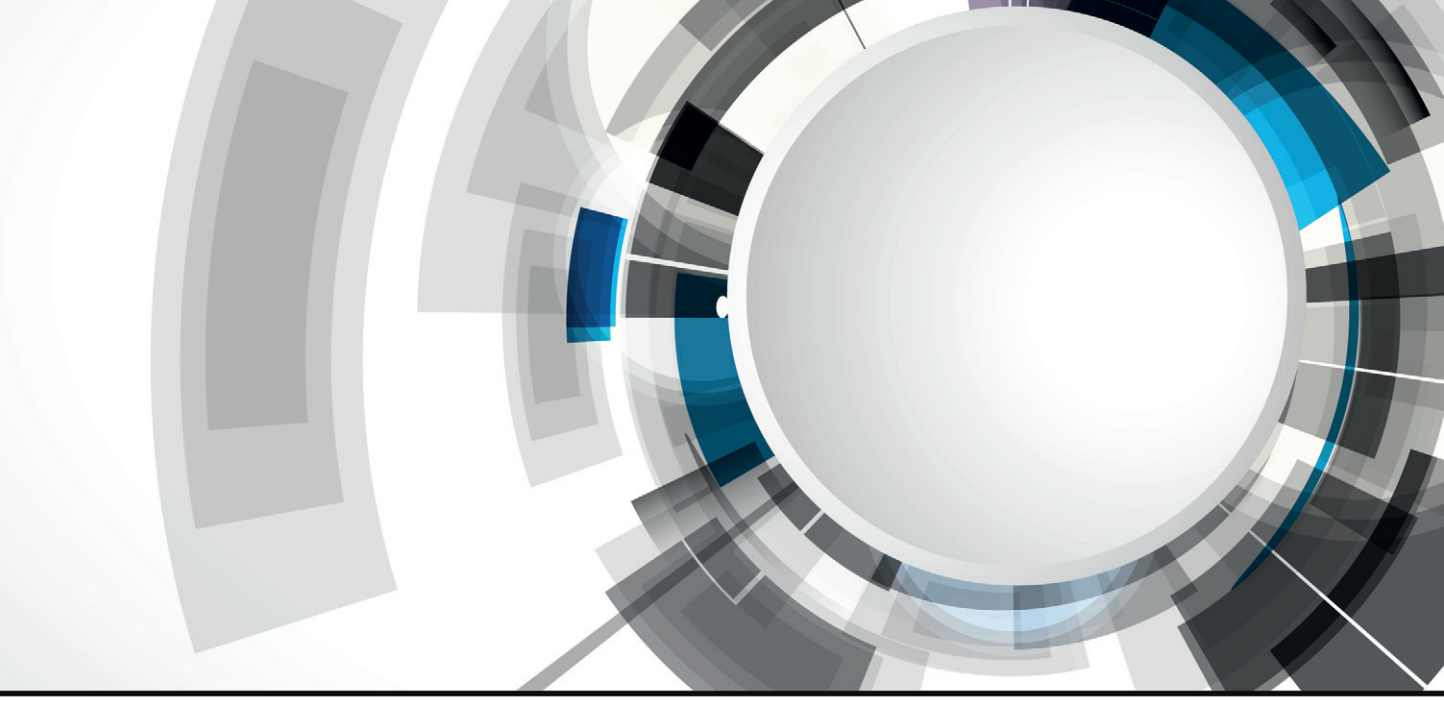


Molecular Docking for Computer-Aided Drug Design

Fundamentals, Techniques, Resources and Applications

EDITED BY
MOHANE S. COUMAR

Molecular Docking for Computer-Aided Drug Design

Fundamentals, Techniques, Resources and Applications

Edited by

MOHANE S. COUMAR
Centre for Bioinformatics
School of Life Sciences
Pondicherry University
Kalapet, Pondicherry, India

ACADEMIC PRESS
An imprint of Elsevier

Academic Press is an imprint of Elsevier
125 London Wall, London EC2Y 5AS, United Kingdom
525 B Street, Suite 1650, San Diego, CA 92101, United States
50 Hampshire Street, 5th Floor, Cambridge, MA 02139, United States
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, United Kingdom

Notices
Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

ISBN: 978-0-12-822312-3

Publisher: Andre Gerhard Wolff
Editorial Project Manager: Tracy I. Tufagaa
Production Project Manager: Kiruthika Govindaraju
Cover Designer: Alan Studholme

Typeset by TNQ Technologies

"கற்றது கை மண் அளவு கல்லாதது உலகளவு"

"Katrathu Kai Mann Alavu, Kallathathu Ulagalavu"

(What you have learnt is a mere handful; what you haven't learnt is the size of the world)
-Thirukkural

This book is dedicated to my parents Selvaraj and Gunasundari, my mentors Dharam Paul Jindal (late) and Hsing-Pang Hsieh, my wife Vasundhara Devi, and my daughters Iniya and Lishitha.

List of Contributors

Azizeh Abdolmaleki, PhD
Department of Chemistry
Tuyserkan Branch
Islamic Azad University
Tuyserkan, Iran

Imlimaong Aier, M.Tech
Department of Bioinformatics & Applied Sciences
Indian Institute of Information Technology – Allahabad
Allahabad, Uttar Pradesh, India

Carolina Horta Andrade, PhD
Laboratory for Molecular Modeling and Drug Design
Universidade Federal de Goiás
Faculdade de Farmácia
Goiânia, GO, Brazil

Tamanna Anwar, PhD
Centre of Bioinformatics Research and Technology
Aligarh, India

Hemant Arya, PhD
Department of Biotechnology
Central University of Rajasthan
Bandar Sindri, Rajasthan, India

Tarun Kumar Bhatt, PhD
Department of Biotechnology
Central University of Rajasthan
Bandar Sindri, Rajasthan, India

Andrzej J. Bojarski, PhD
Maj Institute of Pharmacology
Polish Academy of Sciences
Kraków, Poland

Francesca Cavaliere, PhD student
Molecular Modelling Lab
Department of Food and Drug
University of Parma
Parma, Italy

Sibani Sen Chakraborty, PhD
Department of Microbiology
West Bengal State University
Barasat, West Bengal, India

Jeyaraj Pandian Chitra, PhD
Department of Biotechnology
Dr. Umayal Ramanathan College for Women
Alagappa University
Karaikudi, Tamil Nadu, India

José Correa-Basurto, PhD
Laboratorio de Diseño y Desarrollo de Nuevos Fármacos e Innovación Biotecnológica (Laboratory for the Design and Development of New Drugs and Biotechnological Innovation)
Escuela Superior de Medicina
Instituto Politécnico Nacional
Plan de San Luis y Díaz Mirón
Ciudad de México, Mexico

Mohane Selvaraj Coumar, M.Pharm, PhD
Centre for Bioinformatics
School of Life Sciences
Pondicherry University
Kalapet, Pondicherry, India

Pietro Cozzini, PhD
Molecular Modelling Lab
Department of Food and Drug
University of Parma
Parma, Italy

R. Vasundhara Devi, M.Tech, PhD
Departrment of Computer Science
School of Engineering & Technology
Pondicherry University
Kalapet, Pondicherry, India

Teodora Djikic, PhD
University of Belgrade - Faculty of Pharmacy
Department of Pharmaceutical Chemistry
Belgrade, Serbia

Zarko Gagic, PhD PharmD
University of Banja Luka - Faculty of Medicine
Department of Pharmaceutical Chemistry
Banja Luka, Bosnia and Herzegovina

Jahan B. Ghasemi, PhD
Chemistry Faculty
Drug Design in Silico Laboratory
University of Tehran
Tehran, Iran

Divya Gupta, M.Tech, PhD scholar
Interdisciplinary Biotechnology Unit
Aligarh Muslim University
Aligarh, Uttar Pradesh, India

Department of Life Sciences
Uttarakhand Technical University
Dehradun, Uttarakhand, India

Ravi Guru Raj Rao, PhD student
Structural Biology and Bio-Computing Lab
Department of Bioinformatics
Science Campus
Alagappa University
Karaikudi, Tamil Nadu, India

Sapna Jain, PhD
School of Engineering
University of Petroleum and Energy Studies (UPES)
Dehradun, Uttarakhand, India

Jeyaraman Jeyakanthan, MSc, MPhil, PhD
Structural Biology and Bio-Computing Lab
Department of Bioinformatics
Science Campus
Alagappa University
Karaikudi, Tamil Nadu, India

Asad U. Khan, PhD
Interdisciplinary Biotechnology Unit
Aligarh Muslim University
Aligarh, Uttar Pradesh, India

Sree Karani Kondapuram, PhD scholar
Centre for Bioinformatics
School of Life Sciences
Pondicherry University
Kalapet, Pondicherry, India

Pawan Kumar, PhD
National Institute of Immunology
New Delhi, India

Nathalie Lagarde, PharmD, PhD
Laboratoire GBCM
Conservatoire National des Arts et Métiers
HESAM Université
Paris, France

Florent Langenfeld, PharmD, PhD
Laboratoire GBCM
Conservatoire National des Arts et Métiers
HESAM Université
Paris, France

Neeraj Mahindroo, PhD
School of Health Sciences
University of Petroleum and Energy Studies (UPES)
Dehradun, Uttarakhand, India

Sabrina Silva Mendonca, MSc
Faculdade de Farmácia
Laboratory for Molecular Modeling and Drug Design
Universidade Federal de Goiás
Goiânia, GO, Brazil

Matthieu Montes, PhD
Laboratoire GBCM
Conservatoire National des Arts et Métiers
HESAM Université
Paris, France

José Teofilo Moreira-Filho, PhD
Faculdade de Farmácia
Laboratory for Molecular Modeling and Drug Design
Universidade Federal de Goiás
Goiânia, GO, Brazil

Melina Mottin, PhD
Faculdade de Farmácia
Laboratory for Molecular Modeling and Drug Design
Universidade Federal de Goiás
Goiânia, Brazil

Ayaluru Murali, PhD
Centre for Bioinformatics
Pondicherry University
Kalapet, Pondicherry, India

Mutharasappan Nachiappan, MSc, PhD
Structural Biology and Bio-Computing Lab
Department of Bioinformatics
Science Campus
Alagappa University
Karaikudi, Tamil Nadu, India

Bruno Junior Neves, PhD
Laboratory for Molecular Modeling and Drug Design
Faculdade de Farmácia
Universidade Federal de Goiás
Goiânia, GO, Brazil

Katarina Nikolic, PhD PharmD
University of Belgrade - Faculty of Pharmacy
Department of Pharmaceutical Chemistry
Belgrade, Serbia

Sourav Pal, M.Pharm
Department of Organic and Medicinal Chemistry
CSIR-Indian Institute of Chemical Biology
Kolkata, West Bengal, India

Academy of Scientific and Innovative Research
Ghaziabad, Uttar Pradesh, India

Archana Pan, PhD
Centre for Bioinformatics
School of Life Sciences
Pondicherry University
Kalapet, Pondicherry, India

Jeevan Patra, M.Pharm
School of Health Sciences
University of Petroleum and Energy Studies (UPES)
Dehradun, Uttarakhand, India

Sabina Podlewska, PhD
Department of Technology and Biotechnology of Drugs
Jagiellonian University Medical College
Kraków, Poland

Maj Institute of Pharmacology
Polish Academy of Sciences
Kraków, Poland

Dhamodharan Prabhu, MSc, PhD
Structural Biology and Bio-Computing Lab
Department of Bioinformatics
Science Campus
Alagappa University
Karaikudi, Tamil Nadu, India

G. Pranavathiyani, MSc
Centre for Bioinformatics
School of Life Sciences
Pondicherry University
Kalapet, Pondicherry, India

Sundarraj Rajamanikandan, PhD
Structural Biology and Bio-Computing Lab
Departmentof Bioinformatics
Science Campus
Alagappa University
Karaikudi, Tamil Nadu, India

Muthukumaran Rajagopalan, PhD
Centre for Bioinformatics
Pondicherry University
Kalapet, Pondicherry, India

Amutha Ramaswamy, PhD
Centre for Bioinformatics
Pondicherry University
Pondicherry, India

Manon Réau, PhD
Laboratoire GBCM
Conservatoire National des Arts et Métiers
HESAM Université
Paris, France

Mariadasse Richard, PhD student
Structural Biology and Bio-Computing Lab
Department of Bioinformatics
Science Campus
Alagappa University
Karaikudi, Tamil Nadu, India

Patricia Saenz-Méndez, PhD
Facultad de Química
Computational Chemistry and Biology Group
UdelaR
Montevideo, Uruguay

Faculty of Health, Science and Technology
Department of Engineering and Chemical Sciences
Karlstad University
Karlstad, Sweden

Balasubramanian Sangeetha, PhD
Centre for Bioinformatics
Pondicherry University
Kalapet, Pondicherry, India

Sailu Sarvagalla, PhD
Division of Biology
Indian Institute of Science Education and Research (IISER), Tirupati
Tirupati, Andhra Pradesh, India

Daniela Schuster, Univ.-Prof. Dr.
Institute of Pharmacy
Department of Pharmaceutical and Medicinal Chemistry
Paracelsus Medical University
Salzburg, Austria

Thomas Scior, PhD
Faculty of Chemical Sciences
Laboratory of Computational Molecular Simulations
BUAP
Puebla, Mexico

Asma Sellami, Pharm D
Laboratoire GBCM
Conservatoire National des Arts et Métiers
HESAM Université
Paris, France

Fereshteh Shiri, PhD
Associate Professor
Analytical Chemistry
University of Zabol
Zabol, Iran

Deepanmol Singh, M.Pharm
School of Health Sciences
University of Petroleum and Energy Studies (UPES)
Dehradun, Uttarakhand, India

Yudibeth Sixto-López, PhD
Laboratorio de Diseño y Desarrollo de Nuevos Fármacos e Innovación Biotecnológica (Laboratory for the Design and Development of New Drugs and Biotechnological Innovation)
Escuela Superior de Medicina
Instituto Politécnico Nacional
Plan de San Luis y Díaz Mirón
Ciudad de México, Mexico

Bruna Katiele de Paula Sousa, MSc
Faculdade de Farmácia
Laboratory for Molecular Modeling and Drug Design
Universidade Federal de Goiás
Goiânia, GO, Brazil

Giulia Spaggiari, PhD student
Molecular Modelling Lab
Department of Food and Drug
University of Parma
Parma, Italy

Arindam Talukdar, M.Pharm, PhD
Department of Organic and Medicinal Chemistry
CSIR-Indian Institute of Chemical Biology
Kolkata, West Bengal, India

Academy of Scientific and Innovative Research
Ghaziabad, Uttar Pradesh, India

Veronika Temml, PhD
Institute of Pharmacy
Department of Pharmaceutical and Medicinal Chemistry
Paracelsus Medical University
Salzburg, Austria

Pritish Kumar Varadwaj, PhD
Department of Bioinformatics & Applied Sciences
Indian Institute of Information Technology, Allahabad
Allahabad, Uttar Pradesh, India

Preface

Human health status is of paramount importance in terms of economic productivity, as well as social and mental well-being. However, guaranteeing a good health condition during our life span is surreal. Developments over the past five decades have provided better health care/medicine and increased human life expectancy. Even though there is a huge discord in the average life expectancy of humans residing in different parts of the world, there is an overall trend toward betterment, which is mainly due to better living conditions and the availability of effective and quality medicines. In this respect, one of the most impressive 21st century developments is the decoding of the human genome. This is now coupled with a gigantic leap in our ability to carry out computational work in real time with the help of supercomputers and cloud-based computing. For example, a distributed computing project achieved a speed of 2.43 exaflops (1 exaflop is 10^{18} floating points) during April 2020 for helping to understand COVID-19 related drug targets (data from Folding@home). Such computational power was out of scientist's reach just a few years back.

With a molecular-level understanding of many human diseases, the development of drugs that specifically target and perturb the disease protein is on the rise. However, the process of discovering a drug, even now, relies more on trial and error experimental testing, resulting in long development cycles and expenditure in the range of billions of US dollar. Saving both time and money for discovering drugs could be achieved by incorporating computational approaches in a few of the drug discovery stages. Computer-aided drug design (CADD) is a term applied to a group of techniques associated primarily with the early stages of drug discovery for lead identification and lead optimization; they can speed up the process of identifying molecules for testing in animal models and moving them to clinical trials. Nowadays, CADD techniques are integrated into the iterative process of design, build (synthesize), and experimental testing of the molecules. The most widely used CADD technique is docking, which aims to predict the interaction between two molecules (e.g., a drug and a protein target) in 3D. The predicted interaction between a molecule/drug and a target protein could aid in the identification and development of newer molecules with better interaction to the target in shorter periods. This book brings in experts' from all over the world to discuss their point of view and recent findings in the fundamentals, resources, and application of docking, with a focus on the discovery of new drugs.

I invite students as well as the research community to read and benefit from the book and apply the knowledge to develop better drugs.

January 2021
Pondicherry, India
Mohane S. Coumar

Acknowledgments

I take this opportunity to thank all the authors of the book chapters for their wonderful contribution and support in reviewing the chapters. Also, I thank the Elsevier editorial team members Ms. Mary, Kathy, Tracy, Kiruthika, Kavitha, Sajana, and others for patiently working, guiding, and encouraging me for the past 1 year for this project. Last but not the least, I thank my wife Vasundhara Devi and my daughters Iniya and Lishitha for their understanding and support during my long hours spent with the computer, instead of them.

January 2021
Pondicherry, India
Mohane S. Coumar

Contents

PART I
FOUNDATIONS AND BASIC TECHNIQUES OF DOCKING

1 **Modern Tools and Techniques in Computer-Aided Drug Design**, *1*
Tamanna Anwar, Pawan Kumar and Asad U. Khan

2 **Biomolecular Talks—Part 1: A Theoretical Revisit on Molecular Modeling and Docking Approaches**, *31*
Amutha Ramaswamy, Sangeetha Balasubramanian and Muthukumaran Rajagopalan

3 **Post-processing of Docking Results: Tools and Strategies**, *57*
Sabina Podlewska and Andrzej J. Bojarski

4 **Best Practices for Docking-Based Virtual Screening**, *75*
Bruno Junior Neves, Melina Mottin, José Teofilo Moreira-Filho, Bruna Katiele de Paula Sousa, Sabrina Silva Mendonca and Carolina Horta Andrade

5 **Virtual Libraries for Docking Methods: Guidelines for the Selection and the Preparation**, *99*
Asma Sellami, Manon Réau, Florent Langenfeld, Nathalie Lagarde and Matthieu Montes

PART II
METHODS FOR GENERATING 3D STRUCTURES OF TARGETS

6 **3D Structural Determination of Macromolecules Using X-ray Crystallography Methods**, *119*
Mutharasappan Nachiappan, Ravi Guru Raj Rao, Mariadasse Richard, Dhamodharan Prabhu, Sundarraj Rajamanikandan, Jeyaraj Pandian Chitra and Jeyaraman Jeyakanthan

7 **Electron Microscopy and Single Particle Analysis for Solving Three-Dimensional Structures of Macromolecules**, *141*
Ayaluru Murali

8 **Computational Modeling of Protein Three-Dimensional Structure: Methods and Resources**, *155*
Archana Pan, G. Pranavathiyani and Sibani Sen Chakraborty

PART III
TOOLS, WEB SERVERS, RESOURCES, AND A STEP-BY-STEP GUIDE FOR DOCKING

9 **Resources for Docking-Based Virtual Screening**, *179*
Sailu Sarvagalla, Sree Karani Kondapuram, R. Vasundhara Devi and Mohane Selvaraj Coumar

CHAPTER 1

Modern Tools and Techniques in Computer-Aided Drug Design

TAMANNA ANWAR • PAWAN KUMAR • ASAD U. KHAN

1 OVERVIEW OF COMPUTER-AIDED DRUG DESIGN

The approaches applied in drug development in the present time are very expensive and slow irrespective of the tremendous technological advancements in drug discovery approaches. In such situation of rising pressure of reducing time and cost for safe and effective drug discovery, the focus has moved toward the initial phases of drug discovery and development. Computer-aided drug design (CADD) approaches are now immensely used in the discovery of drug more efficiently and accurately. The cost of discovery and development of drugs can be reduced by 50% with the use of CADD (Xiang et al., 2012).

For more than three decades, CADD approaches have been applied in various stages of drug discovery (Fig. 1.1). Several of the marketed drugs discovered till date have been developed with the help of CADD techniques (Table 1.1). Furthermore, CADD also helps in predicting the novel therapeutic uses of the FDA (Food and Drug Administration) approved drugs; this strategy is termed as "drug repurposing" and will be discussed later in the chapter.

The aim of using CADD approaches is to predict a promising compound that brings a desired effect after binding to the particular biological target. Conventionally, high-throughput screening is used for testing large number of compounds on automated assays to achieve the required effects. In this case, the drug development procedure is not only time-consuming but requires extensive investment. Therefore, to reduce this burden, CADD approaches are applied so that the chemical compounds can be virtually screened first, which will significantly reduce the number of compounds going for experimental screening (Yu & Mackerell, 2017). With the advancement in the information technology (IT), computational power, and availability of big data, recently new approaches have been applied in CADD, which includes machine learning (ML), deep learning (DL), artificial intelligence techniques, and data mining to further enhance the speed and accuracy of drug discovery. In future, drug discovery strategies will very much rely on these advanced IT techniques, which will help in the selection of features (drug and receptor features), image processing, clustering of compounds, etc. For example, to see the drug's impact on patients, ML approaches are used which benefits in the development of drugs that are safe and effective and take less time in the development than the conventional methods. The importance of ML in CADD is well recognized and there are several reports on its successful applications (Khamis & Gomaa, 2015; Vamathevan et al., 2019). In the ML-based approach, large data sets are trained with the help of mathematical framework, which is then applied for the prediction or classification of a new data set (Deo, 2015).

Advancement in the different aspects of computational approaches aid in CADD such as ML approaches help in modeling complex systems that will provide insight into the designing and essential knowledge of molecules. However, DL approaches help in quickly selecting compounds based on pattern recognition, as well as it can be used for early detection of disease and management of the disease. Traditional CADD approaches can be broadly divided into two groups depending upon the availability of the target protein structure: (1) structure-based drug design (SBDD) and (2) ligand-based drug design (LBDD). Availability of the target protein structure provides additional edge in the direct hit to lead optimization process. SBDD includes approaches such as molecular docking, virtual screening (VS), structure-based pharmacophore modeling, and de novo drug design, whereas LBDD approaches include similarity-based screening, quantitative structure–activity relationship (QSAR) modeling, ligand-based pharmacophore modeling, and scaffold hopping (Fig. 1.2).

Molecular Docking for Computer-Aided Drug Design. https://doi.org/10.1016/B978-0-12-822312-3.00011-4

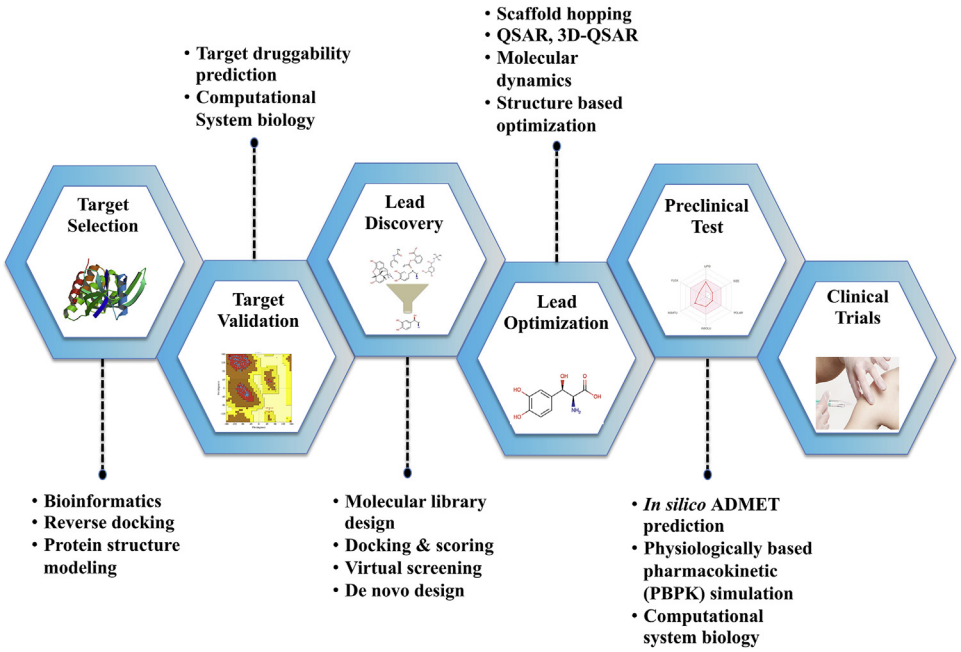


FIG. 1.1 Computer-aided drug design approaches applied in various stages of drug discovery.

TABLE 1.1
List of Drugs Developed with Computer-Aided Drug Design (CADD) Approaches.

Drug	Indication	CADD Approach	Status	References
Saquinavir	Inhibitor of HIV proteases	Structure-based drug design	Approved 1995	Drie (2007)
Nelfinavir	Inhibitor of HIV	Structure-based drug design	Approved 1997	Fischer & Robin Ganellin (2006)
Norfloxacin	Bacterial DNA gyrase Inhibitor	Quantitative structure–activity relationship	Approved 1998	Roy (2015)
Zanamivir	Antiviral (influenza A and B)	Modeling de novo design	Approved 1999	Clark (2006)
Amprenavir	HIV	Protein modeling and molecular dynamics	Approved 1999	Wlodawer & Vondrasek (1998)
Zolmitriptan	Migraine	Pharmacophore modeling	Approved 2003	Clark (2006), Glen et al. (1995)
Dorzolamide	Glaucoma and ocular hypertension	Fragment-based screening	Approved 2012	Grover et al. (2006)

2 CHEMICAL LIBRARIES

Traditionally, for finding a hit against any target in drug discovery, the structure of compounds that can act as inhibitor or activator is required for docking/VS. A high-throughput virtual screening (HTVS) method utilizes chemical databases having millions of compounds to shortlist potential compounds for synthesis. A large number of databases offer structures of chemical compounds, biological targets, and data pertaining to bioactivity for drug discovery. These databases are an

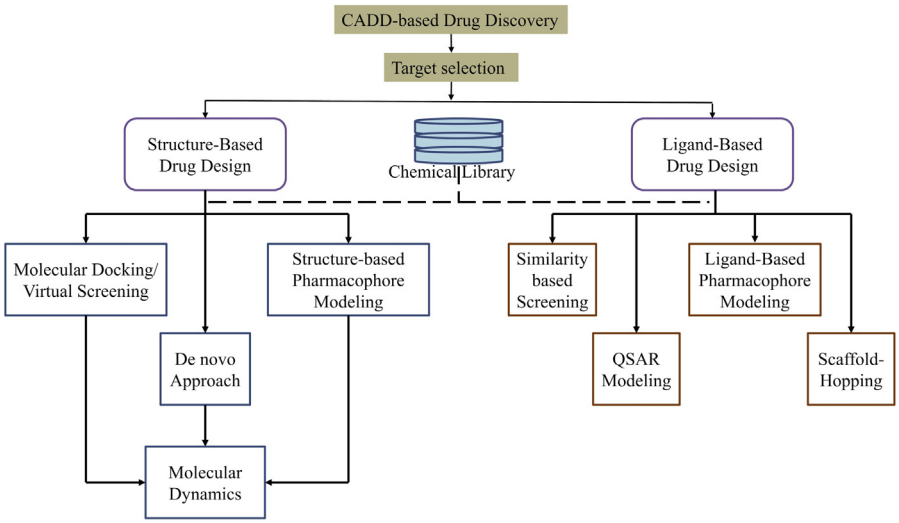


FIG. 1.2 Classification of Computer aided drug design (CADD).

exclusive source for identifying new chemical structures against biological targets. Apart from being a conventional diverse database of chemical structures, considerable attention is given on annotating chemical libraries with a view to provide information on the correlation among the chemical compound and its biological function. Several public and commercial repositories of chemical compounds essential for CADD are highlighted. The information of the drug-like compounds and their physiochemical properties can be retrieved from various databases that are available freely, e.g., PubChem, ZINC, ChEMBL, DrugBank, etc. (Table 1.2). Many resources are also available commercially such as Jubilant BioSys, GVK Bio, and Aureus Pharma. These are large databases of target-centric compounds, focusing mainly on kinases, G protein–coupled receptors, nuclear hormone receptors, or ion channels. The major source of chemical data in these databases comes from patents.

3 STRUCTURE-BASED APPROACHES AND SCREENING

SBDD method utilizes the knowledge of 3D structure of the receptor or target for VS and lead optimization. Thus, for receptors/targets having their crystal structure or modeled structure available, this method can be applied. Types of SBDD methods include molecular docking, structure-based 3D pharmacophore modeling, and de novo drug design methods. It is imperative to check whether the selected target is "druggable," i.e., its biological behavior can be altered by binding small molecule. A target with a very deep, large, and/or highly charged binding pocket is considered unsuitable for SBDD (Fauman et al., 2011). Generally, a structure with high resolution (1.5 Å) and a large ligand binding in its active site is preferred (Rueda, Bottegoni, & Abagyan, 2010).

3.1 Target Structure and Validation

The most extensively used resources of 3D structure determined either by X-ray crystallographic method or nuclear magnetic resonance (NMR) is the Protein Data Bank (PDB) database available at http://www.rcsb.org/pdb. The current version contains 162,529 structures, which is largely determined by X-ray crystallography (88.9%); the fraction of NMR spectroscopy and electron microscopy (EM) determined structures is very low (https://www.rcsb.org/stats/summary) (Berman et al., 2002). In cases where the protein structure is not determined experimentally, again computational approaches can be applied to model the protein structure by homology modeling. The homologous structure is modeled with the help of sequence similarity to the experimentally determined structure of a similar protein. One of the most frequently used software for homology modeling which is freely available is MODELLER (Andrej Šali, 1993). There are several other homology modeling tools/servers available freely for, e.g., Swiss Model, Phyre2, LOMETS, CPHmodels 3.2, I-TASSER, etc.

Among the available solved structures in PDB, X-ray–based crystal structures are still dominating over the other experimental approaches such as NMR and cryo-EM (Cooper et al., 2011). In the drug design

TABLE 1.2
General Resources for Retrieving Chemical Compounds for Docking and Virtual Screening.

Database	Description	License type
ChemSpider http://www.chemspider.com	It is a free database of chemical structures that provides fast text and structure-based searches across 81 million chemical compounds gathered from 278 data sources.	Free
eMolecules Plus https://www.emolecules.com	It contains more than 8 million chemical compounds obtained from the network of global chemical suppliers. The chemicals can be ordered from the website as suppliers are directly connected.	Commercial
ACD (BIOVIA Available Chemicals Directory) https://www.3ds.com/products-services/biovia/products/scientific-informatics/biovia-databases/	It is one of the largest structure-searchable collections of commercially available chemicals in the world, having 10 million unique chemical structures.	Commercial
iResearch Library https://www.chemnavigator.com/cnc/products/iRL.asp	It consists of over 160 million commercially available chemical structures.	Commercial
PubChem https://pubchem.ncbi.nlm.nih.gov/	It is a huge collection of chemical compounds that mostly includes small molecules but macromolecules are also included. PubChem Substance (253 million), PubChem Compound (103 million), and PubChem Bioactives (268 million) are the three components of the dynamically expanding PubChem database.	Free
ZINC https://zinc.docking.org/	It is a large database of 230 million purchasable compounds along with their physicochemical properties. The molecules are available in 3D formats that are ready to dock.	Free
ChEMBL https://www.ebi.ac.uk/chembl/	This database includes bioactive molecules that have properties of drug-like compounds as well as the data of their chemical, bioactivity, and genomic properties are also included. It consists of around 2 million compounds, 13377 drug targets, and 15996368 activities.	Free
BindingDB www.bindingdb.org/bind/index.jsp	It is a publicly available database of binding affinities of small drug-like molecules with their corresponding candidate drug targets. It includes 1,854,767 binding data, for 7493 protein targets and 820,433 small molecules.	Free
PDBeChem https://www.ebi.ac.uk/pdbe-srv/pdbechem/	A database of ligands, small molecules, and monomers referred in Protein Data Bank (PDB) entries. It is consisting of 30899 ligands data.	Free
SuperNatural II http://bioinf-applied.charite.de/supernatural_new/index.php	This database consists of naturally occurring products. It consists of 325,508 natural compounds.	Free
NPACT http://crdd.osdd.net/raghava/npact/	This is a database of 1574 phytochemicals with anticancerous activity.	Free

TABLE 1.2
General Resources for Retrieving Chemical Compounds for Docking and Virtual Screening.—cont'd

Database	Description	License type
DrugBank https://www.drugbank.ca/	The latest version 5.1.5 contains 13548 chemical compounds including 2628 FDA-approved molecules, 1372 approved biologics, 131 nutraceuticals, and over 6363 experimental drugs.	Free
SuperDRUG2 http://cheminfo.charite.de/superdrug2/index.html	This is a database of marketed drugs that consists of 4600 active pharmaceutical ingredients.	Free
GDB-17 http://gdb.unibe.ch	This database consists of 166.4 billion molecules, which are up to 17 atoms of C, N, O, S, and halogens.	Free
KEGG Drug Database https://www.genome.jp/kegg/drug/	It is a compressive database of drugs approved in Japan, the United States, and Europe. It consists of 11,274 drug entries.	Free
SPECS https://www.specs.net/	It contains more than 3,50,000 compounds suitable for synthesis.	Commercial
Maybridge https://www.maybridge.com	It consists of over 53,000 hit-like and lead-like organic compounds.	Free

pipeline, crystallography has gained more importance as this technique is at the heart of SBDD and fragment-based drug design approaches (Cooper et al., 2011). As per the study published by Westbrook et al., 210 new molecular entries (NMEs) are approved by the FDA between 2010 and 2016, and for these NMEs, around 94% of molecular targets are available in the PDB database (Westbrook & Burley, 2019). Very recently, the wwPDB OneDep system has been set up as a single channel for deposition, validation, and biocuration of all incoming structures (Young et al., 2017). OneDep will ensure consistency in the process at the data deposition as well as internal biocuration level.

As the starting structure influences the outcomes in drug designing process, several quality checks are now introduced apart from the structural resolution and R-factor to assess the quality of the experimental structure (Table 1.3). To maintain the data accuracy of the PDB structure, several measures have been taken such as no theoretical structure is now considered from 2006 onwards, structure factor amplitudes/intensities for crystal structures are required with each structural deposition, and each submitted structure should be published in the journal (Kirchmair et al., 2008). At the structure level, the validation matrix is provided to show the accuracy at the structural, geometric, and electron density (ED) level (Fig. 1.3).

ED maps are now provided for all deposited structures and can be used by both experts and novice to assess more about the quality and characteristic of the protein under consideration. Understanding of the user from the ED maps also ruled out the possible biases incorporated by the used modeling procedure, crystallographer expertise, and familiarity. Though ED maps have given the flexibility to the user to analyze the experimental structure carefully, however, the correct representation of the small ligand molecules at the binding site is still a matter of concern. Interpretation of the ligand position binding partly or full, with or without water from the available ED maps, is a laborious task (Smart et al., 2018). Low-resolution structures especially below 3 Å tend to be trickier where water-based interactions play a crucial role between ligand and protein. To emphasize the critical challenges associated with protein–ligand complex crystallography, Smart et al. (2018) have analyzed the PDB ligand and assess the validation report in detail and examined the geometric and ED fit for the same (Smart et al., 2018).

3.2 Molecular Docking and Virtual Screening

One of the most extensively used computational tools in CADD is molecular docking, which is used for determining the complex structure produced by two or more

TABLE 1.3
Tools/Web Server Generally Used for the Protein Structure Validation and Quality Assessment.

Program	Description	Stand-alone/Web server
PQS	Analyze the quaternary protein structures deposited in the Protein Data Bank	Web server
WHAT IF	Tool for protein structure quality checks	Both
Prosa-web	Assess the quality score with respect to known protein structures.	Web server
PROCHECK	Tool to check the stereochemical quality of the protein structure	Stand-alone
PROCHECK—nuclear magnetic resonance (NMR)	Tool to check the stereochemical quality of the NMR protein structure	Stand-alone
MolProbity	Validate the protein structure at different levels	Web server
NQ Flipper	Erroneous Asn and Gln rotamer detection	Web server
PSVS	Protein structure assessment suite	Web server

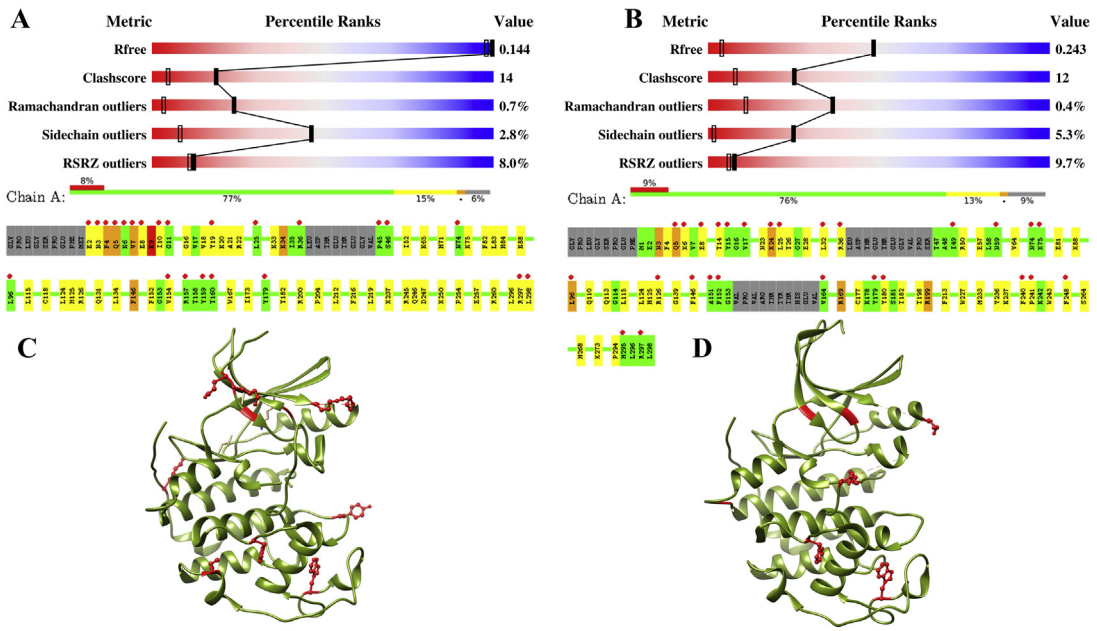


FIG. 1.3 Summary quality metrics available in the wwPDB validation reports. PDB-ID 6GUK (**A** & **C**), and 6Q3C (**B** & **D**) Residues showing the deviation from the experimantal Electron Density Map are shown in red colour (**C** & **D**).

interacting molecules. The docking process involves predicting the 3D conformation of the hit or ligand inside the binding cavity of the target. Several possible ligand poses are generated through molecular docking which are then ranked on the basis of scoring function (SF). The process of simulating the ligand and the

receptor to form a stable complex can be considered as a "lock-and-key model," where the position of key (ligand) is optimized to accommodate into the lock (target binding pocket). The three vital components of molecular docking include the "receptor," the "ligand," and the docking program. The prediction of binding interaction among the protein target and the ligand, the orientation of the ligand in the target's binding pocket, and the scoring of the interaction are achieved by docking programs. The conformational search algorithm explores the poses inside a particular conformational space, while the role of SF is to score each pose that shows its relative binding affinity (Meng et al., 2012). Considerably, the docking program will generate a group of poses for each ligand such that every pose has its own docking score. Generally, the pose that is ranked at the top is considered the best pose of docking; however, the selection of the final pose should not only depend upon the docking score but also on the chemical knowledge and experimental data, if available. The docking program generates the poses by treating the ligand molecule as flexible, and the conformational search algorithm is used for sampling the ligand's torsional degrees of freedom and keeping the target rigid. The accuracy of docking relies on the conformational sampling coverage as well as the SF. Structure-based virtual screening (SBVS) can be done to identify the potential activities available in a large chemical compound database by carrying out docking (Clark, 2008; Schneider, 2010).

3.2.1 Sampling algorithm

The mode of ligand and target binding is possible in several ways as they have six degrees of translational and rotational freedom in addition to the freedom of conformational degrees. Generating all the possible conformations computationally would be highly expensive. Thus, several sampling algorithms were proposed and extensively applied in molecular docking tools (Table 1.4). The ligand is mapped into the active

TABLE 1.4
Docking Programs Used in Computer-Aided Drug Design (CADD) and Their Features.

Docking Program	Characteristic	Sampling Algorithm	Scoring Function	License	References
AutoDock	It is an automated tool for docking consisting of an autogrid, which is used to compute grid, and an autodock, which is used for docking ligands on the grid created by autogrid.	Genetic algorithms, Monte Carlo	Force field based	Open source	Forli et al. (2016)
DOCK	The latest release is built with an improved algorithm to predict binding poses by adding new features like force field scoring enhanced by solvation and receptor flexibility.	Incremental construction, Energy minimization	Force field based	Academic	Ewing et al. (2001)
FRED	An exhaustive search (ES) algorithm is used to identify the ligand's best binding pose in the receptor binding site.	Exhaustive search	Knowledge based	Academic	McGann (2011)
FlexX	It is a tool provided by BioSolveIT for flexible ligand docking. It is fully automated and docking is performed with an incremental construction algorithm.	Incremental construction	Empirical	Commercial	Kramer et al. (1999)

Continued

TABLE 1.4
Docking Programs Used in Computer-Aided Drug Design (CADD) and Their Features.—cont'd

Docking Program	Characteristic	Sampling Algorithm	Scoring Function	License	References
Glide	Glide is a molecular docking suite of software provided by Schrödinger. It offers several modes for virtual screening such as high-throughput virtual screening, standard precision, and extra precision.	Exhaustive search, energy minimization, Monte Carlo	Empirical	Commercial	Friesner et al. (2004)
GOLD	It applies a genetic algorithm for predicting poses of the ligand. It can be configured.	Genetic algorithms	Empirical, knowledge based	Commercial	Verdonk et al. (2003)
ICM	This is an easy-to-use software provided by Molsoft, LLC. The software can be used for chemical clustering, chemical similarity searching, molecular modeling, virtual screening of ligands, fully flexible docking, etc.	Monte Carlo	Empirical	Commercial	Neves et al. (2012)
Surflex-Dock	In Surflex-Dock, the active site ligand is used to produce putative poses, and a combination of similarity searches methods is applied to predict the probable pose of ligand in the binding site.	Incremental construction	Empirical	Commercial	Jain (2007)

site of the target with the help of matching algorithms, on the basis of its shape features and chemical properties. The benefit of matching algorithms is its speed; therefore, active compounds enrichment from vast libraries can be done using this method (Moitessier et al., 2008). This algorithm was used in the older versions of DOCK (Kuntz et al., 1982). Incremental construction algorithm utilizes fragmental and incremental method to place the ligand in the active site. The ligand is fragmented along the rotatable bonds, and then at first the largest fragment is docked inside the binding pocket leading to the addition of rest of the fragments incrementally (Rarey et al., 1996). Other fragment-based algorithms include multiple copy simultaneous search (Eisen et al., 1994) and LUDI (Böhm, 1992a). Programs that implement fragment-based methods comprise DOCK 4.0 (Ewing et al., 2001), FlexX (Rarey et al., 1996), and Surflex (Jain, 2003).

Exhaustive search (ES) is a type of systematic search algorithm, which is used for flexible ligand docking. To perform ES, the ligand's rotatable bonds are systematically rotated at a certain interval, which results in a huge number of ligand conformations. Thus, for initial screening, geometric/chemical constraints are applied after which more accurate refinement procedures are used. FRED (McGann et al., 2003) and Glide (Friesner et al., 2004) are examples of programs that use ES

algorithm. Monte Carlo (MC) and genetic algorithms (GA) belong to the class of stochastic methods. In this class, the conformational space is searched by randomly changing the conformation of the ligand. Both of these algorithms produce a series of random modifications to a ligand or an ensemble of ligands, which is further evaluated on the basis of probability or fitness function. Due to the randomness of conformational sampling, docking is run several times to confirm that the convergence is reached. The programs that apply the MC methods include an earlier version of AutoDock (Goodsell & Olson, 1990), ICM (Abagyan et al., 1994), and Glide (Friesner et al., 2004). GA have been applied in programs such as AutoDock (Morris et al., 1998) and GOLD (Verdonk et al., 2003). Molecular dynamics (MD) (Cornell et al., 1995; Weiner et al., 1984) and energy minimization Mare powerful simulation methods used extensively in MD. These methods are computationally expensive; thus, these methods are applied for refining or rescoring ligand poses produced by other methods. The simulation method is used by the programs DOCK (Kuntz et al., 1982) and Glide (Friesner et al., 2004).

3.2.2 Scoring function

The SF is applied to evaluate the docking poses generated by docking programs to quantitatively measure the quality of the fit (Rajamani & Good, 2007). Along with the evaluation of ligand poses, the SF also evaluates the ligand binding energy and ranks them accordingly to select the best binding ligand. The two main components of any SF are its speed and accuracy. There are three classical categories of SF, i.e., force field (FF)-, empirical-, and knowledge-based SFs. The SF based on FF is calculated on physical atomic interactions like van der Waals (VDW) and electrostatic interactions as well as on bond lengths, bond angles, and dihedrals (Aqvist et al., 2002; Kollman, 1993). The disadvantage with the FF-based SF is its computational speed, which is very slow. Extensions of FF-based SFs include the hydrogen bonds, solvations, and entropy contributions. Further refinement of the result of FF-based docking can be done by applying techniques like linear interaction energy and free energy perturbation (FEP) methods. Empirical SFs are applied to measure the binding free energy (FE) by utilizing various aspects of a protein–ligand complex, for example, hydrogen bond, VDW energy, ionic interaction, hydrophobic effect, binding entropy, etc. (Guedes et al., 2018). Knowledge-based SFs use the experimentally determined structures to get the information of frequencies as well as distance of interatomic contacts in the ligand-protein complex. To improve the accuracy of docking prediction, two or more SFs are applied in some programs, which is referred to as "Consensus Scoring" (Huang et al., 2010). The docking programs applying different SFs are cited in Table 1.4.

Recently, ML-based SFs trained on the complex structures of protein and ligand have gained much attention. This model does not work on predetermined functional forms but is rather developed by supervised learning algorithms (Li et al., 2020). By using the SFs based on ML, the intermolecular binding interactions can be captured implicitly that are difficult to model explicitly. ML-based applications have speedup the inhibitor designing process with desired pharmacodynamics and pharmacokinetic properties compared with the rational *in silico* approaches (Mak & Pichika, 2019). Due to the enormous possibility from the available chemical, genomic, and structural data, its applications are now ranging from the VS-based inhibitor identification, target protein prediction (Kaushik et al., 2020; Zheng et al., 2020), improved consensus docking score development (Ericksen et al., 2017), protein structure prediction (Torrisi et al., 2020), protein–protein interaction prediction (Du et al., 2017), de novo molecule design (Kadurin et al., 2017; Olivecrona et al., 2017), and many more.

ML-based SF used for the prediction of binding affinity performed better than several classical SFs (Ain et al., 2015; Ballester et al., 2014; Khamis et al., 2015). In a very recent study, Su et al. in 2020 have related the performance of six different ML-based SF models to nullify the assumption of overlapping training and test set. The study reports that the performance of the ML models is mostly dependent on the size of the training set used as well as on the content of the training set (Su et al., 2020). However, the docking software does not implement ML-based SFs directly, rather these are generally used for rescoring as these SFs are dependent on training data sets (Zhang, Ai, et al., 2017). The ML-based SFs help in improving the precision of docking done by classical methods by rescoring.

3.2.2.1 Support vector machine. The application of support vector machine (SVM) in SBVS is often done to separate active and inactive ligand poses, and regression model of SVM is applied to predict the binding affinities (Zhang, Ai et al., 2017). A study was done where SVM was combined with the empirical function on the basis of energy terms; as a result, there was an increase in the accuracy of prediction in VS, as well as a correlation among SVM-based and experimental binding affinities was reported (Brylinski, 2013; Kinnings et al., 2011).

Analysis of the HIV protease by ML-based SF SVM-SP performed better than Glide, ChemScore, GoldScore, and X-Score (Li et al., 2011). In another study on 40 DUD2 targets, MIEC-SVM proved to be better than Glide and X-Score (Ding et al., 2013).

3.2.2.2 Random forest. In this classification algorithm, learning is based on multiple decision trees, which is used for classification, regression, etc. The randomness of features is used while building each tree to produce uncorrelated forest with multiple trees, the prediction accuracy of the ensemble of trees is much more than any of the individual trees. Random forests (RFs) have been shown to increase the accuracy of conventional SF by replacing multiple linear regression (Afifi & Al-Sadek, 2018; Wang & Zhang, 2017). In a recent study, RF-based score was developed and compared with five classical SFs. ML-based SF has achieved a very high hit rate at 1% level (55.6%) compared to Vina, which only showed the 16.2% hit rate. Compared to Vina-based predicted activity correlation (Pearson correlation −0.18), RF score has gained Pearson correlation of 0.56 (Wójcikowski et al., 2017).

3.2.2.3 Artificial neural network. Recently, artificial neural network (ANN) has been used extensively in CADD. It is a computational model inspired by biological neural networks. ANN is generally used for QSAR modeling (Cang et al., 2018), but often it is also used to predict binding affinities. An ANN-based SF "NNScore 2.0" predicts binding affinity, as the latest version considers more of binding properties (Durrant & McCammon, 2011). Moreover, NNScore rescoring function can be applied to increase the performance of scoring (Durrant et al., 2013). The prediction accuracy of the classical ANN-based SF can be greatly increased by incorporating techniques such as boosting or bagging (Ashtawy & Mahapatra, 2018). Despite the high precision in the prediction of binding affinity, the ANN-based SFs are incapable of working fine with high dimension data, limiting their application in commercial docking tools.

3.2.2.4 Deep learning. The DL-based SF can extract features from unsupervised data, which is unstructured or unlabeled along with model fitting. The most common model of DL-based SF is convolutional neural network (Ragoza et al., 2017; Wallach et al., 2015), which can be applied for classification of drug binding and prediction of binding affinity (Gomes, Ramsundar, et al., 2017; Stepniewska-Dziubinska et al., 2018). It has been revealed that convolutional neural network models perform better when compared with classical ML models (Bengio et al., 2013), but it is more time-consuming due to the increase in the network complexity of model.

Given a set of training data consisting of an active and inactive compounds, the data can be trained by applying ML-based SFs such as RF-Score (Ballester & Mitchell, 2010), NNScore (Durrant & McCammon, 2011) and SFCscore (Sotriffer et al., 2008; Zilian & Sotriffer, 2013) to find out the known ligands by potency with high accuracy (Wójcikowski et al., 2017). As mentioned earlier, the SF's accuracy can be further improved by applying a hybrid SF that is an integration of different SFs. However, the hybrid SFs are more efficient but more time taking.

3.3 De Novo Drug Design

De novo drug design approach is another most promising SBDD method which allows the generation of the chemical compounds from scratch in the receptor binding site with desirable drug-like properties (Mauser & Guba, 2008; Schneider & Fechner, 2005). Though this approach of novel molecular design is nearly two decades old, its contribution in the drug discovery projects is recently increasing due to its sound applicability and availability of the de novo designing computational program (Schneider & Fechner, 2005). This approach of the drug design process attempts to explore the virtually infinite chemical search space and only captures the building blocks, which is necessary for filling the available interaction space in the substrate binding site (Schneider et al., 2009). So, in the de novo approach, virtual compound generation protocol attempts to imitate the medicinal/synthetic chemist way of designing the virtual compound, while applied SF preform as a virtual assay (Lameijer et al., 2007).

To facilitate the de novo drug designing process, many different tools are published to adapt the multi-objective optimization process (Devi et al., 2015; Nicolaou et al., 2012) and so this approach comes up with many solutions depending upon the initial parameters chosen. Ludi (Böhm, 1992b), LEGEND (Honma et al., 2001), LigBuilder (Wang et al., 2000), BIBuilder (Teodoro & Muegge, 2011), and LiGen (Beccari et al., 2013) are some programs which are developed to assist the de novo drug designing process. As this approach uses all possible combinations to link the available blocks in the respective protein substrate binding site, different sets of rules are formulated to reduce the generated chemical space to a very feasible number of

compounds. Following rules can be implemented to select the de novo chemical hit compound.

(1) Compound should be synthetically accessible
(2) Compound should follow the drug-like/lead-like properties
(3) Generated compounds should be diverse in scaffold

4 LIGAND-BASED APPROACHES AND SCREENING

Contrary to SBDD, LBDD does not require the target 3D structure information, rather the minimum information critical for LBDD method is the knowledge about at least one active compound, which is then utilized for ligand-based virtual screening (LBVS) to pull out similar compounds from databases. This method collects information from the set of reference compounds that are reported in different studies to interact with the target of interest or possess the desired activity. The compounds are represented such that the physiochemical properties relevant to the preferred interaction are retained, while other irrelevant information is excluded. LBDD method for drug discovery is based on "similar property principle" according to which compounds having structural similarity (structure, pharmacophoric features, molecular fields, etc.) will have similar properties. The fundamental approaches for LBDD to identify known actives are either based on chemical similarity or building a model to predict biological activity from chemical structures. LBDD techniques include ligand-based pharmacophore, fingerprint-based similarity methods, and QSAR. The techniques used in LBVS such as substructure mining and fingerprint searches are faster in comparison to SBVS methods like molecular docking. The LBVS technique has helped in finding several promising compounds on the basis of properties such as physiochemical or thermodynamic properties (Forli, 2015). However, the SBVS approach of VS is considered better than LBVS when the target's 3D structure is available (Lyne, 2002). In some cases, where both the target and ligand are known, a hybrid method is used that combines both SBVS and LBVS for achieving better results.

LBVS methods represent compounds with a set of features/descriptors; these descriptors could be either structural or physiochemical and generated with tools based on mechanisms like knowledge-based, molecular mechanics, or quantum mechanics. The molecular descriptors are classified as 1D, 2D, 3D, 4D, etc., according to the chemical structure's dimensionality it is computed from. Several tools are available for computing molecular descriptors which will be discussed later in the chapter. Molecular fingerprint and similarity searches, pharmacophore modeling, and QSAR are the popular approaches of LBDD (Acharya et al., 2010).

4.1 Molecular Fingerprint and Similarity Searches

In this technique, compound libraries are screened based on the molecular fingerprint taken from the known ligands of a particular target to search compounds with similar fingerprint (Vogt & Bajorath, 2011). The theory behind this approach is that the molecules having chemical or physicochemical similarity ought to possess similarity in binding properties (Gomes, Muratov, et al., 2017; Yu & Mackerell, 2017). This approach does not consider the biological activity of the known ligands. Similarity searches are simple but effective and computationally less expensive than pharmacophore modeling and QSAR. In VS, similarity search method is advantageous when only few distinct ligands are known to inhibit a particular target and other methods as pharmacophore screening or structure-based design cannot be applied. The most widely used tool for similarity searching is molecular fingerprint, in which the molecular structure and properties are represented as bit strings. The bit string helps in the identification of presence or absence of molecular features (Xue et al., 2003), which is represented in a quantifiable manner. Every bit in the bit string denotes one molecular substructure/fragment or feature. The bit is fixed to 1 if the fragment is present and 0 if the fragment is absent (Fig. 1.4). The fingerprint-based methods include substructure key–based fingerprints, topological or hashed fingerprints, and circular fingerprints (Cereto-Massagué et al., 2015). The basic difference in these approaches is in the method of translating structural information into the bit string. Each bit represents a certain descriptor or value in substructure key–based fingerprints (Fig. 1.4a) (James et al., 2011). In topological fingerprints, analysis of all the fragments of a molecule is done. Generally, a path is created up to a predefined number of bonds and next all the paths are hashed to build fingerprints. It is likely that the same bit is set by multiple fragments in this method (Fig. 1.4b). The circular fingerprints are also hashed, but here in place of considering paths in the molecule, each atoms environment is documented up to a defined radius. This method is widely applied in VS on the basis of full structure similarity (Fig. 1.4c) (Cereto-Massagué et al., 2015).

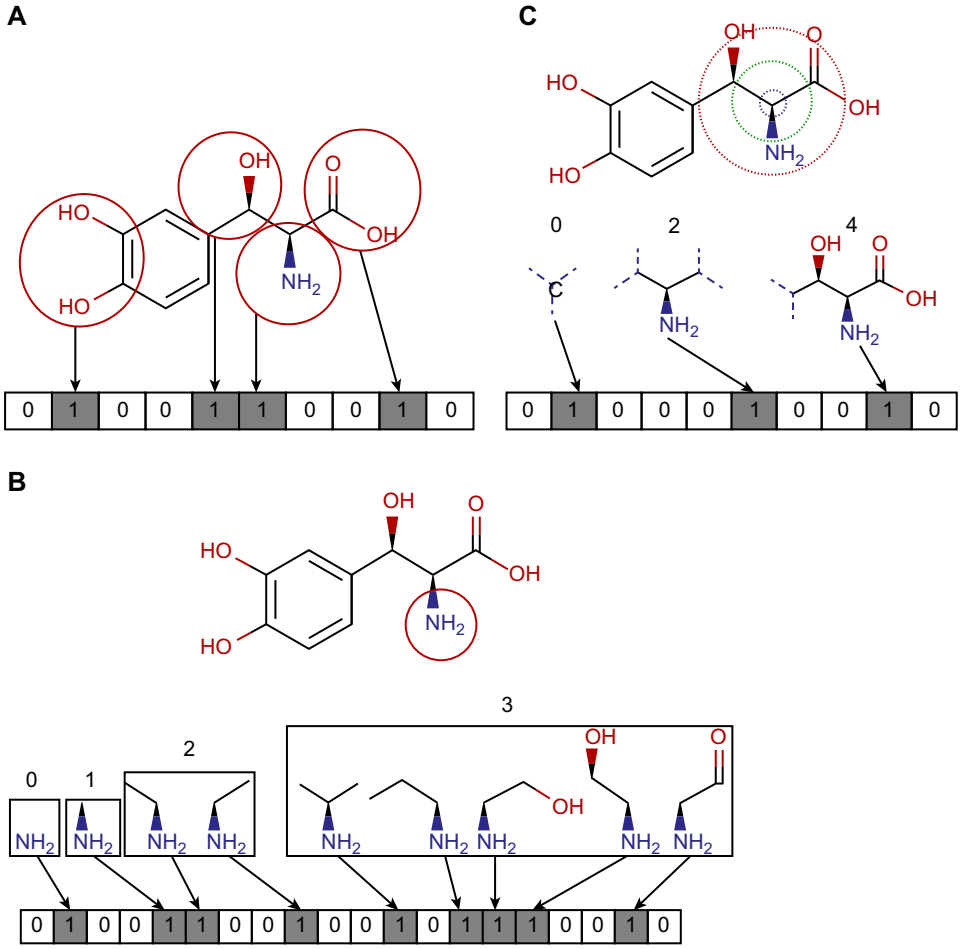


FIG. 1.4 **(A)** An illustration of a substructure key—based fingerprint; molecular substructures represented by bits that are present in the molecule (encircled) are set to 1 and those absent are set to 0. **(B)** Representation of a topological fingerprint. All atoms starting from the amino group of the molecule are shown; the fragment length and subsequent bit in the fingerprint are denoted. Different linear pathway fragments are generated based on the preset number of bonds that are translated into bit strings. **(C)** Representation of a circular fingerprint in which fragment generation starts from a central atom and considers the fragments within a preset radius (e.g., two or four bonds); these fragments are then transformed into bit strings.

Apart from the substructure fingerprint, properties of molecules can also be defined as fingerprint; these property-based fingerprints include functional class fingerprints, pharmacophore fingerprints, reaction fingerprints, etc. The pharmacophore models can also be used as a type of molecular fingerprint. The fragments of the molecule can be transformed into pharmacophoric features; the existence or nonexistence of these features aids in fingerprint creation. However, 3D pharmacophore models are frequently applied to detect chemical functionalities necessary for biological activity as well as for searching large databases of 3D compounds (Cereto-Massagué et al., 2015).

The bit string once created using any of the individual approaches described that the similarity within two molecules is quantified. The molecular similarity can be accessed in different ways; several similarities and distance-based metrics used with fingerprints are mentioned in Table 1.4. Generally, euclidean distance is used for this purpose, but as per the industry standards for molecular fingerprint, Tanimoto coefficient is usually used (Bajusz et al., 2015), which can be evaluated by the formula given in Table 1.5. Tanimoto coefficient lies between the range of 0 and 1; however, sometimes it is also represented in percent. A value $\geq$0.85 of the Tanimoto coefficient represents two compounds that are reasonably similar (Martin et al., 2002).

It has been observed that the longer bit strings perform better in similarity searching as they have a greater amount of stored information (Sastry et al., 2010). Fingerprint similarity search has been

TABLE 1.5
List of Similarity Coefficients and Distances Used for Fingerprint Search.

Similarity/Distance coefficient	Expression	Range
Tanimoto/Jaccard coefficient	$N_c/N_a + N_b - N_c$	0–1
Dice coefficient	$2N_c/N_a + N_b$	0–1
Cosine similarity	$N_c/\sqrt{(N_aN_b)}$	0–1
Euclidean distance	$\sqrt{(Na + Nb - 2N_c)}$	0–N
Hamming distance	$N_a + N_b - 2N_c$	0–N
Russell–RAO coefficient	N_c/m	0–1
Forbes coefficient	N_cm/N_aN_b	0–1
Soergel distance	$N_a + N_b - 2N_c/N_a + N_b - N_c$	0–1

Note: For the fingerprint of two compounds a and b, N_a represents the total number of bits set to 1 in compound a, N_b is the total number of bits set to 1 in compound b, N_c is the number of bits set to 1 in both a and b, and m represents the total number of bits present in the fingerprint.

implemented in various chemical databases for searching similar compounds within a range of defined Tanimoto coefficient, for example, PubChem (Wang, Bryant, et al., 2017), ChEMBL (Bento et al., 2014), ZINC (Irwin & Shoichet, 2005), ChemSpider (Pence & Williams, 2010; Royal Society of Chemistry, 2015), etc. The fingerprint method can be used to study the databases for compound diversity by grouping similar compounds. The software and web servers used for fingerprint-based VS are listed in Table 1.6.

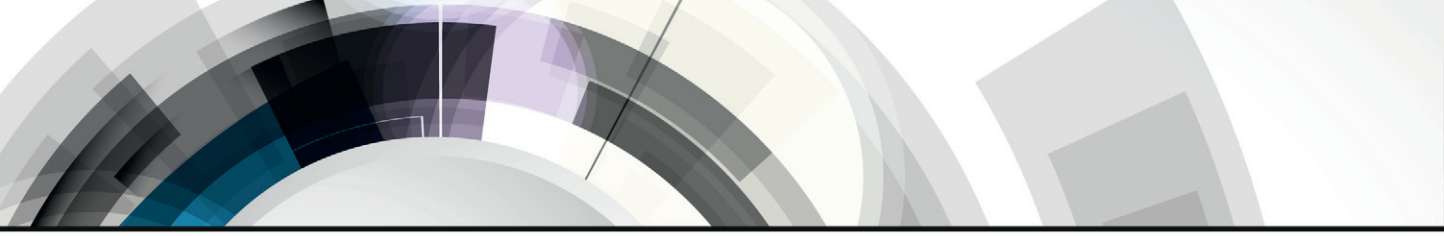
The latest approach in fingerprint-based similarity searching is to use a combination of different VS methods (either fingerprint-based or other VS methods), specifically combining molecular fingerprint similarity method with SBVS (Ahmed et al., 2014; Broccatelli & Brown, 2014; Willett, 2013). As a result of applying a combination of approaches, the compounds performing best will be those that are ranked highest by different methods, leading to an increase in the performance of the VS. Fingerprint-based methods are very extensively used for activity predictions because of their speed, particularly in the area of target fishing, where the query compound is compared with millions of compounds having known activities.

4.2 Pharmacophore Modeling

Most of the biological structures such as proteins or DNA respond to the binding of small chemical molecules, and this response modulates the biological outcomes. How compounds interact with respective protein receptors depends upon the combination of interaction patterns available between protein and ligand molecules. Chemical interactions or chemical features such as hydrophobic, hydrogen bond acceptor, hydrogen bond donor, and ring are the major driving forces in defining the protein–ligand interactions. In the computational drug discovery pipeline, encoding the chemical features in high degree of abstraction is known as 3D pharmacophore features. The term "3D pharmacophore" came into the picture at the starting of the 19th century; however, the concept gradually progressed through many stages, and around the late 80s and early 90s, VS experiments were performed with the help of computational programs (Table 1.7). With time, the pharmacophore concept has evolved from ligand-based approach and receptor-ligand based approach to ab initio receptor-based approach (Kumar et al., 2017; Yang, 2010). With the help of this approach, many successful applications of lead optimization and finding of active molecules have been achieved (Neves et al., 2009; Schuster et al., 2008). Apart from the drug discovery–based application, pharmacophore features are also now in use to design focused chemical library and for scaffold hopping (Shin & Seong, 2013). Apart from the ligand-based pharmacophore modeling, protein–ligand complex–based pharmacophore features are also found to be very valuable in finding the novel inhibitors (Salam et al., 2009; Yang et al., 2009). Apart from the ligand and protein–ligand interaction–based pharmacophore approach, many other pharmacophores perceiving approaches are reported in the literature, and some are detailed below.

4.2.1 Water pharmacophore approach

Water molecules occupied at the unliganded protein binding site are mostly engaged with directional forces or with hydrophobic forces, and over 85% of the protein–ligand complexes have been identified to have one or more bridging water interacting with both protein and ligand (Lu et al., 2007). Most of the time, water-mediated interactions are found to affect the thermodynamic signature of the binding affinity of the ligand (Duan et al., 2017; Spyrakis et al., 2017). Incoming ligand displaces the ordered water molecules from the receptor binding site and consequently disturbs the hydrogen bond network between water and protein. This displacement of the water to the bulk solvent affects the entropy-driven thermodynamic properties of the system (Dunitz, 1994). It thereby



SCIENCE / Life Sciences / Molecular Biology
MEDICAL / Pharmacology

ISBN 978-0-12-822312-3

9 780128 223123

ELSEVIER

ACADEMIC PRESS
An imprint of Elsevier
elsevier.com/books-and-journals