - MT_{ScoreES}(EndState) 623
 - MT_{ScoreE}(Ensemble) 623
 - multiparameter optimization (MPO) 277, 329, 512
 - multiple copy simultaneous search (MCSS)
 - method 84
 - multiple sequence alignment (MSA) 228, 265–266
 - multiple solvent crystal structures (MSCS)
 - method 84
 - Multistate Bennett Acceptance Ratio (MBAR) 10
 - multitask (MT) learning 518
 - Musashi 1 (MSI1) RNA-binding protein 35, 36
- n**
- National Institute of Health (NIH) 397, 398
 - natural language processing (NLP) 220, 228, 256, 348, 410, 517, 631, 655
 - natural products 367, 368, 371, 373, 401, 539
 - Nelder–Mead optimization method 503
 - nested 5-fold cross validation 303
 - net efflux ratio (NER) 501
 - nicotinic acetylcholine receptors (nAChRs)
 - inhibitors 427
 - NIH Molecular Libraries Screening Centre Network (MLSCN) 398
 - no-clone theorem 644
 - nonequilibrium aspects, of drug binding and pharmacokinetics 46
 - nonequilibrium methods 11
 - nonequilibrium statistical mechanics 46–48
 - Non-Equilibrium Switching (NES) 582, 598, 599
 - non-Von-Neumann architecture 629, 636
 - normalized entropy 73
 - normal mode analysis (NMA) 7
 - nsp3 macrodomain 454
 - nuclear magnetic resonance (NMR) 229, 232–234, 662
- o**
- OEPOCKET 590
 - off-target reactivity 561
 - Omidenepag isopropyl (OMDI) 428
 - on-demand libraries 447–449, 457
 - 1D amino acid sequence 227, 243
 - one-step perturbation approach 12, 13
 - ONIOM method 123, 138
 - ONIOM refinement 165–168
 - open-source cartridges 356
 - OpenEye protein preparation tool 587
 - OpenEye Scientific Software 353
 - OpenEye’s ROCS technology 590
 - organic anion-transporting polypeptide 1B1 transporter (OATP) 499
 - orientational entropy 73
 - Orion
 - ADMET prediction and permeability 600–604

Orion (*contd.*)

- drug crystal forms prediction 604–608
- ligand-based virtual screening (LBVS) 581
- Platform 582–585
- predicting small-molecule binding affinity 595–600
- target preparation & structural data organization 585–590
- virtual screening 591–595
- Orion Floe 588, 594, 605
- Ornstein – Zernike equation 70
- orthosteric or allosteric site 588
- osimertinib 129–131, 572
- overall crystallographic structure quality metrics 162
- oxathiazolones 573

p

- pair energy decomposition analysis (PIEDA) 186
 - applications of 189
 - example of 189–192
 - formulation of 186–189
- pair interaction energy (PIE) 185, 186
- partial geometry optimizations 185
- partial MG 548–556
- particle mesh Ewald (PME) method 581
- partition analysis (PA)
 - applications and example 194, 195
 - formulation of 193, 194
- partition analysis of vibrational energies (PAVE)
 - applications of 196, 197
 - example of 200, 201
 - formulation of 196
- PDBbind 258, 264, 475, 476, 623
- pembrolizumab 97–106
- penicillin-binding proteins (PBPs) 133
- Pep-GaMD 22, 37, 39
- peptidyl SARS-CoV-2 main protease inhibitor 140
- periodic boundary conditions (PBC) 13, 185
- pharmacophore hypotheses 94
- pharmacophore modeling 85, 89, 276, 378, 419, 426, 427, 446
- phenix.process_predicted_model 230
- phosphoenolpyruvate (PEP) 427
- physics-based refinement 236, 238
- physiologically-based pharmacokinetic (PBPk) 521
- Platform for Unified Molecular analysis (PUMA) server 380
- Poisson – Boltzmann (PB) 5, 6, 69, 83, 262
- polar neutral molecules 86
- polarizable continuum model (PCM) 185
- polarization energies 185, 186
- polypeptides (proteins) 184
- polypharmacology 214, 215, 382
- positive predictive value (PPV) 515
- potential energy profile 125, 126, 133
- potential of mean force (PMF) 16, 38, 48, 127, 202, 567
- Power User Gateway–Representational State Transfer (PUG-REST) 398
- PPI preference (PPIP) maps 104, 106
- PPI-GaMD simulations 39
- predefined collective variables 21
- predicted local distance difference test (pLDDT) 228
- Predictive MetID (Metabolites identification from chemical structures) 521–522
- primary human hepatocyte (PHH) cytotoxicity 502
- PROTAC conformations 619, 621
- PROTAC-mediated ternary complexes 540, 541, 547–550, 553, 556
- protein data bank (PDB) 241, 258, 399, 421, 444, 476, 587
- protein-drug (un)binding kinetics 45–46
- protein folding 22, 49, 65–68, 240, 245, 625, 636, 654, 655, 668
- protein functional group probability grid maps (PPGMaps) 103
- Protein Kinase Inhibitor Database (PKIDB) 401
- protein–ligand binding free energy 255, 263
- protein–ligand complexes 5, 53, 54, 66, 68, 77, 124, 158, 159, 189, 192, 194, 195, 197, 239, 244, 255–265, 421, 424–426, 428, 454, 476, 477, 540, 547, 595, 596, 623, 624
- protein–ligand (PL) docking and scoring 622
- protein–ligand interactions 66–70, 103, 122, 125, 215, 244, 257, 260, 262, 425, 428, 472, 488, 588, 596, 597, 668
- protein-MG-protein ternary complex 543
- protein-partial MG complex 551, 552, 555
- protein–protein interfaces (PPIs) 541
- proteolysis-targeting chimera database (PROTAC-DB) 401
- PROteolysis-TARgeting Chimeras 401, 539

protonation states 139, 158, 160,
168, 169, 172, 174–177, 259,
446, 477

proton transfer reaction coordinate 130

pseudo-binding constant 503

pseudo-optimization algorithm 374

Psychoactive Drug Screening Program 400

PubChem BioAssay database (PCBA) 477

PubChem bioassays 399, 404

PubChem fingerprint 346, 353

PubChem-Resource Description Framework
(RDF) 398

public ligand libraries

- generated databases (GDBs) 450
- KnowledgeSpace 450–451

PUG-Simple Object Access Protocol (SOAP)
398

pyrido[3,4-b]homotropene (PHT) 427

PyRMD 369

q

QM-driven refinement

- feasibility of 159, 160
- impact on protein
 - selecting protomer states 174, 175
 - structure inspection and modification
172–174
- ligand strain energy 160, 161
- overall crystallographic structure quality
metrics 162
- ZDD of difference density 161, 162

QM/MM MD simulations 135, 140, 571,
572, 574

QM/MM-PBSA scoring 125

qPacker algorithm 664–667

quantitative structure-activity relationship
(QSAR)

- conventional model development
499–509
- deep neural networks methods 280
- development (clinical) stage 499
- discovery phase 498
- machine learning methods 278–280
- mathematical form 278
- technologies and algorithms 516–519

QuantumBio 158, 160, 168, 177, 622, 623

quantum computing

- adiabatic/annealing 647
- classes of applications 643
- cognitive challenges 631–634
- connectivity 651
- cultural challenges 635
- de novo* rational design 662–663

- digital 629, 630
- errors 650, 651
- features 630
- gate-based approaches 667
- gate model 644
- hardware 648–650
- for molecular biology
 - data-driven approaches 655
 - global quantum effects in functional
bio-molecules 654
 - local quantum effects in biochemical
processes 653–654
 - NISQ hardware 652
 - for structural biology 654–655
- peptide and protein design 663
- protein design, optimization problem
663, 664
- qPacker algorithm 665–667
- quantum annealing approach 664, 665
- quantum algorithms 643, 644
- quantum simulators 635, 636
- qubit
 - Bloch sphere 638
 - entanglement 640
 - features 641
 - interference 638, 639
 - multi-particle registers 640
 - nondeterminism 639, 640
 - superposition 638
- scalability 651
- stack 641, 642
- technological challenges 650, 651

quantum Helmholtz machines 660

quantum-inspired machine learning 657

quantum machine learning

- application of 183
- classical ML models 656
- classical information 657, 658
- CQ algorithm types 658
- development and computational
hardware 183
- limitations of 661, 662
- QRAM 658
- quantum clustering model 658
- quantum generative models 660, 661
- quantum information 657, 658
- quantum Kernel model or quantum
support vector machines 659
- quantum neural networks 659, 660

quantum mechanics/molecular mechanics
(QM/MM) approach

- combined QM/MM energy calculations
122, 123

- quantum mechanics/molecular mechanics (QM/MM) approach (*contd.*)
 - for covalent drug design and evaluation 128–133
 - non-covalent inhibitor binding evaluation 124, 125
 - reaction modelling 125–127
- query optimizer 355
- r**
 - radial fingerprinting techniques 347–350
 - random acceleration MD 50, 54
 - random forest model 303, 304, 307, 308, 368
 - reaction-based chemical spaces
 - available building blocks 328, 329
 - reaction sources 327, 328
 - reaction-based enumeration 318, 324, 326–330
 - reaction coordinates (RCs) 16, 21, 47, 50–55, 126, 127, 130, 135, 138, 139, 202, 203, 567, 604
 - reaction modeling, QM/MM 125–127
 - reaction rate prediction
 - absolute unbinding rates 54–56
 - binding rate prediction 56
 - empirical prediction 53, 54
 - error ranges of estimates 52
 - finding reaction coordinates and pathways 52
 - problems with force fields 53
 - reliable benchmarking systems, need for 53
 - real space correlation coefficient (RSCC) 161
 - real-space *R* values (RSR) 480, 481
 - real-space *Z* score of difference density (ZDD) 161
 - Reaxys Medicinal Chemistry (RMC) 397, 402, 403
 - recall or retrieval rate (RTR) 480
 - receptor tyrosine kinases (RTKs) 128
 - reduced graph fingerprint 348
 - reference interaction site model (RISM) 70, 78–79
 - Region-QM refinement approach 162–165
 - relative solvent accessibility (rSASA) 93
 - reliability-density neighborhood approach 301, 302, 304
 - remainder correlation (RC) 186
 - replica exchange molecular dynamics (REMD) 14
 - residual dipolar couplings (RDCs) 232
 - Retro-Score (RScore) 286
 - retrosynthetic combinatorial analysis
 - procedure (RECAP) 322
 - reversible α -ketoamide inhibitor 141
 - reversible covalent inhibition 564, 565, 569–571
 - R-group enumeration 318–321
 - risdiplam 33, 34
 - root mean squared error (RMSE) 481
 - root mean square deviation (RMSD) 476, 480
 - RoseTTAFold 227, 228, 246, 258, 264, 369
 - rule-based methods 521
 - s**
 - SAR-by catalog 343, 344
 - SARS-CoV-2 17, 35, 40, 56, 78, 122, 138–143, 233, 267, 454, 458, 485, 488, 570, 572
 - scaffold enumeration 319, 320
 - scaled/smoothed potential MD 49, 54
 - SciFinder 396, 408
 - scoring functions (SFs) 69, 125, 170, 215, 239, 255, 256, 379, 380, 446, 455, 461, 471–473, 475, 476, 479, 546, 591, 623
 - SE-QM X-ray crystallographic refinement 163
 - secondary structural elements (SSEs) 235
 - semi-empirical DFTB3 method 132
 - semi-empirical QM methods 123–127, 144
 - semiempirical quantum mechanics (SE-QM) methods 158, 159
 - semi-empirical SCC-DFTB method 129, 135
 - Shard Collections 585
 - short-trajectory MD (STMD) 263, 596, 597
 - SHSY5Y cells 428
 - SILCS-Biologics 96, 98, 105, 106
 - SILCS FragMaps 91, 93–96, 98, 100, 101, 103
 - SILCS-Hotspots 91, 93, 95, 96, 98, 102, 103, 105, 106
 - SILCS-MC docking
 - approach 92
 - of Tyr-based activation motifs (ITAMs) 97
 - SILCS-MC method 100
 - SILCS-PPI 96, 98, 103–106
 - SILCS-RNA 91, 95
 - similarity metrics (S) 345, 350–352
 - similarity search 345, 352–353, 357, 369

- Simplified Molecular-Input Line-Entry System (SMILES) standard 318, 623
 - Site Identification by Ligand Competitive Saturation (SILCS) method
 - case studies 97, 98
 - FragMaps construction 99, 100
 - simulations 98, 99
 - and tools 91
 - SiteHopper 590
 - small-molecule calcium sensitizers, for cardiac troponin C 37
 - SMILES 260, 279, 281, 282, 285, 286, 305, 317–319, 348, 349, 375, 379, 400, 446, 448–450, 516, 520, 596, 623
 - solute–water interaction energy 73, 74
 - solvation thermodynamics
 - GIST 72–74
 - 3D-RISM 74, 75
 - watermap 71, 72
 - solvent model density (SMD) method 185
 - Spaya-API 287, 288
 - Spearman's rank-order correlation 481
 - Spearman's rho correlation 482
 - spontaneous binding of risdiplam 34
 - SPRUCE 588–590
 - STAR methods 486
 - StARlite 397, 399
 - STAT3 simulation 621
 - Statistical Assessment of Proteins and Ligands (SAMPL) challenge 482, 483
 - steered MD simulations 50, 262
 - structure-activity relationship (SAR)
 - commercial SAR databases
 - CDDI 403
 - GOSTAR 401, 402
 - Reaxys Medicinal Chemistry (RMC) 402, 403
 - comparison and complementarity of 403–407
 - in drug discovery 407–410
 - open-source databases
 - BindingDB 400
 - ChEMBL database 399
 - DrugBank 399, 400
 - Psychoactive Drug Screening Program 400, 401
 - PubChem 398
 - origins of 396–397
 - structure and interaction based drug design (SIBDD) 183
 - structure-based binding free energy
 - prediction approach 255–268
 - structure-based drug design (SBDD) 35, 83, 121–144, 230, 242, 283, 329, 365, 369, 472, 485, 585
 - structure-based drug discovery (SBDD) 76, 157–177, 239, 257, 258, 590
 - structure-based similarity metrics 351
 - structure-based virtual screening (SBVS)
 - ligand preparation and ligand libraries 444, 446
 - molecular docking 446
 - receptor structure and preparation 444
 - ultra-large ligand libraries 447
 - virtual screenings 446, 447
 - structure–inactivity relationships (SIR) 367
 - substructure-preserving fingerprints 345
 - substructure search
 - atom by atom search 353
 - Ullmann algorithm 353, 354
 - VF2 algorithm 354, 355
 - subsystem analysis (SA)
 - example of 200, 201
 - formulation of 197–199
 - subtractive ONIOM QM/MM calculations 133
 - SuperStar 423, 424, 429, 430, 433
 - superstructure search 345
 - supervised learning 216, 217, 504, 506, 659
 - SureChEMBL 284, 399
 - surface plasmon resonance (SPR) 77
 - SYNOPSIS 322
 - synthetic accessibility score (SAscore) 286, 380
 - synthetic scores 286–288
 - synthon-based virtual screenings 455–457
 - system interaction energy 72, 73
- t**
- Tanimoto distance 303, 304
 - Tanimoto similarity 303, 351–353
 - temperature-accelerated MD 49
 - tensor networks 657
 - therapeutic effects 213, 219
 - thermodynamic cycle 4, 5, 9, 11, 12, 14, 142, 563, 564, 566, 567, 569, 570, 573
 - thermodynamic hypothesis 66
 - thermodynamic integration (TI) 9, 10, 70, 140, 144, 566, 597, 598
 - third power fitting (TPF) 8
 - three-average MM/PBSA (3A-MM/PBSA) 6

- three-dimensional reference interaction-site model (3D-RISM)
 - case studies 78, 79
 - solvation thermodynamics 74, 75
 - 3D shape representations 350
 - time scale problem, of MD simulations 46, 47
 - Toffoli gate 644–646
 - total interaction energy (TIE) 188
 - Tox21 Data Challenge 369
 - traditional drug design and discovery process 275, 276
 - translational entropy 66, 73
 - transition path sampling (TPS) 51
 - transmembrane helix 10 (TM10) 235
 - TROSY-AntiTROSY Encoded RDC (TATER) experiment 233
 - trRosetta 227, 239–241, 243
 - trypsin-benzamidine complex 53, 55–57
 - Tversky similarity 351
 - 2D QSAR models 500
- U**
- ultra-large ligand libraries
 - commercial libraries 447–450
 - public virtual libraries 450–451
 - ultra-large virtual screening (ULVS) approach 443–462
 - umbrella integration 127
 - umbrella sampling (US) 16, 21, 127, 129–132, 134–141, 143, 202, 567
 - uncertainty estimation, for multitask toxicity predictions 303–306
 - UniProt Knowledge Base (UniProtKB) 399, 589
 - unsupervised methods 217
 - USMPep 266
- V**
- vendor libraries 319, 320, 323–326, 338, 449
 - virtual chemical 324
 - VF2 algorithm 353–355
 - VirtualFlow for Ligand Preparation (VFLP) 455, 456
 - virtual ligand screening (VLS) techniques 477
 - virtual screening 373
 - of allosteric modulators of human A₁ adenosine receptor 32–33
 - benchmarking sets for 477–478
 - docking-based ultra-large 451–455
 - machine learning-based 457–460
 - Orion 591–595
 - synthon-based 455–457
 - Virtual Screening of Analog Series (ViSAS) 366, 376
 - V-SYNTHES 457, 458
- W**
- watermap
 - case studies 75–79
 - solvation thermodynamics 71, 72
 - water – water interactions 73
 - WebCSD 420
 - weighted ensembles (WE) 22, 51, 78, 582, 601, 609
 - weighted histogram analysis method (WHAM) 127, 568
 - World of Molecular Bioactivity (WOMBAT) databases 406
- X**
- XModeScore 168–170, 172–177
 - X-ray crystallography 69, 90, 157–177, 213, 229–232, 234, 543, 556, 587, 662
- Z**
- ZDD, of difference density 161, 162
 - 0-order term 188
 - zero-point energy (ZPE) 139, 196
 - ZINC 15 database 372
 - ZINC database 323, 373, 376, 378, 401, 477

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley's ebook EULA.