Dev Bukhsh Singh *Editor*

Computer-Aided Drug Design

Springer

Computer-Aided Drug Design

Dev Bukhsh Singh
Editor

Computer-Aided Drug Design

Springer

Editor
Dev Bukhsh Singh
Department of Biotechnology,
Institute of Biosciences and Biotechnology
Chhatrapati Shahu Ji Maharaj University
Kanpur, Uttar Pradesh, India

ISBN 978-981-15-6814-5 ISBN 978-981-15-6815-2 (eBook)
https://doi.org/10.1007/978-981-15-6815-2

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Foreword

Ever since the cracking of the human genome in the beginning of the present century, scientists have been engaged in locating the drug targets and designing and developing novel drugs through the system's approach. This has resulted in a tremendous reduction in research and production costs. Earlier, the drug design process used to take many decades and was carried out haphazardly without any direction. Already the surge in bioinformatics solutions has redefined the way drug trials are done and making a shift from in vitro to in silico. In this age of multiple drug resistance, in silico drug design could be used to shorten the time of discovery and this issue shall remain the biggest challenge for years to come. In the present fast changing scenario, it is difficult to manage expressive coherence in this rapidly growing area of drug designing.

I am happy that Dr. Dev has ventured to collect twelve well-written chapters and has brought an edited book named "Computer-Aided Drug Design" to be published by Springer Nature, Singapore. I feel that the authors are quite successful in "fusing" the otherwise diverse topics of this fast-emerging area. I am sure that this book will be exceedingly useful for not only under- and postgraduate students but also for research scholars, scientists, and pharma industries involved in developing new drugs.

I hope that the readers of this book shall contribute in the future for making the text more useful for further development of this important field of computer-aided drug design.

Hony. Professor, IIIT-Allahabad
Prayagraj, India

Krishna Misra

Preface

The computer-aided drug design uses computational approaches for analysis of target, screening, and interaction of ligands, simulation of target–ligand complex, optimization of lead compounds, QSAR analysis, and ADMET studies. In structure-based drug designing, ligand molecules are built keeping in mind the binding cavity of the target by assembling small substructures in a stepwise manner. Ligand-based drug designing involves the 2D/3D analysis and chemical modification of ligand known to interact with a drug target of the disease. A large number of computational tools have been developed to fulfill the different objectives in the way of drug designing. There are many successful stories of computer-aided drug designing. This field has attracted many researchers working in diverse fields of knowledge such as chemistry, physics, biology, mathematics, and computer science. In drug designing, systematic and sequential use of different computer-aided drug designing tools/software is required. Much advancement has taken place in the algorithms and approaches of computer-aided drug designing from time to time. The existing limitations of the tools and approaches used for drug designing have also been discussed which can motivate the readers and researchers to overcome such challenges in the future.

The present book "Computer-Aided Drug Design" has been written considering the need for researchers and students working in the domain of computer-aided drug designing. This book not only represents the discussion of recent advances in the field of computer-aided drug designing but also provides a basic knowledge of principles, approaches, and tools used for drug designing. This book includes a discussion of biological database resources used for drug discovery. One chapter is focused on the computational approaches and resources used for vaccine designing. Similarly, a basic discussion and application of machine learning approaches such as genetic algorithm, artificial neural network, and support vector machine have been included. It also explains the basics and use of different biological, physical, and chemical parameters used for modeling, simulation, and ADMET prediction. The chapters provide a summary of related case studies along with the application, merits/demerits, limitations, and future perspectives related to the title. The steps and use of different computational approaches have been explained with the help of simple, suitable, and neat sketches and illustrations. This book is full of a lot of resources that can guide and motivate a learner to proceed for drug designing.

I hope this book will be very helpful in understanding the basics and recent advances in computer-aided drug design. I tried my best effort to present a good quality creation before the readers and other scientific communities. This book will cover the need for a broad spectrum of subjects such as bioinformatics, biotechnology, biochemistry, and pharmaceutical sciences. During the review and editing process, many suggestions, corrections, and suitable addition of new topics have been included. Still, I look forward to your valuable suggestions and feedback related to the content quality of the book.

Kanpur, India Dev Bukhsh Singh

Acknowledgement

I am highly grateful to Prof. (Mrs.) Krishna Mishra (Prayagraj), a renowned scientist and educator for her continuous support, guidance, and motivation. Prof. J.V. Vaishampayan, former Vice-Chancellor of CSJM University, Kanpur has encouraged me a lot to achieve academic excellence. I will always be highly thankful to him for his inspiration, encouragement, and support. I would like to acknowledge the effort of all the authors of this book for their extensive labor, vision, and planning in writing the chapters. I thank the reviewers whose critical comments improved the book in substantial ways. I am highly thankful to Dr. Pankaj Kumar Singh, GBPUA&T, Pantnagar for his technical support and suggestions. I am highly grateful to my parents Mr. Sudhakar Singh and Smt. Radhika Singh and other family members for their wishes, valuable support, and encouragement.

I am also thankful to my colleagues Dr. Manish Kumar Gupta, Dr. Satendra Singh, Dr. P. K. Yadav, Dr. Budhayash Guatam, Dr. Prashant Ankur Jain, Dr. Anil Kumar, Dr. Durg Vijay Singh, Dr. Ajay Kumar Singh, Dr. K. K. Ojha, Dr. Pramod Katara, Dr. Prem Kumar Singh, Dr. R. K. Kesharwani, other friends and staff members for their support. I would like to appreciate the effort of Dr. Rajesh Kumar Pathak, Mr. Rohit Shukla, Mr. Apoorv Tiwari, Mr. Himanshu Avasthi, Mr. Ambuj Srivastava, and Ms. Shikha Agnihotri for their support. I am thankful to Dr. Bhavik Sawhney and the entire team of Springer Nature for their continuous support and cooperation during the entire process of publication.

Contents

Machine Learning Approaches to Rational Drug Design

12

Salman Akhtar, M. Kalim A. Khan, and Khwaja Osama

Abstract

Pharmaceutical industries are multibillionaire setups with a diligent team of scientists, researchers, technical manpower, and investors. A major concern of such industries is to always curtail the time and cost factor associated with them. Bioinformatics involving machine learning (ML) methods have come to the forefront to address this problem. The predictive and statistical efficacy of ML methodologies has even proven to propose better leads than a wet lab pipeline. This chapter aims to give a brief overview of underlying principles of mainly GAs and ANNs as popular ML algorithms and deeper insight into their robust applications in the field of modern day drug design. It also attempts to share the future prospects of such ML techniques and their limitations with possible solutions hereafter.

Keywords

Machine learning · Genetic algorithms · Artificial neural networks · Drug designing · Deep learning · Support vector machines

12.1 Drug Industry

In the current era, drug industries have expanded rapidly and have seen enormous growth in terms of infrastructure, scientific manpower, technical handling, and business product outputs. Health is a major concern nowadays of every livelihood and various health and pharmaceutical products have become a routine part of the daily diet of many individuals. More importantly, a rapid increase in the knowledge

S. Akhtar (✉) · M. K. A. Khan · K. Osama
Department of Bioengineering, Integral University, Lucknow, India

D. B. Singh (ed.), *Computer-Aided Drug Design*,
https://doi.org/10.1007/978-981-15-6815-2_12

and recognition of various diseases, improved diagnostics tools, government promotional health schemes for all and varied networking between scientific researchers around the world has facilitated the need of development and subsequent financial investment in more and more pharmaceutical industrial setups (Zhong et al. 2018; Leelananda and Lindert 2016; Fox and Kriegl 2006).

Usually, the development of a novel drug against a particular disease requires a timeline of 12–15 years and a hugely expensive experimental setup accounting to nearly 200–800 million$. However, with the development of novel and affordable bioinformatics in silico setups, the time and cost required in the development of new drug molecules has been significantly curtailed. Bioinformatics, an in silico science comprising of various computer based prediction, search and optimization methods, has not only opened newer dimensions and directions in novel drug development but has also been instrumental in reducing the cost and time required against the expensive wet lab setups. Normally US Food and Drug Administration (FDA) approves a drug as defined by "a substance intended for use in the diagnosis, cure, mitigation, treatment, or prevention of disease affecting the structure or any function of the body." In a simpler sense, we can say a drug molecule as a chemical compound or more recently organic peptides which has the potential to interact with a specific biological target. Target usually includes any of the biological macromolecules that may be proteins, nucleic acids, carbohydrates, and lipids. Proteins among these stand to be the most prominent target accounting for nearly 95% of drugs designed against them followed by nucleic acids (3–4%), lipids, and carbohydrates (Yosipof et al. 2018; Huang et al. 2017).

12.2 Drug Discovery Pipeline

The important steps involved in the process of drug discovery are as described below (Fig. 12.1).

Fig. 12.1 Different phases of drug designing

12.2.1 Target Discovery

Target identification is the major step in any drug discovery process. The clinical relevance, therapeutics, and disease-causing potential of the selected target actually govern the efficacy of the designed drug molecule inside the body. The step may include a blend of diverse biological assays to high-throughput screening studies in the identification of a single target for a particular disease (Huang et al. 2018; Leelananda and Lindert 2016).

12.2.2 Target Validation

Target validation is carried out mainly by applying biotechnological innovational wet lab studies involving gene knock out in animal models, small molecule agonists/antagonists, aptamers, siRNA, ribozymes, and neutralizing antibodies. The basic aim of target validation studies is to validate the therapeutic efficacy of the selected target and to scientifically reveal that the selected target possesses significant disease-causing potential.

12.2.3 Lead Identification

Lead compounds are all those chemical compounds that possess some but not all properties of the drug molecule. These are the starting clinical agents from where potential drug candidates can be synthesized. The discovery of lead compounds is fastened up by employing virtual HTS studies, where millions of chemical compounds are screened using computer-generated models. Molecular docking along with pharmacophore modeling offers faster, effective, and cost-efficient screening solutions in large drug discovery projects (Schneider and Bohm 2002).

12.2.4 Lead Optimization

This process involves the customized structural and functional modification of lead compounds to enhance their inhibition potential against a particular target. Involving various structure-based and ligand-based drug designing strategies and quantitative structure–activity relationship (QSAR) studies it promotes structural and functional modification of leads for its increased efficacy. This is often the most critical and diligent step in drug discovery.

12.2.5 Preclinical Phase

It involves the studies on animal systems to access the dosage and adverse side effects associated with drug molecules.

12.2.6 Clinical Trials

The clinical lead development involves 4 phases:

12.2.6.1 Phase I

A small group of normal healthy volunteers is selected in this phase and the developed drug is tested on them to assess its safety, tolerability, dosage levels, pharmacokinetics, and pharmacodynamics inside the human body. 80% of drugs are reported to fail in Phase I clinical trial.

12.2.6.2 Phase II

Phase II involves the controlled clinical studies in a hospital conducted on a group of patients, to obtain average data on the efficacy and dosage level of the drug. The trial is carried out in an unbiased manner using a placebo and newer drug simultaneously, in a group of patients with particular indications and symptoms of the disease.

12.2.6.3 Phase III

Phase III studies involve randomized controlled trials on large patient groups in medical institutions and hospitals. Market regulatory affairs and commercial launch decisions of the drug are also accompanied in this step.

12.2.6.4 Phase IV

It involves long term monitoring process after the drug is launched in the market. Post-launch this may be mandated or initiated by the pharmaceutical company in assistance with a team of medical representatives and medicos. It is destined to detect long term rare or serious adverse effects of medicine over a large patient population.

12.3 Dimensions and Complexity of the Problem and Role of ML Techniques

With the ever-increasing number of potential lead compounds derived through virtual screening studies and its subsequent attempt to prioritize them based on their functional groups, physicochemical descriptors, and biological activities has compelled the researchers to use statistical function optimizers and prediction algorithms in handling such big data. The dimensionality in terms of thousands of molecular descriptors and complexity in terms of libraries of millions of compounds guarantees the application of computers as inevitable rather than any sort of manual calculations.

Chemoinformatics, a built-up control in which meaningful data are extricated, processed, and extrapolated from chemical structures plays an important role in solving such a problem (Lo et al. 2018). Chemoinformatics is the utilization of informatics strategies to fix chemical issues including chemical information retrieval and extraction, database exploration, and molecular chart mining (Varnek and Baskin 2011). The design of new drugs is an extremely difficult and complex

issue (Schneider and Schneider 2016). The scalable search space for new and unfamiliar particles is one of the major barriers in drug design and configuration. Scientific experts need to choose and analyze particles from this extensive space to discover molecules that are dynamically active towards the respective target protein. Drug discovery is an enormously expensive and sluggish procedure representing various formidable difficulties. In this manner, technological advances in drug research and preclinical improvement could bring down the expenses of conveying another novel drug to the market (Pu et al. 2019).

In silico strategies try to quicken this procedure and diminish the expensive late-stage breakdown by utilizing the chemical information produced through computational techniques and high-performance assays to reveal hidden connections among the data. Screening vast virtual libraries of compounds with improved biological properties such as explicitness and selectivity toward the respective target, lower toxicity, or reduced cost help to discover new chemical or synthetic inhibitors. With the advancement of "big data" from HTS and combinatorial synthesis, ML has turned into a key apparatus for drug designers to extract synthetic information from substantial compound libraries to devise drugs with essential biological features (Lo et al. 2018). Other modules of chemoinformatics additionally incorporate computer-aided drug synthesis, chemical space investigation, pharmacophore, and scaffold testing, library preparation, etc. (Kapetanovic 2008). Various ML-based strategies thus have been created and ubiquitously used to identify and develop new drugs having superior biological activities to uncover complex relations between chemical and their biological targets. Mathematical mining of chemical graphs enables the inference of a group of 2D or 3D chemical descriptors bundled in a variety of ML models and probabilistic tasks as chemical fingerprints. Integrating numerous data types and sources or the so-called data fusion procedures which collate hereditary, structural, and pharmacological information in different level of organisms seems to be critical for the disclosure of more efficacious and safer drugs (Chen et al. 2018; Searls 2005).

In recent years, ML techniques have gained significant attention as a prominent research and development tool in rational drug designing approaches. Precisely genetic algorithms (GA) and artificial neural networks (ANN) have come to the forefront to uptake the diversified challenges faced in pharmaceutical industries (Zhong et al. 2018; Lavecchia 2015). GAs in the category of evolutionary algorithms and ANNs as a case of artificial intelligence (AI) have seen immense applications as ML techniques in searching and optimizing compounds, predicting active site and binding conformations, the establishment of quantitative structure–activity relationships, gene prediction, pharmacophore analysis, design of combinatorial libraries, and so forth (Goswami et al. 2018; Sayeed et al. 2016; Gupta et al. 2014; Arif et al. 2013). Furthermore, applications of NNs in drug design are concerned in areas like lead discovery; designation and estimation of biological activity; and A (absorption), D(distribution), M(metabolism), E(excretion)/Tox(toxicity) assets; multidimensional data processing; compound library analogy; combinatorial library striking similarity and diversity analysis; HTS data analysis (Terfloth and Gasteiger 2001).

Techniques of ML can be comprehensively named as supervised or unsupervised learning. Training data are assigned to labels for supervised learning, and once prepared; the model can foresee labels for specific data inputs. Regression analysis, random forests, naive Bayes, ANN, k-nearest neighbor (kNN), SVM algorithms are some popular instances of supervised ML models. While unsupervised ML methods gain directly from unlabeled data of molecular patterns, i.e. it needs only input data with no relating output factors. Here, the basic pattern or structure in the data is determined for the use of future analysis and prediction. Independent components analysis (ICA) and principal components analysis (PCA) are the common algorithms of unsupervised learning. These problems could be further categorized into clustering and association groups (Yang et al. 2018).

Capacity in handling the large amount of data employing immense computational power, these ML strategies are certainly proving a state-of-art in the field of rational drug design.

12.4 Genetic Algorithms

GA is a family of population-based computational models that are inspired by nature's natural process of evolution. GAs come under the category of search and optimization algorithms and are often viewed as function optimizers to solve a varied kind of problems. Initially given by John Holland in 1975, GAs have come a long way in providing solutions to the problems which have multiple inputs and multiple outputs (Goswami et al. 2018; Mandal et al. 2007).

A simple implementation of the normal GAs begins with a population of individuals which are encoded on a simple chromosome like data structures. One then evaluate these structures and allocate reproductive opportunities to these individuals based on their relative fitness. Selection is applied to this population so that the individuals with good solutions to the problems are given more chances to reproduce than the individuals with poorer solutions. The selection process generates an intermediate population after which recombination and mutation are applied to generate the next population (offspring). The process of evaluation, selection, recombination, and mutation from parent to offspring constitutes one generation of the GA. GAs likewise nature run for a number of generations and provide an optimal considerable solution to the problems, much better than their initial parent solution.

12.4.1 Working of Genetic Algorithms

Usually, there are two major components of GA:

1. Problem encoding.
2. Evaluation function.

The first assumption that is typically made is that the variable representing the parameters is represented in the form of binary strings (0, 1). This means the variable is discretized in the search space to some power of two (2^n). After creating the initial population, each binary string is then evaluated and assigned a fitness value. The fitness is defined by

$$\text{Relative fitness} = f_i / < f >$$

where f_i = evaluation associated with string i and $<f>$ = average evaluation of all the strings in the population.

The fitness value obtained normally translates the measure of performance of each individual into the number of copies passed onto the next generation. The obtained fitness function generates a population of current parent individuals, which are now subjected to various GA operators to create the next generation of offspring.

12.4.2 Genetic Algorithm Operators

12.4.2.1 Natural Selection Operator

Selection is normally applied to the current population to generate the intermediate population (F_1 generation) in such a way that the highly fit individuals with good solutions to problems are given more share or reproductive opportunities into the next generation against lesser fit individuals (Fig. 12.2). There are a number of ways to do the selection.

Parameter	Fitness
P1	15
P2	10
P3	8
P4	3
P5	2
P6	2
P7	1
P8	1

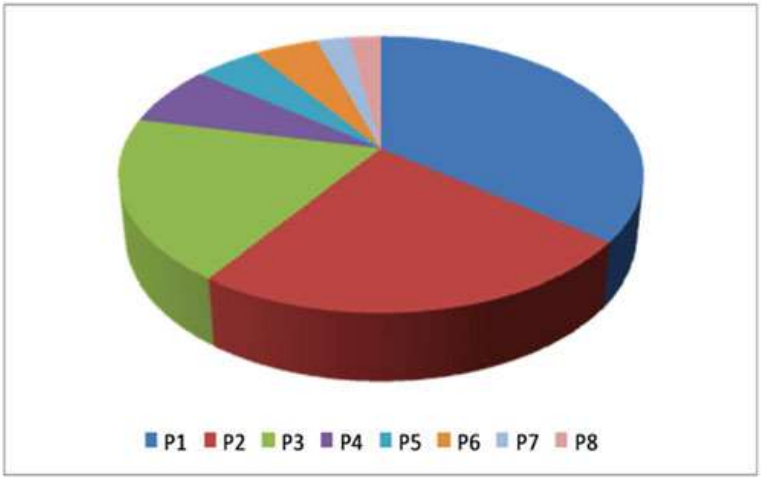


Fig. 12.2 Application of selection operator in GA

12.4.2.2 Stochastic Sampling with Replacement

We might visualize the population as mapping onto a Roulette wheel, where each individual is assigned its space which is directly proportional to its relative fitness. Iteratively spinning this roulette wheel, individuals are chosen and assigned to the intermediate generation in an unbiased manner.

12.4.2.3 Stochastic Universal Sampling

The population is laid down in random order on a simple pie (360°) graph, where again each individual is assigned its space, which is directly proportional to its relative fitness. Next, an outer Roulette wheel with N equally spaced pointers is placed over this pie graph. A single spin of this outer Roulette wheel unbiasedly picks all N individuals as members of the intermediate population.

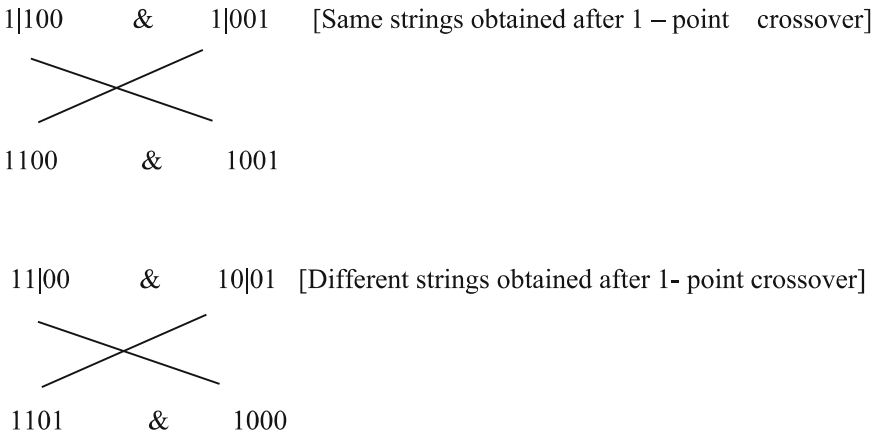
12.4.2.4 Crossover/Recombination Operator

After the selection process has been carried out, the construction of the F_1 generation is complete and now crossover can be implemented. Crossover or recombination is a natural phenomenon that occurs during the late pachytene stage of meiosis and generally involves the exchange of chromosomal segments between the paternal and maternal chromosomes. The prime benefit of the crossover operator is

1. It introduces important genetic diversity in the current population.
2. It naturally preserves the critical information of the parents.

Crossover is normally applied in GAs by randomly recombining a pair of strings with a certain probability $P_{c.}$

Picking a pair of strings and recombining them by using 1-point crossover has the potential to generate newer solutions to the problems or may also subside as like its previous solutions.



12.4.2.5 Mutation Operator

After recombination, we can apply a mutation operator. A mutation is simply interpreted to actually mean flipping of a bit to form a novel combination. For each bit in the population, mutate with some low probability *Pm*.

The mutation is generally seen as a necessary evil in nature and therefore is typically applied as less than 1% in F_1 generation. The mutation has a capacity to introduce a novel solution to the problem which a normal recombination operator even after running for multiple generations fails to do so.

After the completion of selection, recombination, and mutation, the next population can be evaluated. The process keeps on running for multiple generations (>25,000 generations) to provide an optimal solution in the end in the successful compilation of GA.

12.5 Artificial Neural Networks

ANN is an information processing paradigm, which works in a way as the human biological nervous system works. Unlike GA, ANNs are purely a case of AI and are inspired by the human brain mechanism of information processing and exchange (Fig. 12.3). The key element of any ANN is its novel design of information processing structure, destined to solve a particular problem.

Likewise, the human nervous system, ANNs are composed of a large number of interconnected processing elements (artificial neurons) that work in unison to provide an optimal response to a specific stimulus/problem (Fig. 12.4). ANN works by learning from examples, by making adjustments in synaptic connections among its artificial processing elements/neurons. ANNs have been largely used in data recognition, pattern recognition, and prediction problems owing to their capacity for self-intelligence (Gupta et al. 2012). However, the results of ANNs are expected to be unpredictable and cannot be seen to perform miracles but if used sensibly in correct design ANNs have been seen to produce amazing results often.

Neural network simulations have witnessed its establishment even before the advent of computers but due to lack of funding and enthusiasm, this field suffered frustration and disrepute. The first artificial neuron was even developed by Warren McCulloch and Walter Pits in 1943 but later in a research paper published by Minsky and Papert in 1969, the concept of ANNs suffered a major setback and significant limitations.

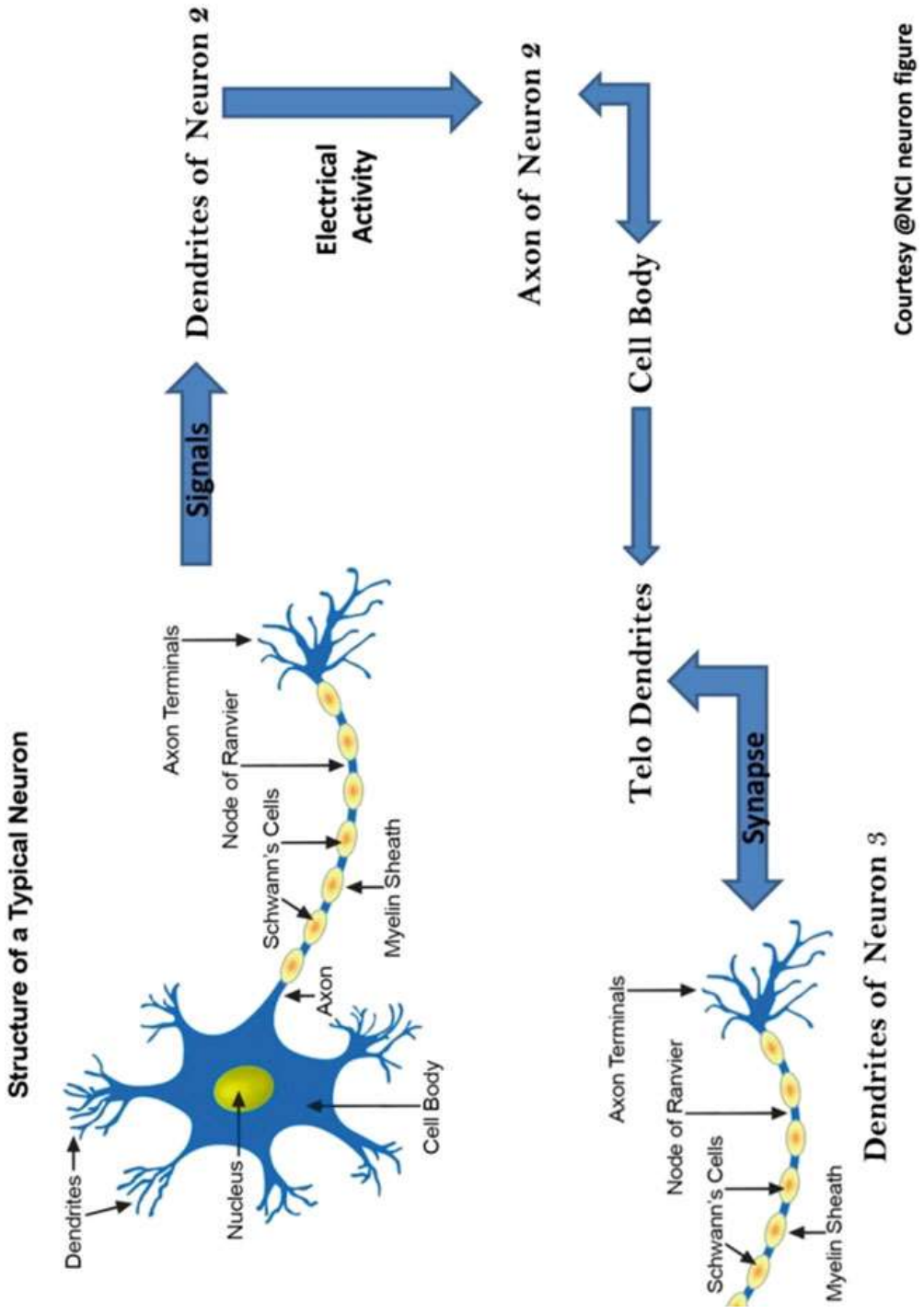


Fig. 12.3 Working of human neuronal cells

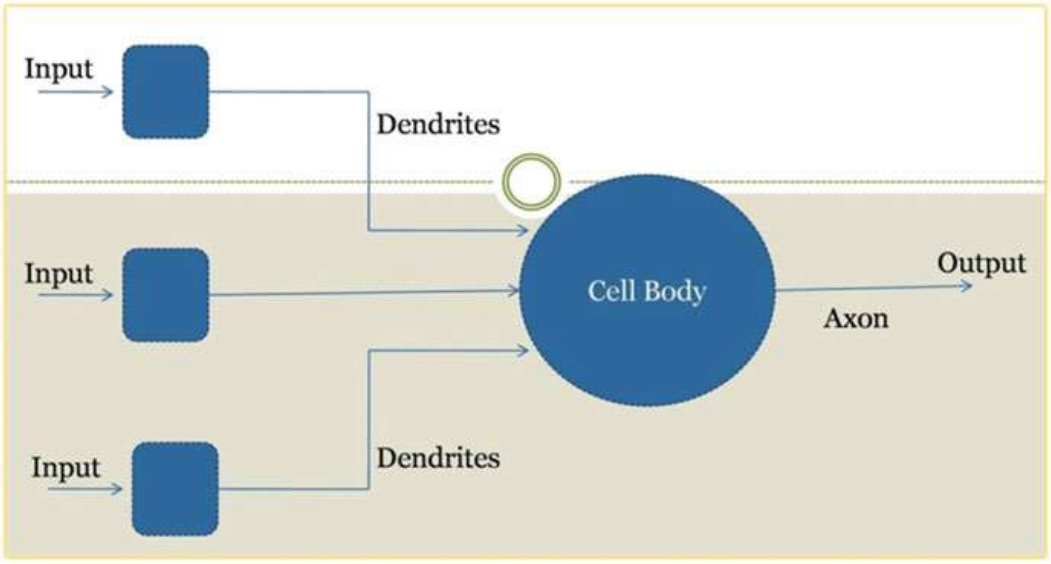


Fig. 12.4 Structure of an artificial neuron

In the current scenario, ANNs have seen a major comeback, and owing to its remarkable ability to derive meaning from complex and indefinite data, they are now been routinely used in data handling and data analysis problems in various sectors. The major advantages of ANNs include its self-organization capacity, adaptive learning mechanism, real-time operation, and fault tolerance attitude. ANN can learn from its training data and automatically create its own organization by making synaptic adjustments. In a state of its dynamic equilibrium, it can produce amazing results (Doyle et al. 2015; Oquendo et al. 2012).

Unlike computers, ANNs are not programmed and are unpredictable. They are better known to solve those problems which it does know how to solve. They can give any result based on the input and do not use a cognitive or algorithmic approach to problem-solving.

12.5.1 How the Human Brain Works?

Learning in the human brain occurs by changing the effectiveness of the synapses which is expressed in the form of change in the influence of one neuron on others. Synapses are the region of contact between two neurons or a neuron with a non-neuronal cell. The transmission of information between neurons may be electrical or chemical depending on the myelination of the nerve fibers.

12.5.2 A Simple Artificial Neuron

An artificial neuron is simply a device with multiple inputs but a single output. It works in basically two modes of operation

1. Using mode: When a taught input pattern is detected, its associated output becomes its natural output.
2. Training mode: In case when the input is not a part of a taught pattern, the neuron is trained to fire or not based on the learning from the taught input patterns. Learning from the taught input pattern is accomplished via the implementation of a firing rule.

Firing rule is an important concept of ANNs and provides its high flexibility. This rule helps ANN to calculate for a particular neuron to fire or not for a given input using a simple hamming distance technique. An example pattern is presented here for its detailed understanding. A 3-input neuron is taught to fire output 1 for the given inputs as 111 or 101. Similarly, it is taught for output 0 if the inputs are 000 and 001. Therefore before the application of firing rule, the truth table could be seen as:

X1	0	0	0	0	1	1	1	1
X2	0	0	1	1	0	0	1	1
X3	0	1	0	1	0	1	0	1
OUT	0	0	0/1	0/1	0/1	1	0/1	1

Implementing the firing rule, we take a pattern, for example, 010. It differs from 000 in 1 element and 001 in 2 elements. Similarly, it differs from 101 in 3 elements and 111 in 2 elements. Therefore, applying the hamming distance concept to the nearest input it belongs to 000. The output associated with 000 is 0, so after the learning process, the output associated with 010 also stands to be 0, instructing the neuron not to fire for this particular input.

By applying the firing rule to every column of this truth table, the new truth table generated can be represented as:

X1	0	0	0	0	1	1	1	1
X2	0	0	1	1	0	0	1	1
X3	0	1	0	1	0	1	0	1
OUT	0	0	0	0/1	0/1	1	1	1

In this way, the firing rule helps the neurons to understand the output for a given novel input based on the learning from the already present taught input patterns. A more complicated neuron model includes the application of weighted inputs. The weight is usually a number, which is multiplied with input to gives its total weighted input value, and if they exceed a certain threshold value, the neurons fire else not.

The commonest type of ANNs comprises of three layers of networking (Fig. 12.5):

1. Input layer: represents the raw data fed into the network.
2. Hidden layer: activity depends upon the input units and weights of connection between input and hidden layer.

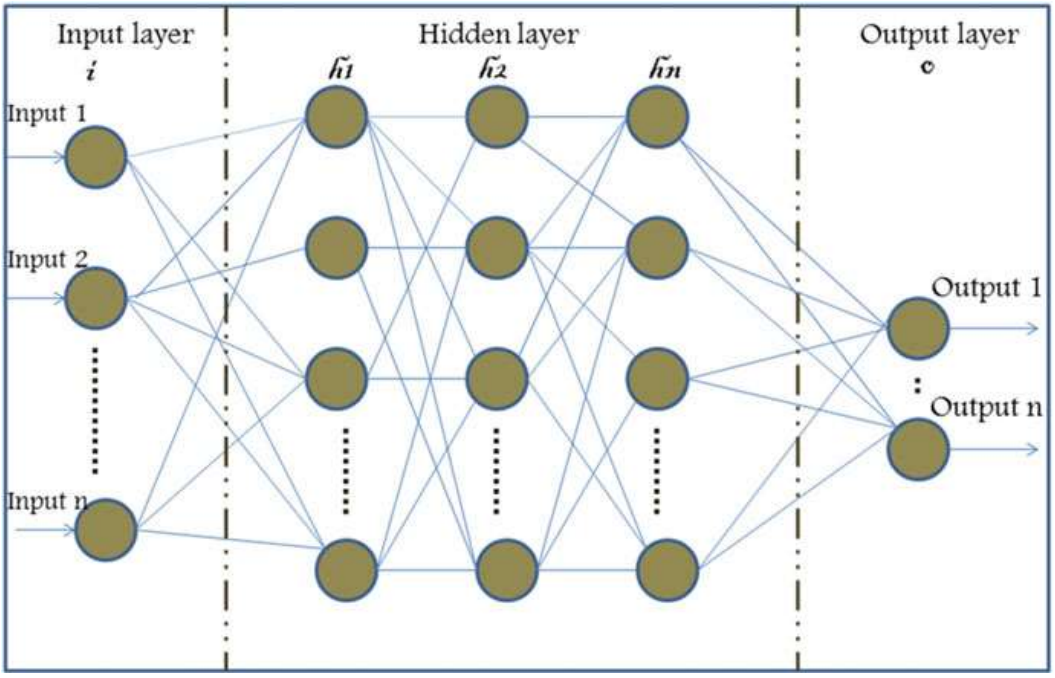


Fig. 12.5 Network layers in artificial neuron network

3. Output layer: activity depends upon the hidden layer and weights of connections between hidden and output units.

Hidden layers are just like a black box and are free to construct their own representation of data and reach a dynamic equilibrium stage (Patra et al. 2017).

12.5.3 Architecture of ANNs

There are two types of ANN architectures:

1. Feedforward networks: These networks allow the signals to travel in a unidirectional manner, i.e. from input to output only. They are straight forward networks and no feedback is provided to the back end layers or the layers on the same level. Also referred to as bottom up or top down networks, these feedforward networks are regularly used in data and pattern recognition problems.
2. Feedback networks: These networks allow the signals to travel in both directions by the introduction of loops in the network. They are a very powerful network and can get extremely complicated in their action. Their signal state keeps on changing continuously until it reaches a dynamic equilibrium. Also referred to as interactive or recurrent networks, they are well known to solve more complex problems.

ANNs are further known to work via supervised and unsupervised learning techniques.

12.6 Deep Learning(DL)

Deep learning or Deep neural networks (DNN), a class of ML algorithms uses ANNs with many layers of nonlinear processing units for learning data representations. DL is a highly adopted method to address the activity prediction problems in the first place. In drug discovery, compounds are presented by the same number of molecular descriptors, the straight forward method is to use fully connected DNNs to build models. Ma et al. 2015 applied a DNN on the Merck Kaggle challenge dataset using a large number of 2D topological descriptors, and the DNN showed better performance than the standard RF method. DNNs can handle thousands of descriptors without the need for feature selection; dropout can avoid the notorious overfitting problem faced by a traditional ANN. The study showed that multitask DNN models perform better than single-task models (Duvenaud et al. 2015). Convolutional neural networks (CNNs) are a type of neural network commonly used in image recognition (LeCun et al. 2015). CNNs are used for virtual screening in drug discovery effectively.

ANN has been applied to drug discovery in the context of ligand-based QSAR, and they have not traditionally been used in structure-based virtual screening methods. Currently, the enormous amount of biological data available for training, and the recent evolution of GPU-accelerated computation, and neural network-based techniques have very high potential to transform the in silico prediction of molecular recognition with maximum accuracy.

Durrant et al. created two fast and accurate neural network scoring functions for rescoring docked ligand poses (NNScore 1.0 and 2.0) (Durrant and McCammon 2010). Unlike traditional docking scoring functions, these nonparametric functions are not constrained to predetermined physical formulae or statistical analyses; rather, they "learn" directly from existing experimental data how best to predict binding and so can, in theory, better capture the nonlinear, synergistic relationships among binding determinants. These are the first neural network scoring functions that predict affinity by directly examining atomic resolution ligand–protein interactions. Recently, Abdul et al. developed a quadratic phenotypic optimization platform (QPOP), which provides new or best drug combinations for patients using the small dataset. This QPOP platform uses system-specific experimental data to determine the best drug combinations for a specific disease model or a patient sample rather than previous assumptions of molecular mechanisms of disease (Rashid and Chow 2019). DeepMind Technologies, a subsidiary of Google, collaborated with Royal Free London NHS Foundation Trust to assist in the management of acute kidney injury (Rashid and Hodson 2017). Atom wise (https://www.atomwise.com/) is a pioneer in healthcare AI and is the first DL technology for novel small molecule discovery and has assisted in the invention of new potential medicines for 27 disease targets and is working with top institutions such as Harvard University and Stanford

University, as well as pharmaceutical companies. Exscientia is an AI company that specializes in phenotypic drug discovery (Lee et al. 2017).

DL helps to resolve several prediction issues in bioinformatics. CNN and deep belief network (DBN) are used to unearth RNA-protein binding motifs DNA binding proteins, peptide-MHC binding, etc., stacked sparse autoencoder (SSAE) pooled with a Legendre moment (LM) has been utilized for predicting protein–protein interaction within cells (Huang et al. 2017). DL has also been effectively utilized in the QSAR studies in the form of 2D-QSAR, 3D-QSAR, and multidimensional (nD) QSAR. 2D fingerprint-based ANN (FANN)- QSAR, a novel ANN technique has been reported to calculate biological activities of structurally diverse chemical ligands efficiently. Successful implementation of FANN-QSAR is done to forecast the cannabinoid receptor (CB2) binding activity through data collection from structurally diverse sources (Myint et al. 2012).

DL has profoundly being used in computational biology for predicting protein disorder, enhancement of docked protein complexes, modeling structural properties of protein targets binding to RNA, etc. Compared to other techniques for predicting the mechanism of action with the help of high content image analysis data, it was observed to be advance to SVM with 87.62. The various framework of DL such as DNN, DBN, and RNN is currently being used for analyzing gene expression, genomic sequencing, and prediction of protein structure (Ekins 2016).

NiftyNet, a modular DL pipeline has emerged as the latest accelerated DL technique for solving problems related to medical imaging and computer-assisted intrusion. Gibson et al. utilized capability of NiftyNet to build analysis application like for layerwise separation of different organs that were obtained from computed tomography (CT), applied regression technique to foresee computed tomography attenuation maps from images of magnetic resonance of brain and creation of simulated images of ultrasound for defined poses of anatomy (Gibson et al. 2018).

12.7 Support Vector Machines

SVMs are by far the most ubiquitously used ML procedures in the design and discovery of drug case studies. SVM classifies various compound varieties to predict new molecules, biological activity (from regression models), and rank substances in virtual screening assays. SVM approaches involve several critical challenges, such as selecting kernel functions and optimizing parameters for specific issues. Integrating SVMs with other approaches has ended up being an excellent strategy for studying medicinal chemistry (Maltarollo et al. 2019).

Various ML algorithms, viz. PCA can help to identify various alveolar cells, BP—back error propagation algorithm can be used for predicting the secondary structure of the protein. CNN can assist in the detection of DNA sequence variation sites. In the field of data science, ML is a strategy to design complex models and algorithms for prediction (Niculescu 2003). A profound convolutional neural system can identify hereditary variations in aligned next-generation sequencing read data by

learning factual relationships (probabilities) between pictures of read pileups framing putative variation points and ground-truth genotypes (Yang et al. 2018).

12.8 Artificial Intelligence and Drug Discovery

The AI pioneered in 1950 discussed a technique that could sense, reason, and think like people. Later on with the advancement in computer processing power and huge datasets now targeted AI has been developed which is more focused and accurate. AI has attracted pharmaceutical industries for de novo designing of peptides and chemical compounds against a particular target, along with its retrosynthesis. AI helps in predicting the physicochemical properties (i.e. ADMET) of candidate drug compounds among the large dataset. Several machine learning technologies like Random forests (RF), SVM, or Bayesian learning have been implied for optimizing the process of compound designing (Hessler and Baringhaus 2018). AI has transformed the methods of a pathway or target identification to treat diseases. In a study, it was shown that possibility of predicting therapeutic targets using a computational prediction application known as "open targets" a platform consisting of gene-disease association data, and it was reported that animal models exhibiting a disease-relevant phenotype with a neural network classifier of greater than 71% accuracy provided the most predictive power (Ferrero et al. 2017). IBM Watson, an AI platform for drug discovery has identified five new RNA-binding proteins (RBPs) linked to the pathogenesis of a neurodegenerative disease known as amyotrophic lateral sclerosis (ALS) (Bakkar et al. 2018). AI also contributes to the identification of target-specific virtual molecules and association of the molecules with their respective target while optimizing the safety and efficacy attributes. AI, with the ability to prioritize molecules based on the ease of synthesis also assist in development of tools that are effective for the optimal synthetic route (Segler et al. 2018). ML-based tools used in drug designing are listed in Table 12.1.

AI in the drug-like compound synthesis process has proven accurate in predicting the best sought-after reactions by filling the voids that cause high failure in expected organic synthesis (commonly known as "out of scope" compounds). The voids in organic synthesis are mainly the result of unpredictable steric and electronic effects and incomplete understanding of the reaction mechanism. Although several computers aided organic compound synthesis (CAOCS) systems are available to assist chemists in selecting the synthesis route a new AI platform named 3N-MCTS developed by Seglar et al. has proven to be much faster and better than that of traditional computer-assisted retrosynthesis systems (Segler et al. 2017). An innovative AI tool, SPiDER, has been developed (Rodrigues et al. 2018) as an alternative to chemoproteomics to advance natural products for drug discovery. The SPiDER was used to predict the molecular target of b-lapachone, a clinical-stage natural naphthoquinone with antitumor activity. The platform predicted b-lapachone as an allosteric and reversible modulator of 5-lipoxygenase (5-LO). The prediction is validated using a 5-LO functional assay. Read-across structure-activity relationship

Table 12.1 Common ML-based drug designing tools

Tool	Developer	Brief description and algorithm used	Website	Freely available
Docking				
AutoDock	The Scripps Research Institute	Simulated annealing, GA	http://autodock.scripps.edu	Yes
DOCK	University of California	Incremental construction, merged target structure ensemble	http://dock.compbio.ucsf.edu	Yes
FlexX	BioSolveIT GmbH	Incremental construction, merged target structure ensemble	https://www.biosolveit.de/FlexX	No
FRED	OpenEye Scientific Software	Exhaustive search algorithm	https://docs.eyesopen.com/oedocking/fred.html	Yes
Glide	Schrödinger, Inc.	Conformational expansion, Monte Carlo, torsional search	https://www.schrodinger.com/glide	No
GOLD	The Cambridge Crystallographic Data Centre	GA	https://www.ccdc.cam.ac.uk/solutions/csd-discovery/components/gold	No
ICM-Pro	Molsoft LLC.	Monte Carlo minimization procedure	http://www.molsoft.com	No
Surflex-Dock	Tripos Inc.	Hammerhead's empirical scoring function with morphological similarity	http://www.jainlab.org	No
Homology modeling				
Modeller	University of California	It performs comparative modeling using spatial restraints and performs optimization of protein structure models	https://salilab.org/modeller	Yes
MOE	Chemical Computing Group	It uses a precompiled antibody x-ray database (Fab database) to model	https://www.chemcomp.com	No

(continued)

Table 12.1 (continued)

Tool	Developer	Brief description and algorithm used	Website	Freely available
		F_v region of the immunoglobulin		
Prime	Schrödinger, Inc.	Prime performs comparative modeling using homology modeling and fold recognition	https://www.schrodinger.com/prime	No
SWISS-MODEL	Swiss Institute of Bioinformatics	Fully automated protein homology modeling server	https://swissmodel.expasy.org/	Yes
Molecular dynamics				
Amber	University of California, San Francisco, USA	*A*ssisted *m*odel *b*uilding with an *e*nergy *r*efinement suite is used for molecular dynamics simulations of biomolecules	http://ambermd.org/	No
CHARMM	Harvard University, USA	It simulates the peptides, proteins, prosthetic groups, ligands, nucleic acids, lipids, and carbohydrates in aqueous, crystals, and membrane environments	https://www.charmm.org	No
Desmond	D. E. Shaw Research	It uses novel parallel algorithms and numerical techniques to perform molecular dynamic simulations	https://www.deshawresearch.com	Yes
GROMACS	University of Groningen, The Netherlands	Performs molecular dynamic simulations of biomolecules like proteins, lipids, and nucleic acids	http://www.gromacs.org	Yes
NAMD	University of Illinois, USA	NAMD is used for the simulation of large biomolecules	http://www.ks.uiuc.edu/Research/namd/	Yes
Quantum mechanics				
GAMESS	Iowa State University, USA	The general atomic and molecular	https://www.msg.chem.	Yes

(continued)

Table 12.1 (continued)

Tool	Developer	Brief description and algorithm used	Website	Freely available
		electronic structure system (GAMESS) is a general ab initio quantum chemistry package	iastate.edu/gamess	
Gaussian	Gaussian Inc.	It is used for estimating analytic frequency, geometric optimizations and IR, Raman, VCD and ROA spectra of large biomolecules	https://gaussian.com	No
Jaguar	Schrödinger Inc.	Jaguar uses density-functional theory (DFT) and local second-order Møller–Plesset perturbation theory to compute a comprehensive array of molecular properties	https://www.schrodinger.com/jaguar	No
MOPAC	Stewart Computational Chemistry	It is used to calculate vibrational spectra, thermodynamic quantities, isotopic substitution effects, and force constants for molecules, radicals, ions, and polymers	http://openmopac.net	Yes
NWChem	Environmental Molecular Sciences Laboratory	Ground and excited solutions of many-electron Hamiltonian are obtained utilizing density-functional theory, many-body perturbation approach, and coupled cluster expansion. It is used to analyze potential energy surface and perform dynamical simulations	http://www.nwchem-sw.org	Yes

(continued)

Table 12.1 (continued)

Tool	Developer	Brief description and algorithm used	Website	Freely available
ADMET prediction				
ADMET Predictor	Simulations Plus, Inc.	ANN, SVM, Kernel partial least squares (KPLS), and multiple linear regression (MLR)	https://www.simulations-plus.com/software/membraneplus/admet-predictor/	No
StarDrop	Optibrium, Ltd	Pareto optimization, GA	https://www.optibrium.com/stardrop	No
Percepta Platform	Advanced Chemistry Development, Inc.	It provides two predictive algorithms—Classic (based on Hammett-type equation) and GALAS	https://www.acdlabs.com/products/percepta	No
ADMEWORKS Predictor	Fujitsu Kyushu Systems	It is a virtual screening system with a simultaneous evaluation of ADMET properties	https://www.fujitsu.com	No
Sarchitect	Syngene	Bayesian methods, ANN, SVM, decision trees and forests, and other algorithms are used for model building and prediction	https://www.syngeneintl.com	No
QikProp	Schrödinger, Inc.	Monte Carlo statistical mechanics simulation is used to correlate different descriptors to experimental properties	https://www.schrodinger.com/qikprop	No
Derek Nexus	Lhasa, Ltd	Uses the knowledge base of Lhasa for accurate toxicity predictions	https://www.lhasalimited.org/products/derek-nexus.htm	No
PASS	V. N. Orechovich Institute of Biomedical Chemistry under the aegis of the	Uses 20,000 principle compounds from MDDR databases	http://195.178.207.233/PASS/index.html	No

(continued)

Table 12.1 (continued)

Tool	Developer	Brief description and algorithm used	Website	Freely available
	Russian Foundation of Basic Research			
Hazard Expert Pro	CompuDrug, Ltd	A neural network-based approach is used to model the relationship between human cytotoxicity and atomic descriptors	https://www.compudrug.com/hazardexpertpro	No
VolSurf+	Molecular Discovery, Ltd.	It correlates128 molecular descriptors with the pharmacokinetic properties of the drugs	https://www.moldiscovery.com/software/volsurf	No
Bioclipse	Uppsala University, Sweden and European Bioinformatics Institute	It is a Java based workbench, which provides molecular editing and visualization as well as prediction of physio-chemical properties of the molecules	https://sourceforge.net/projects/bioclipse	Yes
MetaDrug	GeneGo, Inc.	It is an Oracle based software, which uses several network-building algorithms like Dijkstra's algorithm to predict ADME/Tox properties of drugs	https://omictools.com/metadrug-tool	No
TIMES	OASIS-LMC	It is a heuristic algorithm used to predict the toxicity of metabolites based on their metabolic maps	http://oasis-lmc.org	No
Molecular visualization				
UCSF Chimera	RBVI, University of California, San Francisco, USA	Provides visualization and analysis of molecular structures and related data.	https://www.cgl.ucsf.edu/chimera	Yes
Jmol	University of Notre Dame, USA	It is a free tool for academicians and	http://jmol.sourceforge.net	Yes

(continued)

Table 12.1 (continued)

Tool	Developer	Brief description and algorithm used	Website	Freely available
		researchers written in Java for 3D visualization of molecules		
PyMOL	Schrödinger Inc.	It is a molecular visualization tool for animating 3D structures	https://pymol.org	No
Swiss-Pdb Viewer (Deep View)	Swiss Institute of Bioinformatics	Active sites of the proteins can be compared and structural alignment can be performed. Several proteins can be analyzed at the same time	https://spdbv.vital-it.ch	Yes
VMD	University of Illinois, USA	It can model, analyze, and visualize biomolecules. It includes multiple sequence alignment and can be used for both sequence and structure data	https://www.ks.uiuc.edu/Research/vmd	Yes

(RASAR), an AI tool, links molecular structures and toxic properties by mining a large database of chemicals (Luechtefeld et al. 2018).

ANNs are further extensively used in the interpretation of analytical data, drug, and dosage form design through bio pharmacy to clinical pharmacy. In pharmaceutical, supervised associating networks are applied as an alternative to conventional response surface methodology (Bourquin et al. 1997; Rojas 2013).

12.9 Applications of ANNs, GAs, and Other ML Algorithms in Drug Discovery

The cost of discovering and developing a drug has since escalated from US$ 800 million in the year 2001 to the current estimated figure of US$ 3 billion (DiMasi et al. 2015). Finding successful new drugs is daunting and predominantly the most difficult part of drug development. ML and other computational technologies are playing a vital role in hunting new drugs quicker, cheaper, and more effective.

ML uses experimental data to optimize clustering or classification of samples or features to develop augment or verify models that can be used to predict the behavior

or properties of systems (Cuperlovic-Culf 2018). In recent years bioinformatics and metabolism analyses have witnessed a variety of ML methods including self-organizing maps, SVM, the kernel machine, Bayesian networks, or fuzzy logic. ML has optimized metabolic network models and their analysis with the availability of enormous genomics and metabolomics data. ML also has been successfully applied for the determination of drug side effects using the data available for human diseases, drugs, and their associated phenotypes to determine metabolically associated side effect predictors (Shaked et al. 2016). Recently, as part of the consortium for metabonomics toxicology (COMET), relational and logic-based ML has successfully utilized to provide causal explanations of rat liver cell responses to toxins using high throughput NMR metabolomics data (Tamaddoni-Nezhad et al. 2006; Chen et al. 2008).

ML techniques prominently in the form of ANNs and GAs have been used in chemoinformatics, computational biology, and pharmaceutical research. The models built with these ML methods are been routinely used for robust external predictions.

Commonly the ANNs are employed in drug design to build predictive models involving:

(a) Hepatotoxicity, genotoxicity profiling.
(b) Pharmacokinetics; clearance, and permeability studies.
(c) Activity predictions against target proteins.
(d) ADME-Tox (absorption, distribution, metabolism, excretion, and toxicity) and biological activity prediction and classification studies.
(e) Lead discovery and similarity, diversity analysis of combinatorial libraries.
(f) HTS data analysis.

ML techniques also find their application in representation, classification, and categorization studies of small lead molecules in drug discovery. In this, the molecules are usually seen as fingerprints featuring its molecules substructures, molecular spaces, bonds, and interatomic distances, represented in the form of binary vectors. At the technical architectural level, these small molecular representations are further searched for their similar substructural and molecular spatial arrangements like compounds. The obtained results are ranked based on their biological activity by the application of various ML prediction filters (Panteleev et al. 2018; Terfloth and Gasteiger 2001).

GAs usually found such applications in the automated generation of small organic molecules. The compounds here are represented as SMILES strings. The electronic, lipophilicity, and shape parameters of these compounds are used to carry out virtual screening studies against the database. A specific GA based search and scoring function then generates a list of similar natural and synthetic compounds as potent leads (Douguet and Thoreau 2000).

There are two major benefits of the application of ML techniques in drug designing:

1. Generation of faster hypotheses overcoming the time and cost factor associated with lengthy wet lab studies;
2. Development of better hypothesis: A well planned and diligently designed ML models may propose better leads than conventional methods curtailing the drug discovery time to 8–10 years period.

GA and ANNs have been instrumental at various levels of lead discovery and lead optimization and are routinely used in various drug discovery tools as follows:

12.9.1 Molecular Docking

Docking is used for finding the preferred pose of a ligand with another molecule to form a stable complex. In the simplest sense, in silico interaction studies between any two molecules are referred to as molecular docking. GA is commonly used in docking algorithms as a favorable search and optimization function. It runs for several generations (>25,000) to generate an optimal binding conformation between a protein and ligand molecule. Some common programs involving GA enabled docking include genetic optimization for ligand docking (GOLD), family competition evolutionary approach (FCEA), and Auto dock (Yang and Kao 2000; Morris et al. 1998)

12.9.2 Pharmacophore Modeling

A pharmacophore may be defined as the essential geometric arrangement of atoms or functional groups necessary to produce a given biological response. A strict IUPAC definition of pharmacophore comes as: "A pharmacophore is the ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target structure and to trigger (or to block) its biological response." (Choudhury and Narahari 2019; Seifert et al. 2007).

This method involves the generation of a quantitative pharmacophore model for a structurally diverse set of input compounds. A suitable hypothesis predicting the biological activity of compounds is developed using training set compounds which is further validated with test compounds. The correlation and regression analysis data obtained with training and test set compounds evaluate the predictive efficacy of developed pharmacophore models. A program named as GA for multiple molecule alignment (GAMMA) is a similar function allowing flexible alignment for multiple small molecules applying Newton optimizer based GA (Terfloth and Gasteiger 2001). GAs have also its relevance in the analogue preparation or automated creation of small molecules by altering the SMILES format along with maintaining the physicochemical descriptors and drug-like standards, i.e., electronic properties, lipophilicity, and conformational features for evaluating scoring function to show better interaction with the respective protein (Douguet and Thoreau 2000).

12.9.3 Quantitative Structure–Activity Relationship (QSAR)

QSAR studies involve the mathematical and statistical analysis of how the chemical structure of a compound is related to its biological activity. The chemical structure of the compound is usually defined as a function of its molecular descriptors such as molecular weight, functional groups, atom, and bond counts to more complex topological, geometric, connectivity, and physicochemical parameters. These descriptors can frequently be calculated by several ML-based popular programs such as MOE, DRAGON, Molconn-Z, ADMET predictor, CODESSA, and PowerMV. Once the descriptors are obtained, statistical modeling methods are employed to derive the correlation and regression between the activity and descriptors. In addition, ANNs are applied to these QSAR models for the prediction of physicochemical and pharmacokinetic properties (Terfloth and Gasteiger 2001).

In the recent era, several pharmaceutical companies and research institutions have come up with novel and intelligently developed ML programs to be used in drug discovery. Two such programs have been developed at Merck pharmaceuticals, namely classification & regression at Merck (CREAM) and Merck online computational chemistry analyzer (MOCCA). CREAM is basically a Python-based modeling tool intended for classification and categorization studies, while MOCCA provides predictive models for toxicity, hepatotoxicity, genotoxicity, ADMET, etc. ANNs play an important role in any such programs by optimizing their training parameters at architectural levels, viz. learning rate, weight decay, batch size, loss-function, signal transfer, and feedbacks, etc.

The importance of any such ANN-based ML methods is that it helps to create newer end-points in the analyzed data and it succeeds to provide better statistical performance compared to conventional methods (Jing et al. 2018). Another example of an ML-based predictive model has been developed by researchers at IBM. The model at IBM helps to predict, represent, and identify which diseases are typically linked to which prominent side effects. The graphical representation included orange and blue dots. An orange dot represents the regions of diseases, while blue dot representations imply the specific side effects associated with them. The IBM models have greatly helped to identify and limit the typical side effects associated with specific diseases.

Recently a newer concept has been proposed by Burden et al., whereby neural networks inversion problem for QSAR studies of dihydrofolate reductase inhibitor was solved by applying GA. A small layered feedforward/back-propagation neural network-based QSAR model was developed and maximum activity on the structure-activity surface was determined using a GA. Such development of hybrid algorithms involving both ANN and GA in ML techniques proves to be a milestone in fast prediction and better evaluation studies in the domain of drug discovery. Proper knowledge and diligent application of each of these ML methods may certainly unveil newer dimensions in multi-millionaire pharmaceutical industrial setups in the near future.

12.10 Conclusions

ML algorithms are playing a game changer role in many sectors. ML approaches are useful in designing and development of new biologically active molecules with desired properties for a drug target. In healthcare sectors, many renowned pharmaceutical companies have made a huge investment in AI companies working on ML as a joint venture to come out with healthcare tools and better drugs. With the rapid advancement in computer processing power and huge datasets availability and synthesis planning, ML algorithms have helped in speeding up drug development. In the near future, ML concepts will permanently change the pharmaceutical industry and the way drugs are discovered.

Competing Interest The authors declare that there are no competing interests.

References

Arif JM, Siddiqui MH, Akhtar S, Al-Sagair O (2013) Exploitation of in silico potential in prediction, validation and elucidation of mechanism of anti-angiogenesis by novel compounds: comparative correlation between wet lab and in silico data. Int J Bioinforma Res Appl 965:336–348

Bakkar N, Kovalik T, Lorenzini I, Spangler S, Lacoste A, Sponaugle K, Bowser R (2018) Artificial intelligence in neurodegenerative disease research: use of IBM Watson to identify additional RNA-binding proteins altered in amyotrophic lateral sclerosis. Acta Neuropathol 135 (2):227–247

Bourquin J, Schmidli H, van Hoogevest P, Leuenberger H (1997) Basic concepts of artificial neural networks (ANN) modeling in the application to pharmaceutical development. Pharm Dev Technol 2(2):95–109

Chen J, Muggleton S, Santos J (2008) Learning probabilistic logic models from probabilistic examples. Mach Learn 73(1):55–85

Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T (2018) The rise of deep learning in drug discovery. Drug Discov Today 23(6):1241–1250

Choudhury C, Narahari SG (2019) Pharmacophore modelling and screening: concepts, recent developments and applications in rational drug design. In: Mohan C (ed) Structural bioinformatics: applications in preclinical drug discovery process. Challenges and advances in computational chemistry and physics. Springer, Cham, pp 25–53

Cuperlovic-Culf M (2018) Machine learning methods for analysis of metabolic data and metabolic pathway modeling. Metabolites 8(1):4

DiMasi JA, Grabowski HG, Hansen RW (2015) The cost of drug development. N Engl J Med 372 (20):1972

Douguet DE, Thoreau GG (2000) A genetic algorithm for the automated generation of small organic molecules: drug design using an evolutionary algorithm. J Comput Aided Mol Des 14:449–466

Doyle OM, Mehta MA, Brammer MJ (2015) The role of machine learning in neuroimaging for drug discovery and development. Psychopharmacology 232:4179–4189

Durrant JD, McCammon JA (2010) NNScore: a neural-network-based scoring function for the characterization of protein− ligand complexes. J Chem Inf Model 50(10):1865–1871

Duvenaud DK, Maclaurin D, Iparraguirre J, Bombarell R, Hirzel T, Aspuru-Guzik A, Adams RP (2015) Convolutional networks on graphs for learning molecular fingerprints. In: Advances in neural information processing systems. Curran Associates, Red Hook, pp 2224–2232

Ekins S (2016) The next era: deep learning in pharmaceutical research. Pharm Res 33 (11):2594–2603

Ferrero E, Dunham I, Sanseau P (2017) *In silico* prediction of novel therapeutic targets using gene–disease association data. J Transl Med 15(1):182

Fox T, Kriegl JM (2006) Machine learning techniques for in silico modeling of drug metabolism. Curr Top Med Chem 6(15):1579–1591

Gibson E, Li W, Sudre C, Fidon L, Shakir DI, Wang G, Eaton-Rosen Z, Gray T, Doel R, Hu Y, Whyntie T, Nachev P, Modat M, Barratt DC, Ourselin S, Cardoso MJ, Vercauteren T (2018) NiftyNet: a deep-learning platform for medical imaging. Comput Methods Prog Biomed 158:113–122

Goswami M, Akhtar S, Osama K (2018) Strategies for monitoring and modeling growth of hairy root cultures: an in silico perspective. In: Srivastava V, Mehrotra S, Mishra S (eds) Hairy roots. Springer, Singapore

Gupta MK, Agarwal K, Prakash N, Singh DB, Misra K (2012) Prediction of miRNA in HIV-1 genome and its targets through artificial neural network: a bioinformatics approach. Netw Model Anal Health Inf Bioinf 1:141–151

Gupta CL, Akhtar S, Bajpai P (2014) *In silico* protein modeling: possibilities and limitations. EXCLI J 13:513–515

Hessler G, Baringhaus KH (2018) Artificial intelligence in drug design. Molecules 23(10):2520

Huang G, Li J, Wang P, Li W (2017) A review of computational drug repositioning approaches. Comb Chem High Throughput Screen 20:831. https://doi.org/10.2174/1386207321666171221112835

Huang G, Yan F, Tan D (2018) A review of computational methods for predicting drug targets. Curr Protein Pept Sci 19(6):562–572

Jing Y, Bian Y, Hu Z, Wang L, Xie X (2018) Deep learning for drug design: an artificial intelligence paradigm for drug discovery in the big data era. AAPS J 20(3):58

Kapetanovic IM (2008) Computer-aided drug discovery and development (CADDD): in-silico-chemico-biological approach. Chem Biol Interact 171:165–176

Lavecchia A (2015) Machine- learning approaches in drug discovery: methods and applications. Drug Discov Today 20(3):318–331

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444

Lee EJ, Kim YH, Kim N, Kang DW (2017) Deep into the brain: artificial intelligence in stroke imaging. J Stroke 19(3):277

Leelananda SP, Lindert S (2016) A review of computational methods for predicting drug targets. Beilstein J Org Chem 12:694–2718

Lo YC, Rensi SE, Torng W, Altman RB (2018) Machine learning in chemoinformatics and drug discovery. Drug Discov Today 23(8):1538–1546

Luechtefeld T, Marsh D, Rowlands C, Hartung T (2018) Machine learning of toxicological big data enables read-across structure activity relationships (RASAR) outperforming animal test reproducibility. Toxicol Sci 165(1):98–212

Ma J, Sheridan RP, Liaw A, Dahl GE, Svetnik V (2015) Deep neural nets as a method for quantitative structure–activity relationships. J Chem Inf Model 55(2):263–274

Maltarollo VG, Kronenberger T, Espinoza GZ, Oliveira PR, Honorio KM (2019) Advances with SVM for novel drug discovery. Expert Opin Drug Discovery 14(1):23–33

Mandal AK, Johnson C, Wu F, Bornemeier D (2007) Identifying promisingcompounds in drug discovery: genetic algorithms and some new statistical techniques. J Chem Inf Model 47 (3):81–988

Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, Olson AJ (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. J Comput Chem 19:1639–1662

Myint KZ, Wang L, Tong Q, Xie XQ (2012) Molecular fingerprint-based artificial neural networks QSAR for ligand biological activity predictions. Mol Pharm 9(10):2912–2923

Niculescu SP (2003) Artificial neural networks and genetic algorithms in QSAR. J Mol Struct 622 (1–2):71–83
Oquendo MA, Baca-Garcia E, Artes-Rodriguez A, Perez-Cruz F, Galfalvy HC, Blasco-Fontecilla H, Madigan D, Duan N (2012) Machine learning and data mining: strategies for hypothesis generation. Mol Psychiatry 17(10):956–959
Panteleev J, Gao H, Jia L (2018) Recent applications of machine learning in medicinal chemistry. Bioorg Med Chem Lett 28(17):2807–2815
Patra TK, Meenakshisundaram V, Hung JH, Simmons DS (2017) Neural-network-biased genetic algorithms for materials design: evolutionary algorithms that learn. ACS Comb Sci 19 (2):96–107
Pu L, Naderi M, Liu T, Wu H, Mukhopadhyay S, Brylinski M (2019) eToxPred: a machine learning-based approach to estimate the toxicity of drug candidates. BMC Pharmacol Toxicol 20(2):1–15
Rashid MBMA, Chow EK (2019) Artificial intelligence-driven designer drug combinations: from drug development to personalized medicine. SLAS Technol 24(1):124–125
Rashid J, Hodson H (2017) Google DeepMind and healthcare in an age of algorithms. Health Technol 7(4):351–367
Rodrigues T, Werner M, Roth J, da Cruz EH, Marques MC, Akkapeddi P, Werz O (2018) Machine intelligence decrypts β-lapachone as an allosteric 5-lipoxygenase inhibitor. Chem Sci 9 (34):6899–6903
Rojas R (2013) Neural networks: a systematic introduction. Springer-Verlag, Berlin
Sayeed U, Wadhwa G, Jamal QMS, Kamal MA, Akhtar S, Siddiqui MH, Khan MS (2016) MHC binding peptides for designing of vaccines against Japanese encephalitis virus: a computational approach. Saudi J Biol Sci 25(8):1546–1551
Schneider G, Bohm HJ (2002) Virtual screening and fast automated docking methods. Drug Discov Today 7:64–70
Schneider P, Schneider G (2016) De novo design at the edge of chaos: miniperspective. J Med Chem 59:4077–4086
Searls DB (2005) Data integration: challenges for drug discovery. Nat Rev Drug Discov 4(1):45
Segler MH, Kogej T, Tyrchan C, Waller MP (2017) Generating focused molecule libraries for drug discovery with recurrent neural networks. ACS Cent Sci 4(1):120–131
Segler MH, Preuss M, Waller MP (2018) Planning chemical syntheses with deep neural networks and symbolic AI. Nature 555(7698):604
Seifert MH, Kraus J, Kramer B (2007) Virtual high-throughput screening of molecular databases. Curr Opin Drug Discov Devel 10(3):298–307
Shaked I, Oberhardt MA, Atias N, Sharan R, Ruppin E (2016) Metabolic network prediction of drug side effects. Cell Syst 2(3):209–213
Tamaddoni-Nezhad AR, Kakas CA, Muggleton S (2006) Application of abductive ILP to learning metabolic network inhibition from temporal data. Mach Learn 64(1–3):209–230
Terfloth L, Gasteiger J (2001) Neural networks and genetic algorithms in drug design. Drug Discov Today 6(20):102–108
Varnek A, Baskin II (2011) Chemoinformatics as a theoretical chemistry discipline. Mol Inf 30 (1):20–32
Yang JM, Kao CY (2000) Flexible ligand docking using a robust evolutionary algorithm. J Comput Chem 21:988–998
Yang H, An Z, Zhou H, Hou Y (2018) Application of machine learning methods in bioinformatics. AIP Conf Proc 1967:040015. https://doi.org/10.1063/1.5039089
Yosipof A, Guedes RC, Garcia-Sosa AT (2018) Data mining and machine learning models for predicting drug likeliness and their disease or organ category. Front Chem 6:162
Zhong F, Xing J, Li X, Liu X, Fu Z, Xiong Z, Lu D, Wu X, Zhao J, Tan X, Li F, Luo X, Li K, Chen Z, Zheng M, Jiang H (2018) Artificial intelligence in drug design. Sci China Life Sci 61 (10):1191–1204