SCIENCE CHINA
Life Sciences

FROM CAS & CAE MEMBERS
•REVIEW•

October 2018 Vol.61 No.10: 1191–1204
https://doi.org/10.1007/s11427-018-9342-2

Artificial intelligence in drug design

Feisheng Zhong[1,2†], Jing Xing[1,2†], Xutong Li[1,2], Xiaohong Liu[1,3], Zunyun Fu[1,2], Zhaoping Xiong[1,3], Dong Lu[1,2], Xiaolong Wu[1,2], Jihui Zhao[1,2], Xiaoqin Tan[1,2], Fei Li[1,4], Xiaomin Luo[1], Zhaojun Li[5], Kaixian Chen[1,3], Mingyue Zheng[1*] & Hualiang Jiang[1,3*]

[1]*Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China;*
[2]*School of Pharmacy, University of Chinese Academy of Sciences, Beijing 100049, China;*
[3]*School of Life Science and Technology, ShanghaiTech University, Shanghai 200031, China;*
[4]*Department of Chemistry, College of Sciences, Shanghai University, Shanghai 200444, China;*
[5]*School of Information Management, Dezhou University, Dezhou 253023, China*

Received April 24, 2018; accepted May 22, 2018; published online July 18, 2018

Thanks to the fast improvement of the computing power and the rapid development of the computational chemistry and biology, the computer-aided drug design techniques have been successfully applied in almost every stage of the drug discovery and development pipeline to speed up the process of research and reduce the cost and risk related to preclinical and clinical trials. Owing to the development of machine learning theory and the accumulation of pharmacological data, the artificial intelligence (AI) technology, as a powerful data mining tool, has cut a figure in various fields of the drug design, such as virtual screening, activity scoring, quantitative structure-activity relationship (QSAR) analysis, *de novo* drug design, and *in silico* evaluation of absorption, distribution, metabolism, excretion and toxicity (ADME/T) properties. Although it is still challenging to provide a physical explanation of the AI-based models, it indeed has been acting as a great power to help manipulating the drug discovery through the versatile frameworks. Recently, due to the strong generalization ability and powerful feature extraction capability, deep learning methods have been employed in predicting the molecular properties as well as generating the desired molecules, which will further promote the application of AI technologies in the field of drug design.

drug design, artificial intelligence, deep learning, QSAR, ADME/T

Citation: Zhong, F., Xing, J., Li, X., Liu, X., Fu, Z., Xiong, Z., Lu, D., Wu, X., Zhao, J., Tan, X., et al. (2018). Artificial intelligence in drug design. Sci China Life Sci 61, 1191–1204. https://doi.org/10.1007/s11427-018-9342-2

INTRODUCTION

The process of drug research and development includes drug target identification, target validation, hit to lead generation, lead optimization, the preclinical candidate identification, preclinical study and clinical study (Vohora and Singh, 2017). To develop a novel prescription drug, the mean pre-tax expenditure is approximately 2.558 billion USD (DiMasi et al., 2016) and it takes about 10–15 years (Turner, 2010). However, given the high investment, the estimated clinical approval success rate of innovative small molecules during the drug discovery and development process is still only 13%, with a relatively high risk of failure eventually. The development of computer-assisted drug design technique is among the most ambitious to change this thorny situation based on the rational guidance to the process (Hassan Baig et al., 2015). The drug discovery process and the corresponding computer-assisted drug design methods can be found in the book *Computer-Assisted Drug Design* (Mason, 2007). The

†Contributed equally to this work
*Corresponding authors (Hualiang Jiang, email: hljiang@simm.ac.cn; Mingyue Zheng, email: myzheng@simm.ac.cn)

© Science China Press and Springer-Verlag GmbH Germany, part of Springer Nature 2018

computational methods not only guarantee a systematical assessment of the molecular characteristics (e.g., bioactivity, selectivity, side effects, physicochemical properties, absorption, distribution, metabolism and excretion) at the theoretical level, but also generate lead molecules with favorable properties *in silico*. In addition, the computational methods with multi-objective optimization can be used to decrease the attrition rate of the preclinical candidate compounds. In the field of drug design, artificial intelligence (AI) refers to the application of algorithms that analyze, learn and explain pharmaceutical related big data to discover new drugs, incorporating the development of machine learning in a more integrated and automatic manner (Duch et al., 2007). Due to the development of machine learning methods and the accumulation of chemical and pharmacological data, the AI technology has cut a figure in the field of drug design as a data-driven computational approach. Compared with traditional methods, machine learning based methods, as a branch of AI, do not rely on the theoretical advancement of the complicated physical and chemical concrete principles, but put more focus on the transformation of enormous biomedical big data into new insight and reusable knowledge. Common algorithms for machine learning include logistic regression (LR), naive Bayesian classification (NBC), *k* nearest neighbor (KNN), multiple linear regression (MLR), support vector machine (SVM), probabilistic neural network (PNN), binary kernel discrimination (BKD), linear discriminant analysis (LDA), random forest (RF), artificial neural network (ANN), partial least-squares (PLS), principal component analysis (PCA), and so on (Lavecchia and Giovanni, 2013; Melville et al., 2009). Recently, AI technologies, especially the deep learning models, show great prospects in drug design owing to their powerful generalization and feature extraction capability. Traditional machine learning methods use manually designed features, while the deep learning methods can automatically learn features from the input data, which can transform simple features into complex features through multi-layer feature extraction. In addition, the deep learning methods usually have less generalization errors than the traditional machine learning methods, which helps them getting more satisfactory results on some benchmark or competitive tests. For example, George Dahl's team won the Merck Molecular Activity Challenge by applying the AI technology especially the deep learning method (Ma et al., 2015). Due to the above advantages, the deep learning method as a data mining method has shown great promise in the field of drug design. The deep learning methods mainly consist of deep neural network (DNN), convolutional neural network (CNN), recurrent neural network (RNN), autoencoder, restricted Boltzmann machine (RBN). A quick overview of deep learning algorithms can be found elsewhere (Angermueller et al., 2016; LeCun et al., 2015; Schmidhuber, 2015), with a more detailed introduction to deep learning techniques in the book *Deep Learning* (Goodfellow et al., 2016).

This review introduces the AI methods involved in the field of drug design, and gives a special focus on the application of deep learning methods in the discovery and development of new drugs. The drug discovery, drug design topics and AI models are listed in Figure 1. In addition, this review also introduces machine learning strategies related to drug design scenarios including methods of the molecular representation, transfer learning for low data, the cross-validation method and the skills of training the deep neural networks. Finally, this review summarizes the applications of AI in the field of drug design and gives a prospective to the future of AI in drug discovery and development.

THE APPLICATION OF AI IN DRUG DESIGN

Protein structure and function

Prediction of protein folding from sequence

Most diseases are related to protein dysfunctions. By studying the structures of proteins, the structure-based drug design strategies can be used to discover the active small molecules towards the protein targets. However, measuring the three-dimensional (3D) structures of the proteins will cost a lot of time and money at present, and it is meaningful to develop algorithms to predict the 3D structure of a protein. Although the sequence information of most proteins is available, it is still an unresolved problem to make accurate *de novo* prediction of their 3D structures. Recently, owing to the powerful capability of feature extraction, deep learning technologies have been applied to predict the secondary structure (Spencer et al., 2015), backbone torsion angle (Li et al., 2017) and residue contacts of proteins (Wang et al., 2017). For example, the deep learning method by combining one-dimensional (1D) with two-dimensional (2D) CNN to predict the residue contacts outperformed others in CASP12 (12th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction) (Schaarschmidt et al., 2017; Wang et al., 2017). The architecture of deep learning may accurately learn the relationship between the sequence and the structure through feature extraction. Currently, it is still a distant goal to precisely predict 3D structures of proteins, and the deep learning method has shown great promise on promoting the development in this field.

Prediction of protein-protein interactions

The protein-protein interactions (PPIs) are critical for many biological processes and related to many diseases (Falchi et al., 2014; Scott et al., 2016). A PPIs database—the String database, contains about 1.4 billion PPIs obtained by both experimental and bioinformatics methods (Szklarczyk et al.,

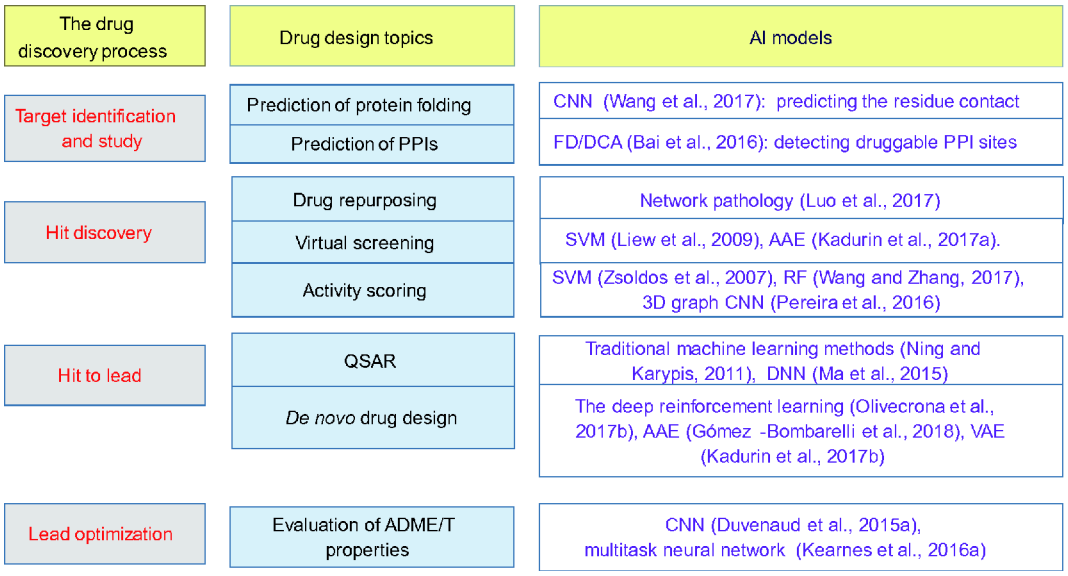

Figure 1 The drug discovery, drug design topics and AI models.

2015).

The PPI interface is defined as the protein-protein binding sites composed of many residues (Cukuroglu et al., 2014). It can become a new class of the drug targets which are different from the traditional drug targets such as G-protein coupled receptors (GPCRs), ion channels, kinases, and nuclear receptors (Higueruelo et al., 2013; Santos et al., 2017). For example, there are 1,756 non-peptide inhibitors among 18 families of PPIs reported in the iPPI-DB (inhibitors of protein-protein Database) (Labbé et al., 2016). As a new class of targets, PPIs will extend the target space and promote the development of the small molecule drugs (Shin et al., 2017). Compared with the traditional methods, targeting PPIs may reduce the adverse effects because it can increase the biological selectivity of regulatory effects (Valkov et al., 2012). For instance, compound DC_AC50 can block copper ion transport within cells by binding with the copper-transfer interfaces, and inhibit specifically tumor cell proliferation without affecting normal somatic cell survival at the same time (Wang et al., 2015).

To achieve the idea of the drug design based on the structure of protein-protein complex, it is important to study the interface of PPI. Unfortunately, in most cases the exact PPI information is limited (Xue et al., 2015), which gives rise to many computational methods for predicting the interface of PPI. The method based on the template is easier and more reliable owing to the conservation of PPI interfaces (Zhang et al., 2010). For example, eFindSite (Maheshwari and Brylinski, 2016), a web server for PPI interfaces prediction, uses template-based, residue-based and sequence-based features to develop SVM and NBC models. Based on the principle of the complementarity, the protein-protein docking methods (e.g., ZDOCK (Chen et al., 2003) and SymmDock (Schneidman-Duhovny et al., 2005)) can be used to predict the interface of PPI when the structure of two interactive proteins is available (Vakser, 2014). Among these methods, the challenging issue is how to predict the conformational change when two unbound proteins become a combined one. Deep learning methods can extract the most relevant sequence features to predict PPI interfaces, which shows an obvious improvement compared to other machine learning methods such as SVM (Du et al., 2016).

Considering the large buried area (1500–3000 Å^2) of the interface (Scott et al., 2016), it is necessary to search for the druggable sites or local region in the interface. The hot spots may be the druggable sites because it contributes to a large amount of binding free energy (Cukuroglu et al., 2014). Bai et al. used fragment docking and direct coupling analysis (FD-DCA) to detect the druggable PPI sites (Bai et al., 2016). They firstly developed a fragment docking tool called iFitDock, which can be used to seek the druggable hot spots in PPI interfaces. Then, the small hot spots were clustered to form candidate binding sites. Finally, the scoring function based on the evolutionary conservative level was employed to find the best protein-protein binding sites. Altogether, the hot spots in the PPI interface are promising drug targets and it is meaningful to develop computational approaches for identifying the hot spots and designing small modulators targeting PPI interfaces.

Hit discovery

Drug repurposing

Drug repurposing, also called drug repositioning, is defined as the process to find novel indications of the approved drugs (Ashburn and Thor, 2004; Lotfi Shahreza et al., 2017), which

can reduce the time and risk of drug development (Ashburn and Thor, 2004). The drug repurposing is feasible because most drugs may have multiple targets (Klaeger et al., 2017) and the targets may correspond to multiple effects, which is showing the high diversity of drug-disease relationship. For example, Metformin, which was approved to treat the type 2 diabetes, may extend lifespan (Cabreiro et al., 2013; De Haes et al., 2014; Martin-Montalvo et al., 2013).

Drugs and diseases are two core elements to repurpose a drug. There are other elements related to the drug repurposing, such as drug targets and disease gene. Due to the diversity of the interaction, the network analysis can be used to depict the interactions of these elements (Lotfi Shahreza et al., 2017). In terms of drug design, there are nine kinds of important networks: gene regulatory networks, metabolic networks, protein-protein networks, drug-target networks, drug-drug networks, drug-disease networks, target-disease networks, drug-adverse effect networks and disease-disease networks (Lotfi Shahreza et al., 2017). The basic hypothesis of the network-based method is that the similar drugs often have the similar targets or effects (Yamanishi et al., 2008). The information of the individual network is limited and partial, thus it is necessary to integrate multiple networks to form the heterogeneous network for repurposing one drug. In particular, it is important to combine drug repurposing with the drug target prediction because the target can be regarded as a bridge from the drug to the disease (Wang et al., 2014). DTINet, a heterogeneous network integrating the information of multiple networks through the network diffusion algorithm and the dimensionality reduction approach, was used to predict the new target and indications (Luo et al., 2017). For example, this method suggested that telmisartan, alendronate and chlorpropamide might have novel cyclooxygenase inhibitory effect. These effects were later verified experimentally by evaluating the expression of proinflammatory factors, and these three drugs thus provide high quality hits for inflammation prevention.

Virtual screening

Virtual screening refers to the application of algorithm and software to find bioactive molecules (hits) from in-house compound collections or commercial chemical libraries, which provides a highly efficient approach to discover novel hits and filter out compounds with unfavorable scaffolds in early drug development (Lavecchia and Giovanni, 2013). Virtual screening methods include docking-based, pharmacophore-based (Kim et al., 2010), similarity searching (Willett, 2006) and machine learning methods (Leelananda and Lindert, 2016). In general, these methods can be classified into two types: structure-based virtual screening and ligand-based virtual screening. Among them, molecular docking has been widely used when the 3D structure of a target protein is available (Chen, 2015). Despite the fact that many successful applications of docking-based virtual screening have been constructed (Talele et al., 2010), there are still obvious limitation of this method. For example, scoring function of docking cannot predict binding affinities accurately because of inadequate consideration of solvation and entropic effects (Huang and Zou, 2010), and the protein flexibility makes the problem even more complicated (Chen, 2015). Moreover, since most docking methods only consider binding affinities and ignore the other parameters such as the residence time (Copeland, 2010), the docking score is not an ideal index for drug efficacy and the false positive rate of the docking-based VS is high (Chen, 2015; Xing et al., 2017).

Unlike the docking-based virtual screening methods, the ligand-based virtual screening methods do not rely on the 3D structural information of the proteins. It tries to map the molecular features (descriptors) to bioactivity classes (Lavecchia and Giovanni, 2013). In this respect, machine learning methods such as SVM have been frequently used for virtual screening (Leelananda and Lindert, 2016; Liew et al., 2009; Melville et al., 2009), which has shown high yields (ratio of predicted known hits) and decreased false-hit rates simultaneously (false hit in predicted hits)(Ma et al., 2009). Recently, deep learning methods have been applied in the virtual screening owing to its fantastic classification capacity, powerful feature extraction ability and low generalization error (LeCun et al., 2015; Unterthiner et al., 2014). For example, sparse distribution of the active compounds in the general database generally wastes a lot of search time at virtual screening (Ma et al., 2008; Segler et al., 2018). To address this issue, a long short-term memory network model based on the similarity between natural language and the simplified molecular input line entry specification (SMILES) was established to generate focused molecule libraries with molecules similar to the training molecules (Unterthiner et al., 2014). The new molecular libraries generated by RNN can be screened with machine learning methods such as DNN and gradient boosting trees. Besides, owing to the powerful generative ability, an adversarial autoencoder (AAE) model was trained based on the NCI-60 cell line assay data (Shoemaker, 2006), which can be applied to generate molecular fingerprints for searching potential anticancer agents (Kadurin et al., 2017a).

Activity scoring

As mentioned above, the core component of molecular docking is the scoring function, which is designed to evaluate the binding affinities of the drug-like molecules towards a target of interest (Huang et al., 2010). Owing to the strong nonlinear mapping ability, machine-learning based scores exhibit better performance by extracting various features effectively, such as the geometric features, chemical features and physical force field features (Khamis et al., 2015). These scores can be considered as data-driven black box models

since they predict the binding affinity or ligand-protein binding interaction from experimental data directly and avoid the study of the complex physical function related to the docking (Ain et al., 2015). Machine learning techniques such as RF and SVM can be used to improve the performance of scoring function. For example, instead of using the linear additive assumption of energy terms, an SVM model described the nonlinear relationship between the individual energy terms derived from the docking program eHiTS and experimental binding affinity data showed improved screening power and scoring power (Kinnings et al., 2011; Zsoldos et al., 2007). Wang and Zhang reported a Δ_{vina}RF parameterization correction method combining the RF with AutoDock scoring function (Wang and Zhang, 2017), which showed an excellent performance compared with GlideScore XP (Repasky et al., 2007). Recently, due to the good performance of CNN in the field of image processing (LeCun et al., 2015), some researchers have attempted to use CNN to extract the features from protein-ligand interactions image so as to predict the protein-ligand affinity. Jimenez et al. employed a 3D graph CNN model to predict ligand-protein binding affinities (Jimenez et al., 2018), which showed that the predictive binding affinities had good correlation with experimental data in the datasets. The real power of deep learning lies in its ability to learn complex and abstract features from basic and primitive features. Therefore, it is important to depict the basic features of the compound-protein complex such as the atom types, atom charge, atom distance and amino types (LeCun et al., 2015). Deep VS, a framework based on CNN, can learn the abstract features from the basic features (the atom context) and it outperformed the traditional docking programs such as ICM (Abagyan et al., 1994) and GLIDE SP (Friesner et al., 2004) on the Directory of Useful Decoys (DUD) in terms of area under the curve of receiver operating characteristic (AUC-ROC) and enrichment factor (Pereira et al., 2016). In principle, the CNN method predicts the binding affinities by extracting the features in the protein-ligand interaction image, which is more akin to a knowledge-based scoring function but with enhanced predictive capability.

Hit-to-lead optimization

QSAR

During the process of hit-to-lead optimization, QSAR analysis can be used to find the potent leading compounds from a series of hits analogues by predicting bioactivity of the analogues. QSAR mainly refers to the use of mathematical methods for studying the quantitative mapping between the structural or physicochemical properties of compounds and their associated biological activities (Esposito et al., 2004). QSAR analysis mainly involves data collection, selection and generation of molecular descriptions, establishment of mathematical models, assessment and interpretation of models, and application of models (Myint and Xie, 2010). Among them, the key issues are the representation of the chemical structure and the mathematical model reflecting QSAR. After selecting the descriptors, it is necessary to find a suitable mathematical model to fit the structure-activity relationship. In 1964, Hansch et al. proposed the well-known Hansch equation which innovatively used linear regression models and physicochemical descriptors (the hydrophobic parameter, the electronic parameter and the steric parameter) to describe the 2D structure-activity relationship, opening the chapter for QSAR study (Hansch and Fujita, 1964). In the same year, Free et al. formulated the Free-Wilson method to describe the relationship between the chemical structure and bioactivity based on hypothesis that the contribution of substituents to the activity of the compound is additive (Free and Wilson, 1964). Compared with Hansch method, Free-Wilson method does not need the physiochemical parameters and it can directly predict the bioactivity from the chemical structure by encoding the chemical structure. With the development of machine learning techniques, various methods have later been used to construct mathematical models (Dobchev et al., 2014; Dudek et al., 2006; Ning and Karypis, 2011), such as RF and SVM. Recently, deep learning methods have been introduced to QSAR modeling owing to the ability of dealing with diverse chemical characters and the merit of extracting features automatically. George Dahl's team won the Merck Molecular Activity Challenge (a Kaggle competition about QSAR problems) in 2012 by the ensemble models consisted of the multi-task DNN, Gaussian progress regression and gradient boosting machine (Ma et al., 2015). Inspired by the Kaggle competition results, Dahl et al. continued to systematically study the multi-task DNN and his results had shown that the multi-task DNN outperformed the single-task neural network because the multi-task method may learn general features by sharing parameters of different but related tasks (Dahl et al., 2014). Ramsundar et al. integrated multi-task neural networks into the DeepChem platform, which eased the use of multi-task neural networks algorithms for drug development (Ramsundar et al., 2017). They also evaluated the performance and discovered that the multi-task deep networks were very robust and outperformed random forests on various tasks. Subramanian et al. employed the DNN with Canvas descriptors to build the classification and regression model to predict the binding affinities of the human β-secretase 1 (hBACE-1) inhibitors (Subramanian et al., 2016). On the validation set, this DNN model yielded good classification ability with an accuracy of 0.82 and it also exhibited favorable regression ability with the coefficient of determination (R^2) of 0.74 and mean absolute error (MAE) of 0.52. In addition, their results have shown that the DNN model with 2D descriptors performed better than the force-field-based

methods (e.g., CoMFA), which is partially due to the powerful generalization capabilities of deep learning models. Clearly, deep learning based QSAR models with improved activity prediction performance will play a more important role in the future hit-to-lead optimization studies.

Generative models for de novo design

De novo drug design means designing new chemical entities to modulate the target of interest (Hartenfeller and Schneider, 2011). The traditional *de novo* design method such as the fragment-based approach can generate new molecules from scratch. However, many of them are difficult to synthesize due to the complexity and impracticality of the molecular structure (Schneider et al., 2017). In addition, it is hard to evaluate their bioactivity because of the shortcomings of scoring functions mentioned previously.

Owing to the strong generative ability and learning capability, deep learning methods have been used to automatically generate new structures with some desired properties (Mullard, 2017). Olivecrona et al. developed the deep reinforcement learning method to tune the RNN to generate the molecules with predicted biological activity (Olivecrona et al., 2017). The SMILES structures of molecules collected from ChEMBL were used to train the RNN for getting the syntax of SMILES, and the RNN can generate the molecules by sampling from the conditional probability distribution of the training set. In reinforcement learning, Agents are the decision makers who take actions under certain states. If an Agent's action leading to a positive reward, the Agent's trend of generating this action will be enhanced (Mnih et al., 2015). SVM was utilized to upgrade the action policy for obtaining the high expected reward by activity scoring based on the ligands in the training set. When the RNN with the deep reinforcement learning model was employed to generate molecules against dopamine receptor type 2, more than 95% of the structures were predicted to be bioactive via the SVM scoring function.

Another application of generation models with deep learning is the use of autoencoders to generate novel molecules. Gómez-Bombarelli et al. combined the variational autoencoder (VAE) and the multilayer perceptron (MLP) to generate new compounds with desired properties automatically (Gómez-Bombarelli et al., 2018). The network consisted of three parts: the encoder, the decoder and the predictor. The encoder transforms discrete SMILES strings into continuous vectors in latent space, and the decoder can make these vectors back to the discrete SMILES strings. The MLP is used to predict the property of the molecules, and the gradient-based optimization can be employed to find the continuous vectors with high predictive value of the property. Owing to the continuity of the vector representation in the latent space, the gradient-based optimization combined with the Bayesian inference can be utilized to quickly identify the molecules with desirable properties. The model has the advantage that it can produce a human-understandable chemical structure with higher predictive activity automatically. However, it also generated many cases of invalid SMILES that do not correspond to valid chemical structures. To overcome this difficulty, Pu et al. utilized the grammar variational autoencoder to make the output more effective by defined SMILES syntax (Pu et al., 2017). More recently, Kadurin et al. introduced an AAE model named druGAN to generate molecular fingerprints, which outperformed the VAE model in terms of the reconstruction error, generation ability and feature extraction capability (Kadurin et al., 2017b).

To evaluate whether a generated molecule is synthetically accessible, Coley et al. defined a synthetic complexity metrics by training a neural network model based on a reaction database (Coley et al., 2018). The principle for scoring synthetic complexity is that the synthetic reaction is a process which will increase complexity of the reactant. For a synthetic reaction, it means that the complexity score of product should be greater than the reactant. Therefore, Coley et al. encoded a chemical reaction into several (reactant, product) pairs and attempted to construct a scoring function for describing the inequality relationship between reactant complexity and product complexity (Figure 2). Since the neural networks have powerful function approximation ability (Andras, 2017), Coley et al. employed 22 million (reactant, product) pairs to train the neural network for learning the scoring function, with their results demonstrating that the learned function (SCScore) could well describe the complexity increase in the synthesis process. This model can not only guide chemists in performing the inverse synthetic analysis but also help them eliminate unrealistic molecules in drug design by evaluating the synthetic complexity.

In silico evaluation of ADME/T properties

Physical and chemical properties

Early identification of molecules with poor physical or chemical properties in a drug discovery pipeline significantly reduces the risk of failure. Many deep learning based approaches have been developed on this topic (Lusci et al., 2013). Duvenaud et al. utilized the CNN-ANN to predict the solubility by extracting information directly from the molecular graph with a good predictive performance consequence (MAE is 0.53±0.07) (Duvenaud et al., 2015). The highlight of this method lies in its interpretability. For example, the fragments contributing to molecule solubility such as hydrophilic R-OH group can be obtained by backtracking the model. Inspired by Duvenaud's work, Coley et al. employed a tensor-based convolutional embedding of attributed molecular graphs method to predict the molecular aqueous so-

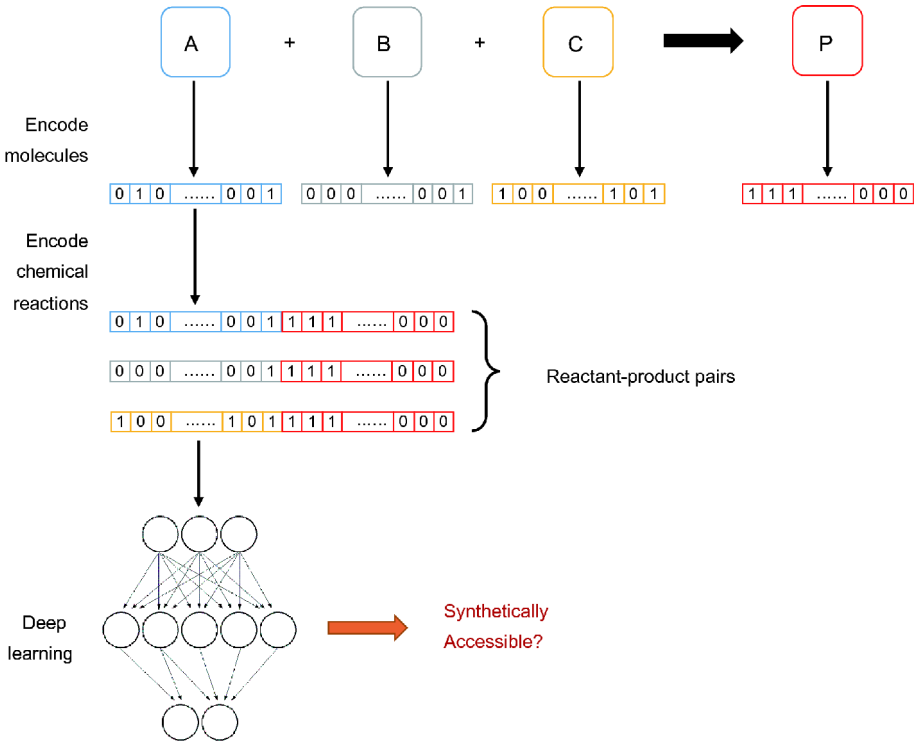

Figure 2 The encoding of the chemical reaction. A, B and C represent the reactants. P represents the main product.

lubility, which outperformed Duvenaud's model (MAE is 0.424±0.005) (Coley et al., 2017). The model employed a molecular tensor integrating the bond-level and atom-level features to describe attributed molecular graph. Compared with the Duvenaud's model, Coley's model used more atom-level information to predict the molecular aqueous solubility.

Since a good correlation was found between oral drug absorption and Caco-2 permeability coefficient (P_{app}) (Artursson and Karlsson, 1991; Hubatsch et al., 2007), predicting the candidate drug P_{app} plays an important role to evaluate the pharmacokinetics properties of candidate drugs. Wang et al. collected 1,272 compounds with Caco-2 permeability data and used Boosting, SVM regression, PLS, and MLR to construct the prediction models with 30 descriptors (Wang et al., 2016). The Boosting model showed the best results with excellent predictive ability (R^2=0.81, root mean square error (RMSE)=0.31) for the test set, and this model followed strictly the Organization for Economic Co-operation and Development (OECD) principles about QSAR/QSPR (OECD, 2014). A sequence of procedures complying with the OECD principle assures the rationality and reliability of the model.

Absorption, distribution, metabolism and excretion

Drug absorption is the process by which drugs enter the bloodstram from the site of administration. The bioavailability is an important pharmacokinetic parameter which reflects the degree of absorption. Predicting the bioavailability of a molecule can guide the medicinal chemist to optimize its absorption properties. Tian et al. collected a dataset including 1,014 molecules and employed the MLR model to predict bioavailability with structural fingerprints and molecular properties (Tian et al., 2011). Genetic function approximation technique was used to make the selection of molecular properties used for model training automatically, and the results gave a good predictive performance, with the correlation coefficient and RMSE of 0.71 and 0.2355, respectively.

Drug distribution is the process by which drugs circulate with blood to interstitial fluid and intracellular fluid following drug absorption (Sim, 2015a). The distribution at steady state (VDss) of a drug is the ratio of its dose *in vivo* to its plasma concentration at steady state. The VDss stands for the extent to which a drug is distributed in the tissue and it is an important index to evaluate the drug distribution. Predicting VDss can guide medicinal chemists to make structural modifications for better pharmacokinetic properties. Lombardo and Jing collected a dataset including 1,096 molecules and constructed PLS and RF models to predict VDss (Lombardo and Jing, 2016). The prediction results of their model on the external test set was not satisfactory as only about 50% molecules are within 2-fold error. Apparently, it is challenging to predict VDss value solely from molecular structural information because there are many unknown

factors that may affect VDss.

After a drug is administrated into the body, it will first pass the metabolism system that might cause the drug for function loss, or produce toxic metabolites in some cases. Prediction of the site of metabolism with high accuracy can guide the structural optimization for ensuring the metabolic stability of the molecule. A large amount of data related to drug metabolism have been collected, and many machine learning methods have been used to predict the sites at which molecules are metabolized by different metabolic enzymes, including cytochrome P450s (CYP450s), aldehyde oxidase, and UDP-glucuronosyltransferases (UGTs). For example, based on a neural networks method, XenoSite (Matlock et al., 2015) can provide the possibility that the site of small molecules metabolized by CYP450s with an overall accuracy of 87% (Zaretzki et al., 2013). In addition, the Xenosite platform also uses a neural network trained on a large database of UGT metabolism to predict UGT sites of the compound metabolism (Dang et al., 2016).

Drug excretion is the process by which drugs and their metabolites are eliminated from the body. Drug metabolites are usually soluble in water and can be easily excreted while some drugs can be directly excreted without metabolism (Sim, 2015b). Lombardo et al. used the PCA method to predict primary clearance mechanism and the model showed good discrimination results between different mechanisms, with a predictive accuracy of 84% (Lombardo et al., 2014). Based on the elimination mechanism prediction model, Lombardo et al. used the PLS model to predict the total human clearance and the PLS model performed well and was competitive with animal scaling methods.

Toxicity and the ADME/T multi-task neural network

During the development of new drugs, pre-clinical and clinical toxicity brings about the attrition of approximately one-third of leading compounds (Guengerich, 2011). Therefore, predicting the toxicity of compounds is helpful to guide the optimization of lead compounds and reduce the risk of failure during drug development. Traditionally, drug toxicity profiles (e.g., liver and kidney toxicity) are predicted by rule-based expert knowledge and structural alerts, which tends to cause false positives and cannot extensively summarize all essential structural features. Recently, due to the ability of handling diverse chemical characters and the merit of extracting features automatically, the deep learning models yield good performance on toxicity prediction. For example, based on the molecular graph encoding convolutional neural networks (MGE-CNN), Xu et al. built an acute oral toxicity prediction model, and the prediction results are better than the previously reported models based on SVM (Xu et al., 2017). The MGE-CNN model is consecutive because the molecular encoding, feature extraction and model construction are carried out in the same process of the neural networks training. In addition, the MGE-CNN model is highly flexible in which molecular fingerprints can be adjusted according to specific problems. Xu et al. mapped the toxicological features of fingerprints back to atomic levels and obtained some highlighted fragments that correspond to structural alerts defined in the ToxAlerts (Sushko et al., 2012). Therefore, as a similarity to Duvenaud's model, the model of Xu et al. is also interpretable. Mayr et al. developed a multi-task DNN model called DeepTox to predict the toxicity and the DeepTox model obviously outperformed other contestants in the Tox21 challenge (Mayr et al., 2016). By sharing the same parameters, the multi-task neural network model was trained to predict many different individual tasks that are highly correlated. Compared with single-task neural network, the performance of multi-task neural network is ordinarily better because of the sharing parameters of different tasks in helping the multi-task model for learning more common features.

Drug absorption, distribution, metabolism, excretion and drug toxicity in the body have some relevance and the multi-tasking neural network can improve the prediction performance of these tasks. Kearnes et al. used ADME/T experimental datasets of the Vertex Pharmaceuticals to compare the single-task and multi-task neural network, and their results suggested that multi-task models would generate better results as expected (Kearnes et al., 2016a).

MACHINE LEARNING STRATEGIES AND AVAILABLE PROGRAMS FOR DRUG DESIGN SCENARIOS

Methods of the molecular representation

In drug design, molecular fingerprints, numbers, ASCII strings and graphs that represent the molecules can be used as the input features of machine learning methods.

The molecular fingerprints encode the molecular attributes as a series of binary bits ("1" indicates that the molecular attribute exists, and "0" indicates that the molecular attribute does not exist). In the field of the drug design, molecular fingerprints are constantly used to predict molecular properties and calculate molecular similarity because it is a simple and effective method to represent the molecules. At present, the molecular fingerprints frequently used as the input of the neural networks are structure-based 2D molecular fingerprints, such as the Molecular ACCess System (MACCS) (Durant et al., 2003), the Extended-Connectivity Fingerprint (ECFP) (Rogers and Hahn, 2010), the Functional Class Fingerprint (FCFP) and the Molprint2D (Bender et al., 2004). For example, MACCS has been employed in training an AAE model to search anti-cancer molecules (Kadurin et al., 2017a).

For a long time, chemists have used 2D molecular graphs

to represent molecular structures and analyze molecular properties qualitatively. Amazingly, the development of AI makes it possible to quantify this process. CNN is a powerful tool to extract features from the molecular graph automatically, which can be used for generating molecular representation in predictions of bioactivity (Wallach et al., 2015), toxicity (Xu et al., 2017), physicochemical properties (Duvenaud et al., 2015) and protein-ligand affinity (Jimenez et al., 2018). Compared with ECFP, the graph convolutional methods are more flexible because the graph architecture can be adjusted based on the given tasks. In addition, the graph convolutional architecture can be combined with neural networks to predict the molecular properties, making the training process, molecular feature extraction and model construction completed simultaneously. The molecular graph CNN fingerprints include Duvenaud's graph convolutional fingerprints based on atomic radiation method (Duvenaud et al., 2015), Kearnes's graph convolutional fingerprints based on atoms, bonds and pairwise relationships (Kearnes et al., 2016), and Coley's graph convolutional fingerprints based on the molecular tensor. The basic principle of the Duvenaud's graph convolutional fingerprints is similar to the ECFP fingerprints and both of them gradually extend molecular substructures by atomic radiation methods. Specifically, Duvenaud et al. first encoded atomic features (e.g., valence, atomic identity, and number of hydrogens) and bond features into vectors, then they used the atomic and bond feature vectors to construct the atomic neighbor features to generate the initial molecular feature vectors. CNN can be used to extract the features from the above initial feature vectors at each iteration, and these values are then summed up as the molecular fingerprints. The underlying atomic and bond features are expert-designed rather than learning from the molecular graph by the AI method. The advantage of Duvenaud's graph CNN is that it can generate the molecular fingerprints suitable for a given task, and it is interpretable because the molecular fragments related to the specific molecular properties can be obtained by backtracking through the neural network nodes. This model has been implemented in the DeepChem toolbox and the results of MoleculeNet benchmark tests suggest that the graph CNN can learn useful molecular features and it often performs better than other models (Wu et al., 2017). In addition to CNN, the recursive neural networks can also be used for molecular representation. For instance, Gregor Urban et al. developed the inner and outer recursive neural networks for graph representation of the molecule (Urban et al., 2018). Compared with Kearnes's method, this method generally gives better prediction results on public data sets of the MoleculeNet benchmark tests (Wu et al., 2017).

The string representations of small molecules include the Wiswesser line-formula notation (WLN)(Smith and Wiswesser, 1975), SYBYL line notation (SLN)(Ash et al., 1997), SMILES (Weininger, 2011) and the International Chemical Identifier (InChI)(Heller et al., 2015). Among them, SMILES is more widely used supported by many programs (e.g., ChemDraw, Cheopy, and RDKit) and databases (e.g., PubChem and ZINC). RNN can be used to learn the coding grammar of SMILES (Segler et al., 2018), which can be converted into the molecular graph. In addition, SMILES can be directly used as an input feature of RNN in predicting the molecular properties (Goh et al., 2017).

Molecular descriptors conventionally refer to the structural or physicochemical properties of a molecule, which can be obtained by molecular encoding or through standard experiments (Todeschini and Consonni, 2009). The detailed description of these descriptors has been reviewed elsewhere (Sahoo et al., 2016). The proper selection of the descriptors is critical for machine learning, which can decrease the amount of computations, increase the model generalization ability and improve the performance and interpretability of the model (Danishuddin and Khan, 2016). The common software to calculate molecular descriptors includes Dragon (Mauri et al., 2006), Cheopy (Cao et al., 2013), PaDEL (Yap, 2011) and Cinfony (O'Boyle and Hutchison, 2008). The summary of the molecular representation is provided in Table 1.

Transfer learning for low data

The deep learning methods have shown a good prospect in drug design due to the strong data mining capabilities. However, the deep learning methods typically require a great amount of training data, which has restricted its application generally. For example, with only a small amount of the activity data available, it is difficult to predict the bioactivity of the new molecules because low data cannot capture an adequate chemical space. The transfer learning method can be used in solving such problems by leveraging existing knowledge obtained from other related data resources. As we know, human experts can apply previously learned knowledge to solve new problems and the ability helps us solve the problem efficiently. One of the directions of AI study is to imitate this ability by a transferring learning scheme (Pan and Yang, 2010). The basic principle of the transfer learning is to apply the knowledge seeking from some previous tasks to a relevant target task with fewer training data. Furthermore, one-shot learning method has been proposed which refers to the deep learning method that requires only a few training samples. It can transfer information between relevant, but different tasks by learning a meaningful distance metric (Vinyals et al., 2016). Altae-Tran et al. developed a one-shot learning method that combined the iterative refinement of long short-term neural networks with the graph CNN for low data learning (Altae-Tran et al., 2017; Goodfellow et al., 2016). The model outperforms the RF and other

Table 1 Summary of the molecular representation

Representation methods	Examples
Molecular fingerprints: MACCS, ECFP, FCFP, Molprint2D, etc.	MACCS was employed as the input and output of the AAE to search anti-cancer molecules (Kadurin et al., 2017a).
Graphs: the molecular graph	CNN graph convolutional representation methods: Duvenaud graph convolution fingerprints (Duvenaud et al., 2015), Kearnes graph convolution fingerprints (Kearnes et al., 2016b), and Coley's graph convolution fingerprints (Coley et al., 2017). Gregor Urban et al. developed the inner and outer recursive neural networks for graph representation of the molecule (Urban et al., 2018).
ASCII strings: SMILES, InChI, SLN, WLN, etc.	Olivecrona et al. developed the deep reinforcement learning method to tune the RNN to generate the molecules with predicted biological activity (Olivecrona et al., 2017). SMILES can be directly used as an input feature of RNN to predict molecular properties (Goh et al., 2017).
Numbers: molecular descriptor	Ma et al. used the DNN to predict molecular bioactivity with the union of the atom pair descriptor and the donor-acceptor pair descriptor (Ma et al., 2015). Mayr et al. developed a multi-task DNN model to predict with the chemical descriptors (Mayr et al., 2016).

methods on the Tox21 and SIDER dataset. However, when the toxicity data is utilized for training a model in predicting a side effects dataset, it will completely fail because the relevance between the two datasets is rather weak.

The cross-validation method

The cross-validation method is employed to evaluate the performance of the model and the common practice is the random-split cross validation. However, the random-split cross-validation method is often too optimistic for the estimation of model predictive effect because it weakens the covariate changes in drug development through mixing different series data (Cortes et al., 2014). Alternatively, the process of the time-split cross-validation was proposed, where the datasets were divided into training and test sets based on the experimental time order of the data (Chen et al., 2012). Sheridan et al. compared different cross validation methods used to estimate performance of the QSAR model and their results showed that the R^2 value given by time-split cross validation method was much closer to the true prospective predictive value (Sheridan, 2013). Guided by this result, Ma et al. applied the time-split cross-validation rather than traditional random-spilt cross-validation to evaluate the performance of the deep neural network in simulating the realistic hit-to-lead process (Ma et al., 2015). For all of these studies, experimental time is a very important parameter, and time-split cross validation should be carried out in drug discovery as long as the data of experimental time information is available.

The skills of training the deep neural networks

Although deep learning models outperform many traditional machine learning methods, they still involve much more parameters and different architectures, which leads to some difficulties while training, especially under the circumstances when the samples are not enough or the feature matrix is sparse. The training algorithm might only get a local optimum and the accuracy is not satisfactory enough. In order to deal with this problem, the unsupervised pre-training method such as deep belief network has been proposed to improve the parameter initialization, and the results suggest that the method was more effective than the random initial values (Ghasemi et al., 2017). The study of Ma et al. indicated that the dropout strategy could effectively avoid overfitting when training the QSAR dataset (Ma et al., 2015). Moreover, compared with the sigmoid action function, ReLU action function is more suitable for the QSAR tasks as its advantages in preventing the gradient disappear and local optimum.

The available AI source code for drug design

In pharmaceutical industry, the commercial potential of computer software for drug design is evident. However, there are still many researchers willing to share their programs on github or other open source platforms, to integrate the AI methods with drug design methods. Several open source implementation of AI-based drug design models are summarized in Table 2. These open source projects will promote the widespread application of artificial intelligence technologies in the field.

CONCLUSION AND FUTURE

The AI technology, especially the deep learning method, can be used to learn pharmaceutical knowledge (e.g., QSAR and chemical structure) from vast amount of the pharmaceutical data. The learned knowledge can be then applied to discover and design the molecule with desired properties, to optimize the molecular properties, and to push forward the clinical

Table 2 Summary of the AI implementation programs in drug design

Programs	Websites	Description
DeepChem	https://github.com/deepchem/deepchem	A free python library that incorporates many high quality AI algorithms for the drug discovery
Neural Graph Fingerprints	https://github.com/HIPS/neural-fingerprint	CNN is used to generate molecular fingerprints to predict molecular properties.
Conv_qsar_fast	https://github.com/connorcoley/conv_qsar_fast	The tensor-basd CNN is used to predict molecular properties.
DeepNeuralNet-QSAR	https://github.com/Merck/DeepNeuralNet-QSAR	Multi-task DNN is used to predict molecular activity.
DeltaVina	https://github.com/chengwang88/deltavina	A rescoring approach combining the RF with AutoDock scoring function
Chemical VAE	https://github.com/aspuru-guzik-group/chemical_vae	An implementation of VAE generation model proposed by Gómez-Bombarelli et al.
ORGANIC (Sanchez-Lengeling, 2017)	https://github.com/aspuru-guzik-group/ORGANIC	A generative model for *de novo* molecule design with desired properties
REINVENT	https://github.com/MarcusOlivecrona/REINVENT	A generative model for *de novo* molecule design by using RNN and reinforcement learning
Open Drug Discovery Toolkit (ODDT) (Wójcikowski et al., 2015)	https://github.com/oddt/oddt	A modular and comprehensive toolkit for use in cheminformatics and molecular modeling
JunctionTree VAE (Jin et al., 2018)	https://github.com/wengong-jin/icml18-jtnn/tree/master/mol-vae	A generative model for *de novo* molecular design based on junction tree VAE
SCScore	https://github.com/connorcoley/scscore	A score evaluating synthetic complexity of the molecule
InnerOuterRNN	https://github.com/Chemoinformatics/InnerOuterRNN	Two kinds of recursive neural networks used to predict molecular properties

approval success rate of the molecule. The AI technology has breathed a new life into the computer-aided drug design owing to its powerful data mining capabilities. However, some difficult problems may still exist: (1) As a data mining technology, the amount of available data directly affects the performance of the related deep learning models since the successful training of deep neural networks highly relies on large amount of data. The development of transfer learning technology may be a potential approach for solving this problem. (2) The deep learning method is a "black box" and the mechanism of the model remains mysterious. The counterfactual probe (e.g., LIME) was used to open the black box of the deep learning method (Voosen, 2017) and the information bottleneck new theory was proposed to explain the mechanism of the deep learning method (Tishby and Zaslavsky, 2015). However, the studies trying to reveal the deep learning models are still in their early stages. (3) Training neural network models involves adjusting many parameters but there are only a few practical guidelines, and the complete theoretical systems for the optimizations of these models are still out of reach.

In the near future, the AI technology will be expected to cover all the aspects of new drug discovery and development. An automated drug development AI platform that integrates theoretical computation results (e.g., molecular docking, molecular dynamics simulation, and quantum chemistry calculation), omics data, chemistry data and biomedical data will emerge and we expect to witness a revolution in the new drug discovery campaign.

Compliance and ethics *The author(s) declare that they have no conflict of interest.*

Acknowledgements *This work was supported by the National Natural Science Foundation of China (21210003 and 81230076 to H.J., 81773634 to M.Z. and 81430084 to K.C.), the "Personalized Medicines—Molecular Signature-based Drug Discovery and Development", Strategic Priority Research Program of the Chinese Academy of Sciences (XDA12050201 to M.Z.), National Key Research & Development Plan (2016YFC1201003 to M.Z.), and the National Basic Research Program (2015CB910304 to X.L.).*

Abagyan, R., Totrov, M., and Kuznetsov, D. (1994). ICM—A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. J Comput Chem 15, 488–506.

Ain, Q.U., Aleksandrova, A., Roessler, F.D., and Ballester, P.J. (2015). Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. WIREs Comput Mol Sci 5, 405–424.

Altae-Tran, H., Ramsundar, B., Pappu, A.S., and Pande, V. (2017). Low data drug discovery with one-shot learning. ACS Cent Sci 3, 283–293.

Andras, P. (2017). High-dimensional function approximation with neural networks for large volumes of data. IEEE Trans Neural Netw Learn Syst 99, 1–9.

Angermueller, C., Pärnamaa, T., Parts, L., and Stegle, O. (2016). Deep learning for computational biology. Mol Syst Biol 12, 878.

Artursson, P., and Karlsson, J. (1991). Correlation between oral drug absorption in humans and apparent drug permeability coefficients in human intestinal epithelial (Caco-2) cells. Biochem Biophys Res Commun 175, 880–885.

Ash, S., Cline, M.A., Homer, R.W., Hurst, T., and Smith, G.B. (1997). ChemInform abstract: SYBYL line notation (SLN): a versatile language for chemical structure representation. ChemInform 28, no.

Ashburn, T.T., and Thor, K.B. (2004). Drug repositioning: identifying and developing new uses for existing drugs. Nat Rev Drug Discov 3, 673–

683.
Bai, F., Morcos, F., Cheng, R.R., Jiang, H., and Onuchic, J.N. (2016). Elucidating the druggable interface of protein-protein interactions using fragment docking and coevolutionary analysis. Proc Natl Acad Sci USA 113, E8051–E8058.
Bender, A., And, H.Y.M., Glen, R.C., and Reiling, S. (2004). Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): evaluation of performance. J Chem Inf Comput Sci 44, 1708–1718.
Cabreiro, F., Au, C., Leung, K.Y., Vergara-Irigaray, N., Cocheme, H.M., Noori, T., Weinkove, D., Schuster, E., Greene, N.D., and Gems, D. (2013). Metformin retards aging in *C. elegans* by altering microbial folate and methionine metabolism. Cell 153, 228–239.
Cao, D.S., Xu, Q.S., Hu, Q.N., and Liang, Y.Z. (2013). ChemoPy: freely available python package for computational biology and chemoinformatics. Bioinformatics 29, 1092–1094.
Chen, B., Sheridan, R.P., Hornak, V., and Voigt, J.H. (2012). Comparison of random forest and pipeline pilot naïve bayes in prospective QSAR predictions. J Chem Inf Model 52, 792–803.
Chen, R., Li, L., and Weng, Z. (2003). ZDOCK: an initial-stage protein-docking algorithm. Proteins 52, 80–87.
Chen, Y.C. (2015). Beware of docking! Trends Pharmacol Sci 36, 78–95.
Coley, C.W., Barzilay, R., Green, W.H., Jaakkola, T.S., and Jensen, K.F. (2017). Convolutional embedding of attributed molecular graphs for physical property prediction. J Chem Inf Model 57, 1757–1772.
Coley, C.W., Rogers, L., Green, W.H., and Jensen, K.F. (2018). SCScore: synthetic complexity learned from a reaction corpus. J Chem Inf Model 58, 252–261.
Copeland, R.A. (2010). The dynamics of drug-target interactions: drug-target residence time and its impact on efficacy and safety. Expert Opin Drug Discovery 5, 305–310.
Cortes, C., Kuznetsov, V., and Mohri, M. (2014). Ensemble methods for structured prediction. Proceedings of 31st International Conference on Machine Learning 2014, 1134–1142.
Cukuroglu, E., Engin, H.B., Gursoy, A., and Keskin, O. (2014). Hot spots in protein–protein interfaces: Towards drug discovery. Prog Biophys Mol Biol 116, 165–173.
Dahl, G.E., Jaitly, N., and Salakhutdinov, R. (2014). Multi-task neural networks for QSAR predictions. Comput Sci, arXiv:1406.1231v1.
Dang, N.L., Hughes, T.B., Krishnamurthy, V., and Swamidass, S.J. (2016). A simple model predicts UGT-mediated metabolism. Bioinformatics 32, 3183–3189.
Danishuddin, and Khan, A.U. (2016). Descriptors and their selection methods in QSAR analysis: paradigm for drug design. Drug Discov Today 21, 1291–1302.
De Haes, W., Frooninckx, L., Van Assche, R., Smolders, A., Depuydt, G., Billen, J., Braeckman, B.P., Schoofs, L., and Temmerman, L. (2014). Metformin promotes lifespan through mitohormesis via the peroxiredoxin PRDX-2. Proc Natl Acad Sci USA 111, E2501–E2509.
DiMasi, J.A., Grabowski, H.G., and Hansen, R.W. (2016). Innovation in the pharmaceutical industry: New estimates of R&D costs. J Health Economics 47, 20–33.
Dobchev, D., Pillai, G., and Karelson, M. (2014). *In silico* machine learning methods in drug development. CTMC 14, 1913–1922.
Du, T., Liao, L., Wu, C.H., and Sun, B. (2016). Prediction of residue-residue contact matrix for protein-protein interaction with Fisher score features and deep learning. Methods 110, 97–105.
Duch, W., Swaminathan, K., and Meller, J. (2007). Artificial intelligence approaches for rational drug design and discovery. CPD 13, 1497–1508.
Dudek, A., Arodz, T., and Galvez, J. (2006). Computational methods in developing quantitative structure-activity relationships (QSAR): a review. CCHTS 9, 213–228.
Durant, J.L., Leland, B.A., Henry, D.R., and Nourse, J.G. (2003). Reoptimization of MDL keys for use in drug discovery. J Chem Inf Comput Sci 34, 1273–1280.
Duvenaud, D., Maclaurin, D., Aguilera-Iparraguirre, J., Hirzel, T., and Adams, R.P. (2015). Convolutional networks on graphs for learning molecular fingerprints. In Proceedings of the 28th International Conference on Neural Information Processing Systems, pp. 2224-2232.
Esposito, E.X., Hopfinger, A.J., and Madura, J.D. (2004). Methods for applying the quantitative structure-activity relationship paradigm. Methods Mol Biol 275, 131-214.
Falchi, F., Caporuscio, F., and Recanatini, M. (2014). Structure-based design of small-molecule protein-protein interaction modulators: the story so far. Future Medicinal Chem 6, 343–357.
Free, S.M., and Wilson, J.W. (1964). A mathematical contribution to structure-activity studies. J Med Chem 7, 395–399.
Friesner, R.A., Banks, J.L., Murphy, R.B., Halgren, T.A., Klicic, J.J., Mainz, D.T., Repasky, M.P., Knoll, E.H., Shelley, M., and Perry, J.K. (2004). Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J Med Chem 47, 1739–1749.
Ghasemi, F., Mehridehnavi, A.R., Fassihi, A., and Pérez-Sánchez, H. (2017). Deep neural network in biological activity prediction using deep belief network. Appl Soft Comput 62, doi: 10.1016/j.asoc.2017.09.040.
Goh, G.B., Hodas, N.O., Siegel, C., and Vishnu, A. (2017). SMILES2Vec: An Interpretable General-Purpose Deep Neural Network for Predicting Chemical Properties. arXiv:1712.02034v2.
Gómez-Bombarelli, R., Wei, J.N., Duvenaud, D., Hernández-Lobato, J.M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T.D., Adams, R.P., and Aspuru-Guzik, A. (2018). Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent Sci 4, 268–276.
Goodfellow, I., Bengio, Y., and Courville, A. (2016). Deep Learning (Cambridge: The MIT Press).
Guengerich, F.P. (2011). Mechanisms of drug toxicity and relevance to pharmaceutical development. Drug Metab Pharmacokinetics 26, 3–14.
Hansch, C., and Fujita, T. (1964). Additions and corrections -ρ-σ-π analysis. A method for the correlation of biological activity and chemical structure. J Am Chem Soc 86, 5710.
Hartenfeller, M., and Schneider, G. (2011). *De novo* drug design. Methods Mol Biol 672, 299–323.
Hassan Baig, M., Ahmad, K., Roy, S., Mohammad Ashraf, J., Adil, M., Haris Siddiqui, M., Khan, S., Amjad Kamal, M., Provazník, I., and Choi, I. (2015). Computer aided drug design: success and limitations. CPD 22, 572–581.
Heller, S.R., McNaught, A., Pletnev, I., Stein, S., and Tchekhovskoi, D. (2015). InChI, the IUPAC international chemical identifier. J Cheminform 7, 23.
Higueruelo, A.P., Jubb, H., and Blundell, T.L. (2013). Protein-protein interactions as druggable targets: recent technological advances. Curr Opin Pharmacol 13, 791–796.
Huang, S.Y., and Zou, X. (2010). Inclusion of solvation and entropy in the knowledge-based scoring function for protein-ligand interactions. J Chem Inf Model 50, 262–273.
Huang, S.Y., Grinter, S.Z., and Zou, X. (2010). Scoring functions and their evaluation methods for protein-ligand docking: recent advances and future directions. Phys Chem Chem Phys 12, 12899–12908.
Hubatsch, I., Ragnarsson, E.G.E., and Artursson, P. (2007). Determination of drug permeability and prediction of drug absorption in Caco-2 monolayers. Nat Protoc 2, 2111–2119.
Jimenez, J., Skalic, M., Martinez-Rosell, G., and De Fabritiis, G. (2018). KDEEP: protein-ligand absolute binding affinity prediction via 3D-convolutional neural networks. J Chem Inf Model 58, 287–296.
Jin, W., Barzilay, R., and Jaakkola, T. (2018). Junction tree variational autoencoder for molecular graph generation. arXiv:1802.04364v2.
Kadurin, A., Aliper, A., Kazennov, A., Mamoshina, P., Vanhaelen, Q., Khrabrov, K., and Zhavoronkov, A. (2017a). The cornucopia of meaningful leads: Applying deep adversarial autoencoders for new molecule development in oncology. Oncotarget 8, 10883.
Kadurin, A., Nikolenko, S., Khrabrov, K., Aliper, A., and Zhavoronkov, A. (2017b). druGAN: An advanced generative adversarial autoencoder model for *de novo* generation of new molecules with desired molecular properties *in silico*. Mol Pharm 14, 3098–3104.

Kearnes, S., Goldman, B., and Pande, V. (2016a). Modeling industrial ADMET data with multitask networks. arXiv:1606.08793v3.
Kearnes, S., Mccloskey, K., Berndl, M., Pande, V., and Riley, P. (2016b). Molecular graph convolutions: moving beyond fingerprints. J Comput Aid Mol Design 30, 1–14.
Khamis, M.A., Gomaa, W., and Ahmed, W.F. (2015). Machine learning in computational docking. Artif Intell Med 63, 135–152.
Kim, K.H., Kim, N.D., and Seong, B.L. (2010). Pharmacophore-based virtual screening: a review of recent applications. Expert Opin Drug Discov 5, 205–222.
Kinnings, S.L., Liu, N., Tonge, P.J., Jackson, R.M., Xie, L., and Bourne, P. E. (2011). A machine learning-based method to improve docking scoring functions and its application to drug repurposing. J Chem Inf Model 51, 408–419.
Klaeger, S., Heinzlmeir, S., Wilhelm, M., Polzer, H., Vick, B., Koenig, P. A., Reinecke, M., Ruprecht, B., Petzoldt, S., Meng, C., et al. (2017). The target landscape of clinical kinase drugs. Science 358, eaan4368.
Labbé, C.M., Kuenemann, M.A., Zarzycka, B., Vriend, G., Nicolaes, G.A. F., Lagorce, D., Miteva, M.A., Villoutreix, B.O., and Sperandio, O. (2016). iPPI-DB: an online database of modulators of protein-protein interactions. Nucleic Acids Res 44, D542–D547.
Lavecchia, A., and Giovanni, C. (2013). Virtual screening strategies in drug discovery: a critical review. CMC 20, 2839–2860.
LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature 521, 436–444.
Leelananda, S.P., and Lindert, S. (2016). Computational methods in drug discovery. Beilstein J Org Chem 12, 2694–2718.
Li, H., Hou, J., Adhikari, B., Lyu, Q., and Cheng, J. (2017). Deep learning methods for protein torsion angle prediction. BMC BioInf 18, 417.
Liew, C.Y., Ma, X.H., Liu, X., and Yap, C.W. (2009). SVM model for virtual screening of Lck inhibitors. J Chem Inf Model 49, 877.
Lombardo, F., and Jing, Y. (2016). *In silico* prediction of volume of distribution in humans. Extensive data set and the exploration of linear and nonlinear methods coupled with molecular interaction fields descriptors. J Chem Inf Model 56, 2042–2052.
Lombardo, F., Obach, R.S., Varma, M.V., Stringer, R., and Berellini, G. (2014). Clearance mechanism assignment and total clearance prediction in human based upon *in silico* models. J Med Chem 57, 4397–4405.
Lotfi Shahreza, M., Ghadiri, N., Mousavi, S.R., Varshosaz, J., and Green, J. R. (2017). A review of network-based approaches to drug repositioning. Brief Bioinform, doi: 10.1093/bib/bbx017.
Luo, Y., Zhao, X., Zhou, J., Yang, J., Zhang, Y., Kuang, W., Peng, J., Chen, L., and Zeng, J. (2017). A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. Nat Commun 8, 573.
Lusci, A., Pollastri, G., and Baldi, P. (2013). Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules. J Chem Inf Model 53, 1563–1575.
Ma, J., Sheridan, R.P., Liaw, A., Dahl, G.E., and Svetnik, V. (2015). Deep neural nets as a method for quantitative structure-activity relationships. J Chem Inf Model 55, 263–274.
Ma, X., Jia, J., Zhu, F., Xue, Y., Li, Z., and Chen, Y. (2009). Comparative analysis of machine learning methods in ligand-based virtual screening of large compound libraries. CCHTS 12, 344–357.
Maheshwari, S., and Brylinski, M. (2016). Template-based identification of protein-protein interfaces using eFindSitePPI. Methods 93, 64–71.
Martin-Montalvo, A., Mercken, E.M., Mitchell, S.J., Palacios, H.H., Mote, P.L., Scheibye-Knudsen, M., Gomes, A.P., Ward, T.M., Minor, R.K., Blouin, M.J., et al. (2013). Metformin improves healthspan and lifespan in mice. Nat Commun 4, 2192.
Mason, J.S. (2007). Introduction to the volume and overview of computer-assisted drug design in the drug discovery process. In Taylor, J.B., and Triggle, D.J., ed. Comprehensive Medicinal Chemistry II (Elsevier), pp. 1–11.
Matlock, M.K., Hughes, T.B., and Swamidass, S.J. (2015). XenoSite server: a web-available site of metabolism prediction tool. Bioinformatics 31, 1136–1137.
Mauri, A., Consonni, V., Pavan, M., and Todeschini, R. (2006). DRAGON software: An easy approach to molecular descriptor calculations. Match Commun Math Comput Chem 56, 237–248.
Mayr, A., Klambauer, G., Unterthiner, T., and Hochreiter, S. (2016). DeepTox: toxicity prediction using deep learning. Front Environ Sci, https://doi.org/10.3389/fenvs.2015.00080.
Melville, J., Burke, E., and Hirst, J. (2009). Machine learning in virtual screening. CCHTS 12, 332–343.
Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. Nature 518, 529–533.
Mullard, A. (2017). The drug-maker's guide to the galaxy. Nature 549, 445–447.
Myint, K.Z., and Xie, X.Q. (2010). Recent advances in fragment-based QSAR and multi-dimensional QSAR methods. IJMS 11, 3846–3866.
Ning, X., and Karypis, G. (2011). *In silico* structure-activity-relationship (SAR) models from machine learning: a review. Drug Dev Res 72, 138–146.
O'Boyle, N.M., and Hutchison, G.R. (2008). Cinfony—combining Open Source cheminformatics toolkits behind a common interface. Chem Cent J 2, 1–10.
OECD. (2014). Guidance document on the validation of (quantitative) structure-activity relationship [(Q)Sar] models. 69, 1–154.
Olivecrona, M., Blaschke, T., Engkvist, O., and Chen, H. (2017). Molecular *de-novo* design through deep reinforcement learning. J Cheminform 9, 48.
Pan, S.J., and Yang, Q. (2010). A survey on transfer learning. IEEE Trans Knowl Data Eng 22, 1345–1359.
Pereira, J.C., Caffarena, E.R., and Dos Santos, C.N. (2016). Boosting docking-based virtual screening with deep learning. J Chem Inf Model 56, 2495.
Pu, Y., Wang, W., Henao, R., Chen, L., Gan, Z., Li, C., and Carin, L. (2017). Adversarial symmetric variational autoencoder. arXiv:1711.04915v2.
Ramsundar, B., Liu, B., Wu, Z., Verras, A., Tudor, M., Sheridan, R.P., and Pande, V. (2017). Is multitask deep learning practical for pharma? J Chem Inf Model 57, 2068–2076.
Repasky, M.P., Shelley, M., and Friesner, R.A. (2007). Flexible Ligand Docking with Glide (John Wiley & Sons, Inc.).
Rogers, D., and Hahn, M. (2010). Extended-connectivity fingerprints. J Chem Inf Model 50, 742–754.
Sahoo, S., Adhikari, C., Kuanar, M., and Mishra, B. (2016). A short review of the generation of molecular descriptors and their applications in quantitative structure property/activity relationships. CAD 12, 181–205.
Sanchez-Lengeling, B., Outeiral, C., Guimaraes, G.L., and Aspuru-Guzik, A. (2017). Optimizing distributions over molecular space. An Objective-Reinforced Generative Adversarial Network for Inverse-design Chemistry (ORGANIC). ChemRxiv Preprint.
Santos, R., Ursu, O., Gaulton, A., Bento, A.P., Donadi, R.S., Bologa, C.G., Karlsson, A., Al-Lazikani, B., Hersey, A., Oprea, T.I., et al. (2017). A comprehensive map of molecular drug targets. Nat Rev Drug Discov 16, 19–34.
Schaarschmidt, J., Monastyrskyy, B., Kryshtafovych, A., and Bonvin, A. (2017). Assessment of contact predictions in CASP12: co-evolution and deep learning coming of age. Proteins 86, https://doi.org/10.1002/prot.25407.
Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural Networks 61, 85–117.
Schneider, G., Funatsu, K., Okuno, Y., and Winkler, D. (2017). *De novo* drug design—Ye olde scoring problem revisited. Mol Inform 36, https://doi.org/10.1002/minf.201681031.
Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., and Wolfson, H.J. (2005). PatchDock and SymmDock: servers for rigid and symmetric docking. Nucleic Acids Res 33, W363–W367.
Scott, D.E., Bayly, A.R., Abell, C., and Skidmore, J. (2016). Small molecules, big targets: drug discovery faces the protein-protein

interaction challenge. Nat Rev Drug Discov 15, 533–550.
Segler, M.H.S., Kogej, T., Tyrchan, C., and Waller, M.P. (2018). Generating focussed molecule libraries for drug discovery with recurrent neural networks. ACS Cent Sci 4, 120-131.
Sheridan, R.P. (2013). Time-split cross-validation as a method for estimating the goodness of prospective prediction.. J Chem Inf Model 53, 783–790.
Shin, W.H., Christoffer, C.W., and Kihara, D. (2017). *In silico* structure-based approaches to discover protein-protein interaction-targeting drugs. Methods 131, 22–32.
Shoemaker, R.H. (2006). The NCI60 human tumour cell line anticancer drug screen. Nat Rev Cancer 6, 813–823.
Sim, D.S.M. (2015a). Drug Distribution (Springer International Publishing).
Sim, D.S.M. (2015b). Drug elimination. In Chan, Y., Ng, K., and Sim, D., ed. Pharmacological Basis of Acute Care (Springer, Cham), pp. 37-47.
Smith, E.G., and Wiswesser, W.J. (1975). The Wiswesser Line-Formula Chemical Notation (New York: McGraw-Hill).
Spencer, M., Eickholt, J., and Jianlin Cheng, J. (2015). A deep learning network approach to *ab initio* protein secondary structure prediction. IEEE/ACM Trans Comput Biol Bioinf 12, 103–112.
Subramanian, G., Ramsundar, B., Pande, V., and Denny, R.A. (2016). Computational modeling of β-secretase 1 (BACE-1) inhibitors using ligand based approaches. J Chem Inf Model 56, 1936–1949.
Sushko, I., Salmina, E., Potemkin, V.A., Poda, G., and Tetko, I.V. (2012). ToxAlerts: a web server of structural alerts for toxic chemicals and compounds with potential adverse reactions. J Chem Inf Model 52, 2310–2316.
Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P., et al. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res 43, D447–D452.
Talele, T., Khedkar, S., and Rigby, A. (2010). Successful applications of computer aided drug discovery: moving drugs from concept to the clinic. CTMC 10, 127–141.
Tian, S., Li, Y., Wang, J., Zhang, J., and Hou, T. (2011). ADME evaluation in drug discovery. 9. Prediction of oral bioavailability in humans based on molecular properties and structural fingerprints. Mol Pharm 8, 841–851.
Tishby, N., and Zaslavsky, N. (2015). Deep learning and the information bottleneck principle. Paper presented at: Information Theory Workshop, arXiv:1503.02406v1.
Todeschini, R., and Consonni, V. (2009). Molecular Descriptors for Chemoinformatics (Wiley-VCH).
Turner, J.R. (2010). New Drug Development (Springer New York).
Unterthiner, T., Mayr, A., Klambauer, G., Steijaert, M., Ceulemans, H., Wegner, J.K., and Hochreiter, S. (2014). Deep learning as an opportunity in virtual screening. Paper presented at: The Workshop on Deep Learning & Representation Learning.
Urban, G., Subrahmanya, N., and Baldi, P. (2018). Inner and outer recursive neural networks for chemoinformatics applications. J Chem Inf Model 58, 207–211.
Vakser, I.A. (2014). Protein-protein docking: from interaction to interactome. BioPhys J 107, 1785–1793.
Valkov, E., Sharpe, T., Marsh, M., Greive, S., and Hyvonen, M. (2012). Targeting protein-protein interactions and fragment-based drug discovery. Top Curr Chem 317, 145–179.
Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., and Wierstra, D. (2016). Matching networks for one shot learning. Papers published at the Neural Information Processing Systems Conference.
Vohora, D., and Singh, G. (2017). Pharmaceutical Medicine and Translational Clinical Research (Academic Press).
Voosen, P. (2017). The AI detectives. Science 357, 22–27.
Wallach, I., Dzamba, M., and Heifets, A. (2015). AtomNet: A deep convolutional neural network for bioactivity prediction in structure-based drug discovery. Mathematische Zeitschrift 47, 34–46.
Wang, C., and Zhang, Y. (2017). Improving scoring-docking-screening powers of protein-ligand scoring functions using random forest. J Comput Chem 38, 169–177.
Wang, J., Luo, C., Shan, C., You, Q., Lu, J., Elf, S., Zhou, Y., Wen, Y., Vinkenborg, J.L., Fan, J., et al. (2015). Inhibition of human copper trafficking by a small molecule significantly attenuates cancer cell proliferation. Nat Chem 7, 968–979.
Wang, N.N., Dong, J., Deng, Y.H., Zhu, M.F., Wen, M., Yao, Z.J., Lu, A.P., Wang, J.B., and Cao, D.S. (2016). 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, 763–773.
Wang, S., Sun, S., Li, Z., Zhang, R., and Xu, J. (2017). Accurate *de novo* prediction of protein contact map by ultra-deep learning model. PLoS Comput Biol 13, e1005324.
Wang, W., Yang, S., Zhang, X., and Li, J. (2014). Drug repositioning by integrating target information through a heterogeneous network model. Bioinformatics 30, 2923–2930.
Weininger, D. (2011). Simplified Molecular Input Line Entry Specification.
Willett, P. (2006). Similarity-based virtual screening using 2D fingerprints. Drug Discovery Today 11, 1046–1053.
Wójcikowski, M., Zielenkiewicz, P., and Siedlecki, P. (2015). Open Drug Discovery Toolkit (ODDT): a new open-source player in the drug discovery field. J Cheminform 7, 26.
Wu, Z., Ramsundar, B., Feinberg, E.N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V. (2017). MoleculeNet: A benchmark for molecular machine learning. arXiv:1703.00564v2.
Xing, J., Lu, W., Liu, R., Wang, Y., Xie, Y., Zhang, H., Shi, Z., Jiang, H., Liu, Y.C., Chen, K., et al. (2017). Machine-learning-assisted approach for discovering novel inhibitors targeting bromodomain-containing protein 4. J Chem Inf Model 57, 1677–1690.
Xu, Y., Pei, J., and Lai, L. (2017). Deep learning based regression and multiclass models for acute oral toxicity prediction with automatic chemical feature extraction. J Chem Inf Model 57, 2672–2685.
Xue, L.C., Dobbs, D., Bonvin, A.M.J.J., and Honavar, V. (2015). Computational prediction of protein interfaces: A review of data driven methods. FEBS Lett 589, 3516–3526.
Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W., and Kanehisa, M. (2008). Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. Bioinformatics 24, i232–i240.
Yap, C.W. (2011). PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem 32, 1466–1474.
Zaretzki, J., Matlock, M., and Swamidass, S.J. (2013). XenoSite: accurately predicting CYP-mediated sites of metabolism with neural networks. J Chem Inf Model 53, 3373–3383.
Zhang, Q.C., Petrey, D., Norel, R., and Honig, B.H. (2010). Protein interface conservation across structure space. Proc Natl Acad Sci USA 107, 10896–10901.
Zsoldos, Z., Reid, D., Simon, A., Sadjad, S.B., and Johnson, A.P. (2007). eHiTS: a new fast, exhaustive flexible ligand docking system. J Mol Graphics Model 26, 198–212.