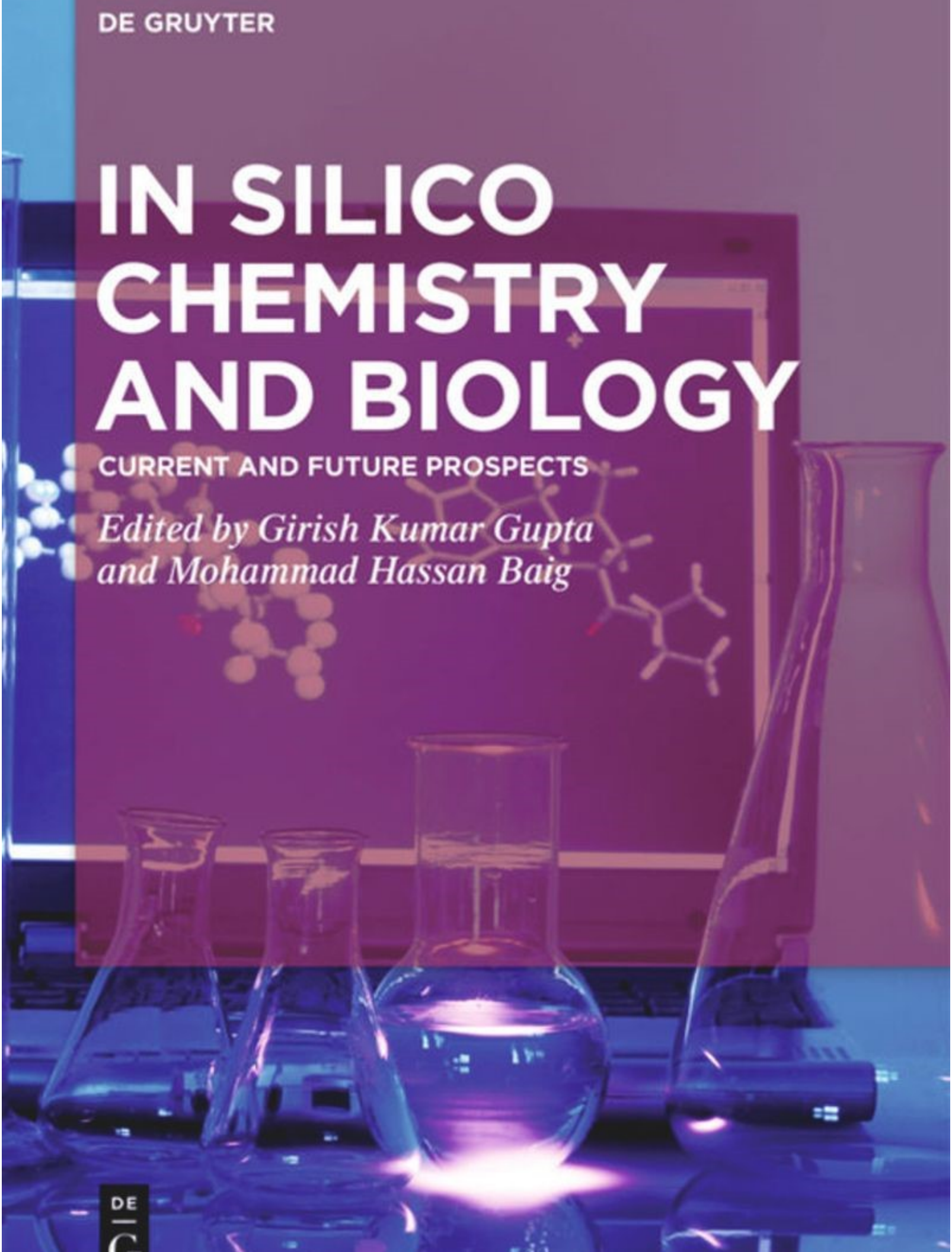
DE GRUYTER



IN SILICO CHEMISTRY AND BIOLOGY

CURRENT AND FUTURE PROSPECTS

Edited by Girish Kumar Gupta and Mohammad Hassan Baig

In Silico Chemistry and Biology

Current and Future Prospects

Edited by
Girish Kumar Gupta and Mohammad Hassan Baig

DE GRUYTER

Editors
Dr. Girish Kumar Gupta
Department of Pharmaceutical Chemistry
Sri Sai College of Pharmacy
Badhani, Pathankot
Punjab
India
and
Research and Development
Sri Sai Group of Institutes
Badhani, Pathankot 145001
Punjab, India
girish_pharmacist92@rediffmail.com

Dr. Mohammad Hassan Baig
Department of Family Medicine,
Gangnam Severance Hospital
Yonsei University College of Medicine
Seoul, Republic of Korea
mohdhassanbaig@gmail.com

ISBN 978-3-11-049517-1
e-ISBN (PDF) 978-3-11-049395-5
e-ISBN (EPUB) 978-3-11-049245-3

Library of Congress Control Number: 2022931619

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.dnb.de.

Cover image: snowflock/iStock/Getty Images Plus
Typesetting: TNQ Technologies Pvt. Ltd.
Printing and binding: CPI books GmbH, Leck

www.degruyter.com

Contents

Mohammad Kalim Ahmad Khan* and Salman Akhtar

Novel drug design and bioinformatics: an introduction

Received March 5, 2021; accepted August 3, 2021

Abstract: In the current era of high-throughput technology, where enormous amounts of biological data are generated day by day via various sequencing projects, thereby the staggering volume of biological targets deciphered. The discovery of new chemical entities and bioisosteres of relatively low molecular weight has been gaining high momentum in the pharmacopoeia, and traditional combinatorial design wherein chemical structure is used as an initial template for enhancing efficacy pharmacokinetic selectivity properties. Once the compound is identified, it undergoes ADMET filtration to ensure whether it has toxic and mutagenic properties or not. If the compound has no toxicity and mutagenicity is either considered a potential lead molecule. Understanding the mechanism of lead molecules with various biological targets is imperative to advance related functions for drug discovery and development. Notwithstanding, a tedious and costly process, taking around 10–15 years and costing around $4 billion, cascaded approached of Bioinformatics and Computational biology viz., structure-based drug design (SBDD) and cognate ligand-based drug design (LBDD) respectively rely on the availability of 3D structure of target biomacromolecules and vice versa has made this process easy and approachable. SBDD encompasses homology modelling, ligand docking, fragment-based drug design and molecular dynamics, while LBDD deals with pharmacophore mapping, QSAR, and similarity search. All the computational methods discussed herein, whether for target identification or novel ligand discovery, continuously evolve and facilitate cost-effective and reliable outcomes in an era of overwhelming data.

Keywords: bioinformatics, homology modelling, LBDD, MD simulation, molecular docking, SBDD

1 Introduction

Nowadays, highly sophisticated tools and techniques are being developed and used to tackle big data generated via various genomics, proteomics, and allied projects

***Corresponding author: Mohammad Kalim Ahmad Khan**, Department of Bioengineering, Faculty of Engineering, Integral University, Lucknow, Uttar Pradesh, 226026, India, E-mail: mkakhan@iul.ac.in. https://orcid.org/0000-0002-8004-1448
Salman Akhtar, Department of Bioengineering, Faculty of Engineering, Integral University, Lucknow, Uttar Pradesh, 226026, India

https://doi.org/10.1515/psr-2018-0158

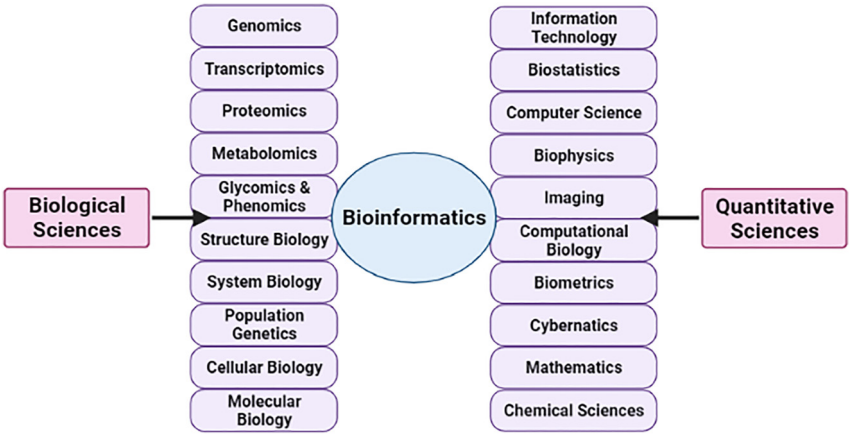


Figure 1: The genesis of bioinformatics.

exploring complicated biological systems, thereby helping to understand the genetic changes affecting health and diseases. However, globally, the scientific community finds it a Herculean task to quickly manage such enormous data and deducing meaningful findings, which is relatively more comfortable and cost-effective. So, according to the need and demand of time, science was developed called highly interdisciplinary bioinformatics, including quantitative sciences such as biostatistics, mathematics, computer science, chemical sciences, biophysics, imaging, computational biology, biometrics and cybernetics, as well as biological sciences such as genomics, transcriptomics, proteomics, metabolomics, glycomics, phenomics, structural biology, system biology, evolutionary biology, population genetics, cellular and molecular biology to uncover patterns and associations within and between all sets of biological data (Figure 1).

Thus, after the augment of bioinformatics, we got two approaches to tackle biological problems. One is a wet-lab experimental approach which is the conventional strategy of biological scientists. Another approach is biomolecular modelling and simulation, often known as computational or *in silico* or dry-lab method. However, wet-lab biology is undoubtedly used to develop better models to describe our understanding of biology, while *in silico* results require validation through wet-lab experimentations. It means that *in silico* biology depends on experimental science to produce raw data for analysis. It, in turn, provides valuable information, clues and meaningful interpretations for further research and development [1, 2]. Thus, this interdisciplinary science is about improving and facilitating the methods and technologies of acquisition, processing, storage, distribution, analysis, interpretation and display of all biological information used by the people to answer the biological questions otherwise unattainable using conventional strategies. However, the term bioinformatics did not mean what it means today. In the early 1970s, Paulien Hogeweg and Ben Hesper coined bioinformatics to explain biotic systems' information processes [2–4].

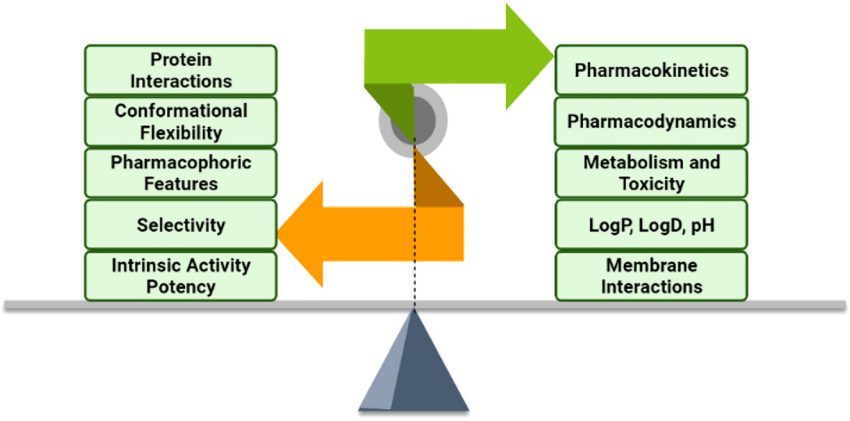


Figure 2: The balance between biological activity and drug-like properties.

Bioinformatics and its intercalated approaches revolutionise and accelerate the entire virtual screening process of lead molecules and their subsequent development into drug molecules. There is a gap between the rate of initial data screening and their conversion into drug-like molecules reaching the market for several reasons, including low biopharmaceutical properties, lack of efficacy, toxicity, and market response.

According to recent data published by Tufts Center for the Study of Drug Development (https://csdd.tufts.edu), the cost of drug discovery and development of new molecules lies somewhere between $2–3 billion that is supposedly very expensive and inadmissible level [5]. However, the overall cost of launching a new drug molecule to market is from the initial drug discovery and design process through various preclinical, clinical trials to registration and regulatory approvals. Moreover, such gaps could be curtailed to a certain extent by accelerating the accuracy and efficiency of lead optimisation techniques by exploiting *in silico* potential and thereby enhancing the balance between activities and drug-likeness properties of lead molecules (Figure 2).

Different tools and techniques in the drug discovery process play an essential role in optimising newly identified small bioactive molecules. Once a molecule is established in the initial phase of the discovery process, we need to streamline biologics' desirable characteristics. The structure-activity relationship (SAR), QSAR, CADD, SBDD, and *de novo* drug design is widely used to optimise lead molecules. There are numerous tools for the characterisation of binding cavities, e.g., estimation of charge distribution, pKa values or lipophilicity calculation, and identification of H-bond donors and acceptors; moreover, various docking tools are used along with 3D structure databases of bioactive molecules with different scoring parameters that attempt to depict the binding propensity of designed molecules. To be considered for further improvement, lead structures should be acquiescent to chemistry optimisation and have desirable drug-like properties.

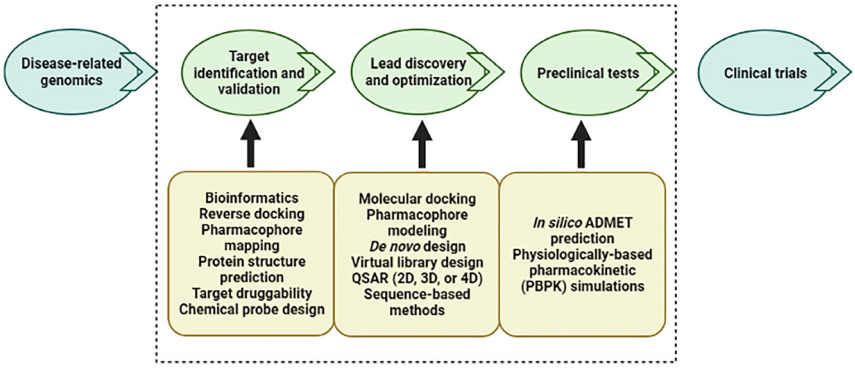


Figure 3: Computational drug discovery approaches are applied in various stages of the drug discovery pipeline.

Although history indicates a different story, many drugs have been discovered by serendipity. However, a deeper understanding of cell biology, genetics, computational tools, target identification methods, and cracking its 3D structure has moved researchers' more rational approach to drug design. The exponential increase in information on biological macromolecule and small molecules and biologics in various databases has increased the application of computational drug discovery, and it is applied to almost every stage in the drug design workflow, which includes target identification and validation, lead discovery, and optimisation and preclinical tests (Figure 3).

CADD uses a more targeted search to improve novel drug compounds' hit rate, which is impossible in traditionally used high throughput screening and combinatorial chemistry. In novel drug design, CADD is mainly used for three primary purposes: (1) filtering more extensive compound libraries into smaller sets. (2) Increasing ADMET properties by guiding lead compounds' optimisation. (3) Designing novel lead compounds by growing starting molecules one functional group at a time or by piecing together fragments into novel chemotypes [6].

Drug or rational drug design is an inventive method of finding new medication based on biological targets' knowledge [7]. The classification of drug design can be studied in two ways: structure-based drug design and ligand-based drug design. In structure-based drug design, the 3D structure and functional role of the target are known. We develop molecules with desirable characteristics towards the target, which can be a protein or nucleic acid. Another approach is ligand-based drug design. It is used when the 3D structure of the target is not known, and we try to develop small molecules with desired properties towards the target [8]. Different techniques used in CADD are summarised in Figure 4.

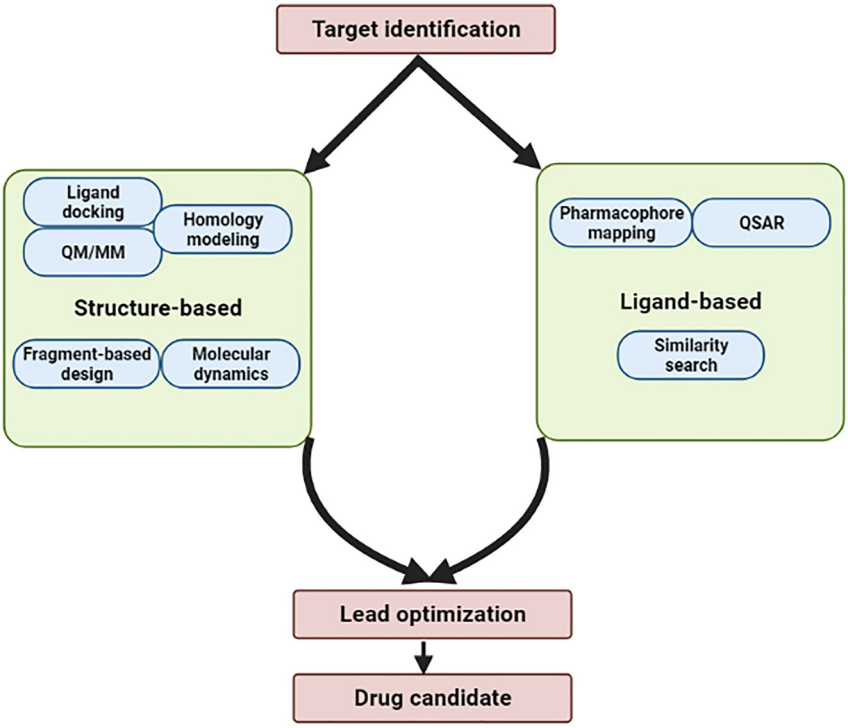


Figure 4: Various techniques used in CADD.

2 Structure-based drug design

With the availability of target structure, we use a structure-based drug design approach. X-ray Crystallography and NMR techniques determine the 3D structures of targets, and the data is stored in a Protein Data Bank (PDB). If the target structure is not known, then we can predict the structure by homology modelling. This approach works on the hypothesis that the molecule's ability to interact with the target and exert the desired effect is due to its binding on a specific binding site on the biological target. Molecules that share the same kind of interaction with the binding site exert the same biological effect. Hence, novel compounds can be found using the interaction with the binding site. Some of the molecules developed using structure-based drug designs are mentioned in Table 1.

2.1 Homology modelling

If the target's 3D structure is unknown, we can use homology modelling to determine the structure using a template solved structure, also known as the comparative

Table 1: Drug molecules developed using Bioinformatics interdisciplinary approaches.

Year	Generic Name	Manufacturer/Supplier	Drug Target	Techniques
1981	Captopril	Bristol Myers Squibb	ACE inhibitor	SBDD
1989	Zanamivir	Glaxo Smith Kline	Neuraminidase	SBDD
1994	Dorzolamide	Merck & Co., Inc.	Carbonic anhydrase	Fragment-based screening
1995	Saquinavir	Hoffman- La Roche	HIV-1 protease	SBDD
1997	Nelfinavir	Hoffman- La Roche	HIV-1 protease	SBDD
1998	Raltitrexed	AstraZeneca	Thymidylate synthase	SBDD
1999	Amprenavir	GlaxoSmithKline	HIV-1 protease	Protein modelling & MD simulation
2000	Isoniazid	Amsal Chem	Inhibin, alpha (InhA)	SBVS & pharmacophore modelling
2006	Dasatinib	Bristol Myers Squibb	Tyrosine kinase	SBDD
2007	Raltegravir	Merck & Co., Inc.	HIV-1 integrase	SBDD
2010	STX-0119	Sigma Aldrich	STAT3	SBVS
2011	Boceprevir	Schering-Plough	Serine protease	SBDD
2013	Pim-1 Kinase Inhibitors	Tocris Bioscience	Pim-1 Kinase	Hierarchical multistage VS
2014	Epalrestat	Ono Pharmaceutical Co., Ltd.	Aldose Reductase	SBVS & MD simulation
2015	Flurbiprofen	Abbott	COX-2	Molecular docking
2010	STX-0119	Sigma Aldrich	STAT3	SBVS
2017	Steglatro	Merck	SGLT2 inhibitor	SBVS
2018	Herzuma	Celltrion, Inc. and Teva Pharmaceutical Industries Ltd.	HER2/neu receptor antagonist	SBVS & pharmacophore modelling
2019	Rozlytrek	Genentech, Inc.	Tyrosine kinase inhibitor	SBVS
2020	Pemazyre	Incyte Corporation	FGFR	SBVS
2021	Lupkynis	Aurinia Pharmaceuticals Inc.	Calcineurin-inhibitor	SBVS

modelling of protein. The prediction of 3D structure can be made by several approaches, depending on the availability of template sequences with significant sequence identity. If there is no template available with significant sequence identity to the target sequence, we use *de novo* methods or *ab initio* methods [9–11]. If the similarity between the query sequence and template is low (<25%), then we can use a fold recognition approach to find a protein with much more similar folding to the target protein [12, 13]. If the similarity between query and template sequence is more than 35%, we can use homology or comparative modelling for 3D structure prediction [14–16]. If the alignment quality and sequence similarity are correct, then this method can be successful. According to a general rule, if the sequence identity between template and the target sequence is above 50%, then the model developed is good enough

for drug discovery, when it is between 25 and 50%, then the model is good enough for mutagenesis experiment, and when it is between 20 and 25%, then it is not good enough [17].

The basic protocol to build a homology model involves identifying the template, sequence alignment, model coordinates generation, optimisation and model validation. The process starts with identifying at least one protein, a known 3D structure, as a template for the target protein. If the target's protein family is unknown, we look for structures in the data bank of amino acid sequences, which can be compared to the target, using algorithms like BLAST [18] and FASTA [19]. We can also use data banks of amino acid sequences like GenBank [20], SwissProt [21] and Protein Identification Resource (PIR) [22]. The templates are then aligned to the target sequence using software like PSI-BLAST [18], FASTA [19], Multalign [23], Spdb viewer [24], the profile software's Bio Shell [25] and Muster [26]. When we align a part of the query sequence to a target sequence, it is called local alignment. In global alignment, we align all the sequences of query and target sequence. Local alignment is used for detecting possible templates, and global alignment is used for model construction. Sequence identities >25% suggest that template and target have similar 3D structures, and hence template is suitable for modelling [27]. A sequence identity of >60% means that the resulting homology model is accurate and similar to experimentally derive structures because folding in a protein is more highly conserved than its amino acid sequence [27].

After model generation, it is optimised and minimised to remove or minimise the unfavourable interaction between non-covalently bonded atoms. After energy minimisation, molecular dynamics simulations are recommended using force fields and taking into account that calculations are restricted to avoid deviation from the original template and loss of similarity to the experimental model, followed by validation of the constructed model. There are many ways for validation, but the primary methods are based on stereochemical analysis in the same way as is done for experimental structures. The stereochemistry of the model can be verified by software like PROCHECK [28], WHAT CHECK [29], PROSA [30] and Molprobity [31, 32]. Phylogenetically similar proteins have a similar sequence, and homologous proteins have similar structures due to their conserved sequence. Sequence alignment and template structure will help generate a structural model of target protein [33]. Some of the popular modelling software is SWISS-MODEL [34] and MODELLER [35]. The MODELLER [15] can construct transmembrane protein more efficiently, and the Swiss model [36] is used for polar proteins.

Homology modelling has been instrumental in drug design. There are many examples of its usage. In one case, the crystal structure of CXCR4 was used as a template to develop a model of chemo-attractant receptor OXE-R [37]. The binding mode of antihypertensive drugs to angiotensin II receptor type 1 was predicted in another example [38]. Later, crystal structure determination led to validating the homology model and analysing the binding mode of active compounds using MD and

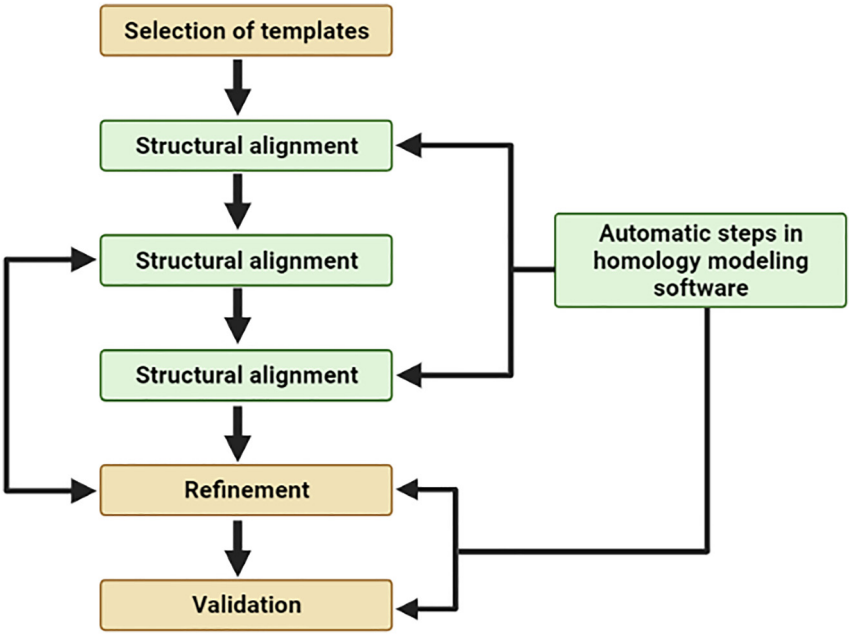


Figure 5: A general protocol for homology modelling.

pharmacophore modelling [39]. The general protocol of homology modelling is shown in Figure 5 [40].

2.2 Ligand docking

Docking is one of the most highly used methods in structure-based drug design. It is the most efficient method in designing, discovering, and synthesis of therapeutic drugs. The molecular docking approach can model the ligand and target at the atomic level to characterise ligand and describe fundamental processes [41]. Two steps can perform docking: the first step is the sampling conformation of ligand in the target's active site, followed by ranking this conformation by scoring function. Monte Carlo [42] and genetic algorithm [43] are typically used.

In the Monte Carlo method, several ligands pose through bond rotation, rigid body translation or rotation is generated. An iterative process of collecting a predefined quantity of conformations, which pass the energy-based selection criterion, is done after saving and modifying the subsequent confirmation in the loop. An earlier version of AutoDock [44], ICM [45] and QXP [46] use the Monte Carlo method.

Another class of well-known stochastic methods is the Genetic Algorithm [43]. Darwin's theory of evolution inspired the genetic algorithm. Genes, which are binary

strings, are encoded forms of the degree of freedom of ligand. Chromosomes, made up of genes, represent the ligand pose. Mutation and cross-over are two genetic operations in GA. The exchange of genes between two chromosomes happens during the cross-over, and sudden random change to the gene is caused by mutation. A new ligand structure is formed when the genetic operators affect the genes. Assessment of new structures is done by scoring function, and the structure which crosses the threshold can be used for the next generation [47]. AutoDock [43], GOLD [48], DIVALI [49], and DARWIN [50] use genetic algorithms.

The scoring function helps to separate correct from contorted poses and to separate binders from inactive molecules. Scoring functions are of two types: force field-based scoring function [51], empirical scoring function [52], and knowledge-based scoring function [53].

Assessing the binding energy by calculations of non-bonded interactions is used in classical force field-based scoring functions [40, 50, 51]. An extension of force field-based scoring functions will also consider hydrogen bonds, solvations and entropy contributions. DOCK [54], GOLD [48], and AutoDock [43] use such functions.

In empirical scoring functions [52], binding energy breaks into several energy components: hydrogen bond, ionic interaction, hydrophobic effect and binding entropy. LUDI [55], PLP [52], ChemScore [56] are examples derived from empirical scoring functions.

Statistical analysis of ligand-protein complexes crystal structures is used to obtain the interatomic contact frequencies and distances between the protein and ligand in knowledge-based scoring functions [53]. Examples of knowledge-based scoring functions are PMF [57], DrugScore [58] and Bleep [59]. There are following three types of docking methodologies that are being used conventionally.

1. Induced fit docking: Ligand and receptor, both are considered flexible in this. The ligand binds flexibly to the active site in receptor protein for maximum bonding forces between them.
2. Lock and key docking: According to this, both receptor and ligand are rigid, and they show tight binding with each other.
3. Ensemble Docking: This approach explains the complexity and flexibility of conformational states of proteins. Multiple protein structures are utilised as an ensemble for docking with the ligand.

Molecular docking is widely used in the drug discovery process to find novel compounds against drug targets. The flow chart of molecular docking is shown in Figure 6.

2.3 Fragment-based drug design

The fragment-based technique is a promising drug design approach to identifying chemical compounds with a low molecular weight that can bind effectively with the

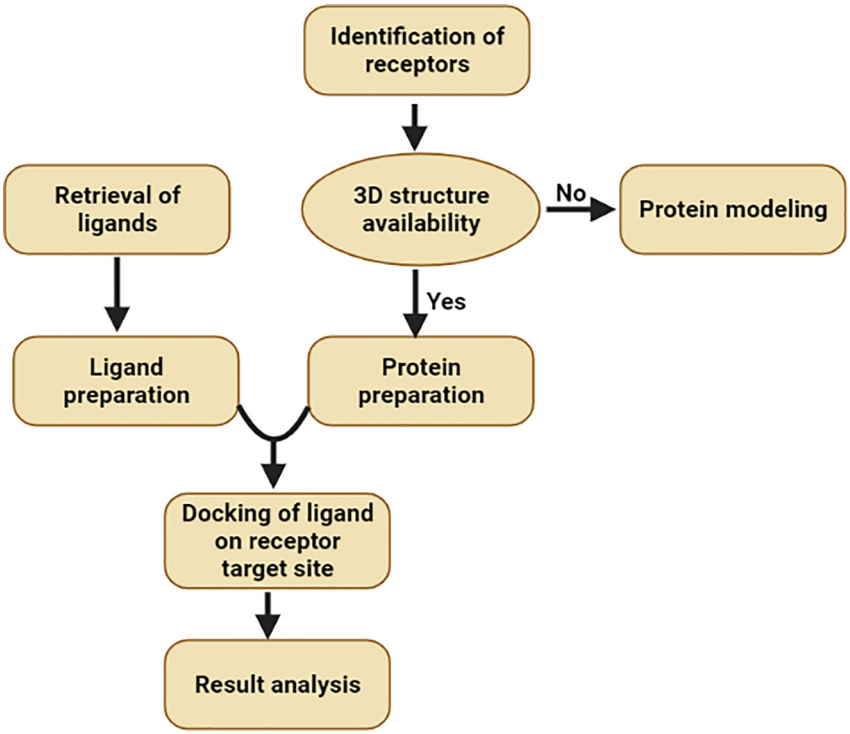


Figure 6: Basic steps of molecular docking.

molecular targets [60]. Most therapeutic targets are proteins, but there are few examples of molecular targets other than proteins, e.g., nucleic acids, nucleoproteins, lipoproteins and glycoproteins [61, 62]. One of the critical principles underlying fragment-based drug design (FBDD) is that screening small chemical compounds increases the likelihood of finding HIT relative to screening large and complex molecules [63].

A variant of virtual screening, emerging *in silico* lead discovery method, is FBDD. A low molecular weight fragment of the complete compound is introduced in the binding pocket of the receptor. A lead candidate is grown, using these fragments as starting material. New leads are formed by sequentially joining together molecules. There are three sources of fragments: natural products, biologically active drugs and compounds with novel scaffolds [64]. Fragments usually have a molecular weight of <250 Da and Log $P < 3$ [65].

Later on, optimisation of fragment HIT is done. There are two commonly used approaches for optimising fragment hits into lead-like compounds: Fragment growing and fragment linking. There are several strategies for developing fragments into lead compounds, and the techniques and strategies are continually evolving. To grow the fragment, we start with a fragment in the ligand-binding site and expand to interact

with the pocket side chain. Fragment linking/merging is the second strategy. First, the fragments have optimally interacted with the pocket, and then fragments are covalently joined, which will most likely form a novel scaffold. A fragment usually has a low binding affinity, grown into a high-affinity ligand through the drug design [66].

Zelboraf was developed in 2011 using FBDD and approved by the FDA to treat late-stage melanoma. Venetoclax is another example of the FBDD-derived drug that was approved by the FDA later. Overall, it is possible to design drugs that conventional SAR-based methods could not predict.

2.4 Molecular dynamics

Proteins are dynamic molecules. A structural snapshot is insufficient to study the protein interaction with a ligand or identify the binding site. Classical molecular dynamics is a physical method to study the forces and motion of atoms and molecules following Newtonian physics. A force field is used to calculate the system's energy and estimate forces between interacting atoms. Then, successive configurations of the evolving system are generated during MD simulation by integrating Newton's laws of motion, providing trajectories that can specify the position and velocities of particles over time. Using these MD trajectories, properties such as kinetic measures, free energy, and other macroscopic properties can be calculated. This method was initially conceived within theoretical physics in the late 1950s, but it now finds application in materials science, chemical physics, modelling biomolecules and drug discovery [67, 68].

Upon assigning Amber [69] or CHARMm [70], Newtonian mechanics-based force fields, molecular dynamics simulations can calculate a trajectory of conformations as a function of time.

MD simulations can model chemical bonds and atomic angles using simple virtual springs, and dihedral angles can be modelled using the sinusoidal function. Non-bonded forces such as hydrophobic interaction and electrostatic interactions can be calculated using Coulomb's law. When used with experimental data, these simulations can significantly affect the process of drug discovery.

The main advantage of MD simulation is in treating structural flexibility and entropic effects explicitly. MD simulation will help determine kinetics and thermodynamics associated with binding and recognition of drug-target because of better hardware and algorithm. In some cases, MD simulation can be done before docking to generate a conformer different from the crystal structure.

Furthermore, MD runs can be done after docking for *in silico* assessment of the predicted binding modes of the top-ranking compounds as a final filter. It can guide chemical synthesis for hit optimisation. Despite being computationally expensive for docking large compounds, MD simulations are being used for hit discovery and optimisation, and their use is increasing steadily. MD is a tool for mapping molecular

fragments to binding sites on targets. It can validate the predicted binding mode. MD is also used for docking natural ligands and known drugs [71].

3 Ligand-based drug design

A ligand-based drug design approach is used without target structure information and availability of one or more bioactive compounds [72, 73]. The information about the bioactive molecule includes chemical structure, biochemical properties and physico-chemical properties. Sometimes, these methods are considered more successful than the structure-based approach [74].

Databases can be screened to find molecules with similar fingerprints as known ligands [75]. Pharmacophore modelling can find common structural features of ligands, which can then screen for molecules with these features [57]. QSAR can build models to predict the activity of a novel molecule [59]. The pharmacophore model can only indicate the activity conferring features of an active ligand, but the relationship between biological activity and the chemical or physical property of the ligand can be studied using the QSAR model. There are several approaches in ligand-based drug design: similarity search, pharmacophore modelling, and QSAR (Table 2).

Table 2: Drug molecules optimised by QSAR and Pharmacophore modelling.

Year	Generic Name	Brand Name	Manufacturer	Drug Target	Technique
1986	Norfloxacin	Noroxin	Merck & Co., Inc.	Fluoroquinolone antibacterial	Lead optimisation by QSAR
1996	Donepezil	Aricept	Pfizer Inc.	AChE inhibitor	Lead optimisation by QSAR
2007	Aliskiren	Tekturna	Novartis	Renin inhibitor	Pharmacophore modelling
2007	Raltegravir	Isentress	Merck & Co., Inc.	HIV-I integrase	Pharmacophore modelling
2010	liraglutide	Victoza	Novo Nordisk A/S	GLP-1 receptor agonist	Lead optimisation by QSAR
2015	Eluxadoline	Viberzi	Allergan, Inc.	Mu-opioid receptor agonist	Pharmacophore modelling
2017	Enasidenib	Idhifa	Celgene Corporation	IDH2 inhibitor	Pharmacophore modelling
2019	Afamelanotide	Scenesse	Clinuvel Pharmaceuticals Ltd.	MC1R agonist	QSAR
2021	Tepotinib	Tepmetko	EMD Serono, Inc.	MET inhibitor	Pharmacophore modelling

3.1 Similarity search

Similarity search is performed when a single bioactive compound is available. The basic principle of similarity searching is to use the backbone of the lead molecule to screen the database for similar compounds. The fingerprints method is used to represent a molecule so that it can be effectively compared against molecules. The fingerprints method depends on chemical information of compounds, which gives a highly qualitative approach for searching for more potent ligands.

Fast *in silico* selection of libraries focussed on target from repositories is a desirable and cost-effective approach. A quick 2D similarity search can be done on compound databases on the availability of the structure of active compounds. The concept behind the similarity search is that the homologous sequences are likely to have similar properties [76, 7776]. 2D similarity search is a method of choice whenever several active reference compounds and databases are available [78].

3.2 Pharmacophore mapping

A pharmacophore is an essential geometric arrangement of a functional group or atom for generating a given biological response. According to IUPAC, A pharmacophore is the ensemble of steric and electronic features necessary to ensure plausible molecular interactions with a specific biological target structure and trigger or block its biological response [79]. A pharmacophore search will find molecules with different overall chemistries, but they have the functional group incorrect geometry. Searching 3D databases for molecules that contain the pharmacophore is the most common use of this technique.

Pharmacophore mapping is one of the significant drug design elements in cases where target 3D structure is unavailable. Virtual screening is the most common application of pharmacophore. However, the pharmacophore concept is also helpful for ADME-tox modelling, side effect, and off-target prediction, as well as target identification. Furthermore, pharmacophore can be combined with molecular docking simulations to improve virtual screening [80]. It can now be used for lead optimisation. It can also align molecules based on the 3D arrangement of chemical features or develop predictive 3D QSAR models. Raltegravir, an HIV-1 integrase inhibitor, has been developed using pharmacophore modelling. Aliskiren, a renin inhibitor for the treatment of hypertension, has been developed using pharmacophore modelling.

3.3 Quantitative structure-activity relationship

It is a method that relates chemical structures to biological or chemical activity using mathematical models [81]. The general QSAR workflow involves gathering a set of

active and inactive molecules against the target and producing descriptors that describe their physicochemical and structural properties.

The mathematical model is then used to correlate descriptors and experimental activity, which will serve as a predictive tool for new entities. A QSAR model allows us to determine the effect of a particular property on ligand activity, but a pharmacophore model involves an active ligand's main features. For example, the QSAR model can tell if a property positively or negatively affects ligand activity, but the pharmacophore model cannot provide such information.

QSAR deals with the 2D and 3D descriptors. Electronegativity, atom distribution, molecular weight, volume, rotatable bonds, interatomic distances, atom types, molecular walk counts, aromaticity and solvation properties. Descriptors either may be structural or physicochemical that can be explained at different complexity levels. Norfloxacin, a fluoroquinolone antibacterial, and Donepezil, an AChE inhibitor, have been optimised by using QSAR.

4 Quantum mechanics/molecular mechanics

QM/MM is a molecular modelling technique that aims to access quantum chemistry accuracy while the other part of the system is treated at a lower level of theory. QM/MM improves the description of polarizability, charge transfer effect, which corrects the electrostatics' overestimation in enzyme-inhibitor free energy.

QM/MM supports the description of metalloenzymes which is challenging to make with solely MM. It can explore chemical bond formations through QM/MM, which will provide helpful information towards designing small-molecule inhibitors. Only MM computation can be more accurate than QM/MM. Hence, a decision to use the primary QM method or MM force field should be made following the underlying chemistry of the biological system [82].

5 Proteochemometrics modelling

Proteochemometrics (PCM) is a hybrid approach using the information of both ligand and target molecules in order to produce desirable results [83, 84]. The biological effect is conceived through the amalgamation of ligand and target attributes simultaneously. Upon merging target and ligand data in the single frame of the machine learning predictive model, someone will be able to start determining the most probable treatment for a given genotype, viz., personalised medicine. Alone neither ligand nor protein can achieve such valuable robustness [83, 84].

6 Deep learning approach

Intelligence exhibited by computers is called machine intelligence or artificial intelligence. Traditionally, experiments have been costly and time-consuming, and various machine intelligence has been used to guide them in drug design history. Over the past decade, QSAR has been used to quickly and cost-effectively identify active molecules from several million compounds.

Identifying new therapeutic molecules from huge data libraries of chemical molecules requires sophisticated tools to hone the drug designing process. Towards this direction, the machine learning tools are combined with deep learning techniques. It has efficiently handled enormous data generated from drug discovery approaches [85]. It includes nonlinear and direct preparing units, which change depictions from lower to higher levels [86].

The process of drug design can be attributed to (i) identification of new chemical molecules, (ii) designing of protein targets, (iii) analysis and elucidation of genetic factors and (iv) biomolecular modelling of pharmacokinetics and pharmacodynamics.

1. Identifying new chemical molecules: The simple way to sharpen drug design is to build computational models to find new small drug compounds from large chemical libraries, thereby facilitating drug discovery. The deep learning methods can accelerate identifying new lead and drug molecules via computational VS [87–90]. The technique of deep learning is also helpful to create different molecular fingerprints [91] or focused molecule libraries [92] and to model pharmacokinetic properties of potential drug molecules [93].
2. Designing of protein targets: Deep learning approaches are being used to explore and discover the structure and function of the protein. The biological function of target molecules can be predicted directly from their raw 3D electron density and electrostatic potential field [94].
3. Analysis and elucidation of genetic factors: Massive amount of genomics data have been developed due to next-generation sequencing (NGS) technology, and they fit well with deep learning approaches. Hence deep learning methods have been used in the development of precision medicines [95], genomics modelling for drug repurposing [96], and sequence specification prediction [97].
4. Pharmacokinetics and pharmacodynamics modelling. Deep learning methods have been used to interact with different complexes, such as homogeneous complexes [47] and drug-protein [98].

7 Summary and outlook

It is of paramount importance for academicians, scientists, pharmaceutical and biopharmaceutical industries to reduce the cost depleted in drug discovery R&D that

might be curbed by developing ultra-high-throughput and sophisticated interdisciplinary tools and techniques designing such new chemical entity (NCE) whose failure rate is low. The data exhibits that 9 out of 10 new NCEs remain unsuccessful in clinical experiments, and the failure rate in Phase I, II and III is about 50 %, 30 %, and 25–50 %, respectively. Moreover, these data may be higher in cases of neurological and some other complicated disorders. Thus, in general, a drug molecule's overall success rate is about 3–8 % that is extremely low. As per the latest statistics of DrugBank online release dated 3rd January 2021, there are 14,556 drug entries, including 2698 approved small molecule drugs, 1473 approved biologics (proteins, peptides, vaccines and allergenics), 131 nutraceuticals, and over 6653 experimental (discovery-phase) drugs (https://go.drugbank.com). However, the permutation and combination of possible compounds are 1060. It means that many compounds are available for screening and subsequent development into the lead and drug molecules. Thus, integrated approaches of Bioinformatics are not only intended to analyse a plethora of biological data cognate to target identification, validation and optimisation of new drug candidates, but other fields viz., structure prediction, gene prediction and phylogenetic analysis are also exploiting its potential to uncover patterns and association within and between them.

Moreover, bioinformatics and its allied sciences play a vital role in understanding complex biological systems. The living systems are highly interconnected and well-organised and regulated by specific phenomenal mechanisms. They can be seen beautifully even in a tiny single cell with many molecules that harmonise to keep it healthy via different molecular interactions. Such molecular interactions generate enormous data to be analysed and interpreted to get meaningful information, thereby fostering data-powered hunting of biological treasure, which is otherwise difficult to achieve through traditional means.

Acknowledgement: The authors would like to thank Ms. Shalini Maurya (DBT-JRF) for her assistance during manuscript writing.
Author contributions: All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.
Research funding: None declared.
Conflict of interest statement: The authors would like to declare no conflict of interest.

References

1. Lotka AJ. Elements of physical biology. Baltimore: Williams and Wilkins Company; 1925.
2. Hogeweg P. The roots of bioinformatics in theoretical biology. PLoS Comput Biol 2011;7:e1002021.
3. Hogeweg P, Hesper B. The alignment of sets of sequences and the construction of phyletic trees: an integrated method. J Mol Evol 1984;20:175–86.
4. Hogeweg P. Simulating the growth of cellular forms. Simulation 1978;31:90–6.

5. Prasad V, Mailankody S. Assessing pharmaceutical research and development costs—reply. JAMA Intern Med 2018;178:588.
6. Veselovsky A, Ivanov A. Strategy of computer-aided drug design. Curr Drug Targets - Infect Disord 2003;3:33–40.
7. Arlington S. Pharma 2005-an industrial revolution in R&D. Eugene: Pharmaceutical Executive; 2000.
8. Mandal S, Moudgil M, Mandal SK. Rational drug design. European Journal of Pharmacology 2009; 625:90–100.
9. Lee J, Scheraga HA, Rackovsky S. New optimisation method for conformational energy calculations on polypeptides: conformational space annealing. J Comput Chem 1997;18:1222–32.
10. Liwo A, Pincus MR, Wawak RJ, Rackovsky S, Ołdziej S, Scheraga HA. A united-residue force field for off-lattice protein-structure simulations. II. Parameterisation of short-range interactions and determination of weights of energy terms by Z-score optimisation. J Comput Chem 1997;18: 874–87.
11. Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Protein: Struct Funct Genet 1999;34:82–95.
12. Sippl MJ, Weitckus S. Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a database of known protein conformations. Protein: Struct Funct Genet 1992;13:258–71.
13. Bryant SH, Lawrence CE. An empirical energy function for threading protein sequence through the folding motif. Proteins: Struct Funct Genet 1993;16:92–112.
14. Johnson MS, Srinivasan N, Sowdhamini R, Blundell TL. Knowledge-based protein modeling. Crit Rev Biochem Mol Biol 1994;29:1–68.
15. Webster D, Sánchez R, Šali A. Comparative protein structure modeling: introduction and practical examples with modeller. In: Protein structure prediction. New Jersey: Humana Press; 2003:97–129 pp.
16. D'Alfonso G, Tramontano A, Lahm A. Structural conservation in single-domain proteins: implications for homology modeling. J Struct Biol 2001;134:246–56.
17. Wong WC, Maurer-Stroh S, Eisenhaber F. Not all transmembrane helices are born equal: towards the extension of the sequence homology concept to membrane proteins. Biol Direct 2011;6:57.
18. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–10.
19. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 1988;85:2444–8.
20. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. Nucleic Acids Res 2003; 31:23–7.
21. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 2000;28:45–8.
22. Barker WC. Protein Information Resource: a community resource for expert annotation of protein data. Nucleic Acids Res 2001;29:29–32.
23. Barton GJ, Sternberg MJE. Flexible protein sequence patterns. A sensitive method to detect weak structural similarities. J Mol Biol 1990;212:389–402.
24. Guex N, Peitsch MC. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 1997;18:2714–23.
25. Gront D, Blaszczyk M, Wojciechowski P, Kolinski A. BioShell Threader: protein homology detection based on sequence profiles and secondary structure profiles. Nucleic Acids Res 2012;40:W257–62.
26. Wu S, Zhang Y. MUSTER: improving protein sequence profile-profile alignments by using multiple sources of structure information. Proteins: Struct Funct Bioinformat 2008;72:547–56.

27. Rost B, Sander C. Bridging the protein sequence-structure gap by structure predictions. Annu Rev Biophys Biomol Struct 1996;25:113–36.
28. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Crystallogr 1993;26:283–91.
29. Hooft R, Vriend G, Sander C, Abola EE. Errors in protein structures [3]. Nature 1996;381:272.
30. Sippl MJ. Recognition of errors in three-dimensional structures of proteins. Proteins: Struct Funct Genet 1993;17:355–62.
31. Davis IW, Leaver-Fay A, Chen VB, Block JN, Kapral GJ, Wang X, et al. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. Nucleic Acids Res 2007;35(Web Server): W375–83.
32. Chen VB, Arendall WB, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, et al. MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr Sect D Biol Crystallogr 2010;66:12–21.
33. Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A. Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct 2000;29:291–325.
34. Schwede T. SWISS-MODEL: an automated protein homology-modeling server. Nucleic Acids Res 2003;31:3381–5.
35. Webb B, Sali A. Comparative protein structure modeling using MODELLER. Curr Protoc Bioinf 2016; 54:5.6.1–37.
36. Roy A, Yang J, Zhang Y. COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. Nucleic Acids Res 2012;40:W471–7.
37. Wu B, Chien E, Mol C, Fenalti G, Liu W, Katritch V, et al. Structures of the CXCR4 chemokine. Science 2010;330:1066–71.
38. Matsoukas M, Cordomí A, Ríos S, Pardo L, Tselios T. Ligand binding determinants for angiotensin II type 1 receptor from computer simulations. J Chem Inf Model 2013;53:2874–83.
39. Zhang H, Unal H, Gati C, Han GW, Liu W, Zatsepin NA, et al. Structure of the angiotensin receptor revealed by serial femtosecond crystallography. Cell 2015;161:833–44.
40. Santos Filho OA, Bicca De Alencastro R. Modelagem de proteínas por homologia. Brazil: Quimica Nova; 2003, 26.
41. McConkey BJ, Sobolev V, Edelman M. The performance of current methods in ligand-protein docking. Curr Sci 2002;83:845–56.
42. Goodsell DS, Lauble H, Stout CD, Olson AJ. Automated docking in crystallography: analysis of the substrates of aconitase. Proteins: Struct Funct Genet 1993;17:1–10.
43. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, et al. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. Journal of Computational Chemistry 1998;19:1639–62.
44. Goodsell DS, Olson AJ. Automated docking of substrates to proteins by simulated annealing. Proteins: Struct Funct Genet 1990;8:195–202.
45. Abagyan R, Totrov M, Kuznetsov D. ICM—a new method for protein modelling and design: applications to docking and structure prediction from the distorted native conformation. J Comput Chem 1994;15:488–506.
46. McMartin C, Bohacek R. QXP: powerful, rapid computer algorithms for structure-based drug design. J Comput Aided Mol Des 1997;11:333–44.
47. Meng X-Y, Zhang H-X, Mezei M, Cui M. Molecular docking: a powerful approach for structure-based drug discovery. Curr Comput Aided Drug Des 2011;7:146–57.
48. Verdonk M, Cole J, Hartshorn M, Murray C, Taylor R. Improved protein-ligand docking using GOLD. Protein: Struct Funct Genet 2003;52:609–23.
49. Clark KP, Ajay. Flexible ligand docking without parameter adjustment across four ligand-receptor complexes. J Comput Chem 1995;16:1210–26.

50. Taylor JS, Burnett RM. DARWIN: a program for docking flexible molecules. Protein: Struct Funct Genet 2000;41:173–91.
51. Aqvist J, Luzhkov V, Brandsdal B. Ligand binding affinities from MD simulations. Acc Chem Res 2002;35:358–65.
52. Verkhivker GM, Bouzida D, Gehlhaar DK, Rejto PA, Arthurs S, Colson AB, et al. Deciphering common failures in molecular docking of ligand-protein complexes. J Comput Aided Mol Des 2000; 14:731–51.
53. Feher M, Deretey E, Roy S. BHB: a simple knowledge-based scoring function to improve the efficiency of database screening. J Chem Inf Comput Sci 2003;43:1316–27.
54. Ewing TJA, Makino S, Skillman AG, Kuntz ID. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. J Comput Aided Mol Des 2001;15:411–28.
55. Böhm HJ. LUDI: rule-based automatic design of new substituents for enzyme inhibitor leads. J Comput Aided Mol Des 1992;6:593–606.
56. Eldridge M, Murray C, Auton T, Paolini G, Mee R. Empirical scoring functions: the development of a fast-empirical scoring function to estimate the binding affinity of ligands in receptor complexes. J Comput Aided Mol Des 1997;11:425–45.
57. Langer T, Hoffmann R. Virtual screening an effective tool for lead structure discovery. Curr Pharmaceut Des 2005;7:509–27.
58. Lapinsh M, Prusis P, Gutcaits A, Lundstedt T, Wikberg J. Development of proteo-chemometrics: a novel technology for the analysis of drug-receptor interactions. Biochim Biophys Acta Gen Subj 2001;1525:180–90.
59. Scitor T, Medina-Franco J, Do Q-T, Martínez-Mayorga K, Yunes Rojas J, Bernard P. How to recognise and workaround pitfalls in QSAR studies: a critical review. Curr Med Chem 2009;16:4297–313.
60. Baker M. Fragment-based lead discovery grows up. Nat Rev Drug Discov 2012;12:5–7.
61. Moumné R, Catala M, Larue V, Micouin L, Tisné C. Fragment-based design of small RNA binders: promising developments and contribution of NMR. Biochimie 2012;94:1607–19.
62. Warner K, Homan P, Weeks K, Smith A, Abell C, Ferré-D'Amaré A. Validating fragment-based drug discovery for biological RNAs: lead fragments bind and remodel the TPP riboswitch specifically. Chem Biol 2014;21:591–5.
63. Hann M, Leach A, Harper G. Molecular complexity and its impact on the probability of finding leads for drug discovery. J Chem Inf Comput Sci 2001;41:856–64.
64. Boyd S, de Kloe G. Fragment library design: efficiently hunting drugs in chemical space. Drug Discov Today Technol 2010;7:e173–80.
65. Hajduk P. Fragment-based drug design: how big is too big? J Med Chem 2006;49:6972–6.
66. Katsila T, Spyroulias G, Patrinos G, Matsoukas M. Computational approaches in target identification and drug discovery. Comput Struct Biotechnol J 2016;14:177–84.
67. Allen MP, Tildesley DJ. Computer simulation of liquids (Oxford science publications) SE - Oxford science publications. London: Oxford University Press; 1989.
68. Frenkel D, Smit B, Tobochnik J, McKay S, Christian W. Understanding molecular simulation. Comput Phys 1997;11:351.
69. Wang J, Wolf R, Caldwell J, Kollman P, Case D. Development and testing of a general Amber force field. J Comput Chem 2004;25:1157–74.
70. Vanommeslaeghe K, Hatcher E, Acharya C, Kundu S, Zhong S, Shim J, et al. CHARMM general force field: a force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. J Comput Chem 2010;31:671–90.
71. Zhao H, Caflisch A. Molecular dynamics in drug design. Eur J Med Chem 2015;91:4–14.
72. Jain A. Virtual screening in lead discovery and optimisation. Curr Opin Drug Discov Dev 2004;7: 396–403.

73. Reddy AS, Pati SP, Kumar PP, Pradeep HN, Narahari SG. Virtual screening in drug discovery - a computational perspective. Curr Protein Pept Sci 2007;8:329–51.
74. Stumpfe D, Ripphausen P, Bajorath J. Virtual compound screening in drug discovery. Future Med Chem 2012;4:593–602.
75. Ertl P, Schuffenhauer A, Renner S. Chemoinformatics and computational chemical biology, Bajorath J, editor. Totowa, NJ: Humana Press; 2011, 672 p.
76. Klopmand G. Concepts and applications of molecular similarity, by Mark A. Johnson and Gerald M. Maggiora, eds., John Wiley & Sons, New York, 1990, 393 pp. Vol. 13, J Comput Chem 1992. 539–540 pp.
77. Willett P. Similarity-based virtual screening using 2D fingerprints. Drug Discov Today 2006;11: 1046–53.
78. Tovar A, Eckert H, Bajorath J. Comparison of 2D fingerprint methods for multiple-template similarity searching on compound activity classes of increasing structural diversity. ChemMedChem 2007;2:208–17.
79. Wermuth CG, Ganellin CR, Lindberg P, Mitscher LA. Glossary of Terms Used in Medicinal Chemistry (IUPAC Recommendations 1997). Annu Rep Med Chem 1998;385–95. https://doi.org/10.1016/S0065-7743(08)61101-X.
80. Qing X, Lee X, De Raeymaeker J, Tame J, Zhang K, De Maeyer M, Voet A. Pharmacophore modeling: advances, Limitations, and current utility in drug discovery. J Recept Ligand Channel Res 2014;7: 81.
81. Esposito EX, Hopfinger AJ, Madura JD. Methods for applying the quantitative structure-activity relationship paradigm. In 2004. p. 131131–213. https://doi.org/10.1385/1-59259-802-1.
82. Barbault F, Maurel F. Simulation with quantum mechanics/molecular mechanics for drug discovery. Expert Opin Drug Discov 2015;10:1047–57.
83. van Westen G, Wegner J, IJzerman A, van Vlijmen H, Bender A. Proteochemometrics modeling as a tool to design selective compounds and for extrapolating to novel targets. MedChemComm 2011;2: 16–30.
84. Bongers B, IJzerman A, van Westen G. Proteochemometrics – recent developments in bioactivity and selectivity modeling. Drug Discov Today Technol 2019;32–33:89–98.
85. Zhang L, Tan J, Han D, Zhu H. From machine learning to deep learning: progress in machine intelligence for rational drug discovery. Drug Discov Today 2017;22:1680–5.
86. Lecun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44.
87. Ma J, Sheridan RP, Liaw A, Dahl G, Svetnik V. Deep neural nets as a method for quantitative structure-activity relationships. J Chem Inf Model 2015;55:263–74.
88. Yuan Y, Zhao Z, Hu R, Li J, Zhang R, Lu J. Using deep learning for compound selectivity prediction. Curr Comput Aided Drug Des 2016;12. https://doi.org/10.2174/1573409912666160219113250.
89. Unterthiner T, Mayr A, Klambauer G, Steijaert M, Wegner JK, Ceulemans H. Deep learning as an opportunity in virtual screening. In: NIPS workshops. Montreal: Neural Information Processing System Foundation, Inc; 2014.
90. Pereira J, Caffarena E, Dos Santos C. Boosting docking-based virtual screening with deep learning. J Chem Inf Model 2016;56:2495–506.
91. Kadurin A, Aliper A, Kazennov A, Mamoshina P, Vanhaelen Q, Khrabrov K, Zhavoronkov A. The cornucopia of meaningful leads: applying deep adversarial autoencoders for new molecule development in oncology. Oncotarget 2017;8:10883–90.
92. Segler M, Kogej T, Tyrchan C, Waller M. Generating focused molecule libraries for drug discovery with recurrent neural networks. ACS Cent Sci 2018;4:120–31.
93. Hughes T, Miller G, Swamidass S. Modeling epoxidation of drug-like molecules with a deep machine learning network. ACS Cent Sci 2015;1:168–80.

94. Goh G, Hodas N, Vishnu A. Deep learning for computational chemistry. J Comput Chem 2017;38: 1291–307.
95. Liang M, Li Z, Chen T, Zeng J. Integrative data analysis of multi-platform cancer data with a multimodal deep learning approach. IEEE/ACM Trans Comput Biol Bioinf 2015;12:928–37.
96. Wen M, Zhang Z, Niu S, Sha H, Yang R, Yun Y. Deep-learning-based drug-target interaction prediction. J Proteome Res 2017;16:1401–9.
97. Alipanahi B, Delong A, Weirauch M, Frey B. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol 2015;33:831–8.
98. Kwon S, Yoon S. DeepCCI. In: Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics. New York, NY, USA: ACM; 2017: 203–12 pp.