Chapter

Artificial Intelligence-Based Drug Design and Discovery

Yu-Chen Lo, Gui Ren, Hiroshi Honda and Kara L. Davis

Abstract

The drug discovery process from hit-to-lead has been a challenging task that requires simultaneously optimizing numerous factors from maximizing compound activity, efficacy to minimizing toxicity and adverse reactions. Recently, the advance of artificial intelligence technique enables drugs to be efficiently purposed *in silico* prior to chemical synthesis and experimental evaluation. In this chapter, we present fundamental concepts of artificial intelligence and their application in drug design and discovery. The emphasis will be on machine learning and deep learning, which demonstrated extensive utility in many branches of computer-aided drug discovery including de novo drug design, QSAR (Quantitative Structure–Activity Relationship) analysis, drug repurposing and chemical space visualization. We will demonstrate how artificial intelligence techniques can be leveraged for developing chemoinformatics pipelines and presented with real-world case studies and practical applications in drug design and discovery. Finally, we will discuss limitations and future direction to guide this rapidly evolving field.

Keywords: artificial intelligence, chemoinformatics, data mining, drug discovery

1. Introduction

The path of drug discovery from small molecule ligands to drugs that can be utilized clinically has been a long and arduous process. Starting with a hit compound, the drugs need to be evaluated through multiple *in vitro* and cell-based assays to improve the mechanism of actions followed by mouse models to demonstrate appropriate *in vivo* and transport properties. Mechanistically, the drugs not only need to exert enough binding affinity to the disease targets, but also necessitate proper transport through multiple physiological barriers to enable access to these targets. Other problems like chemical toxicity, often induced by off-targets interactions with unintended proteins as well as pharmacogenetic, where genetic variation influences drug responses all need to be considered in drug design. Therefore, these multifaceted problems in drug discovery often posed significant challenges for drug designers. Recently, the rise of artificial intelligence approach saw potential solutions to these challenges. A sub-umbrella of artificial intelligence called machine-learning has taken a central stage in many R&D sectors of pharmaceutical companies that allows drugs to be developed more efficiently and at the same time mitigate the cost associated with the required experiments [1]. Given some observations of chemical data, machine learning can be used to construct a predictor by learning compound properties from extracted features of compound structures and interactions. Because this approach does not require a mechanistic

understanding of how drugs behave, many compound properties like binding affinity and other transport and toxicity problems can be accurately forecasted in this way before they are synthesized [2]. Furthermore, by simultaneously tackling the Pharmacokinetics/Pharmacodynamics (PK/PD) problems using artificial intelligence, we can expect that the effort and time required to bring a drug from bench to bedside can be substantially reduced. In this regard, the artificial intelligence approach has now become an essential tool to facilitate the drug discovery process.

2. Chemoinformatic for drug discovery

2.1 Chemical formats

To facilitate the discussion on artificial intelligence and machine learning in drug discovery and design, it is necessary to understand the type of format and data presentation commonly used for chemical compounds in chemoinformatics. Chemoinformatics is a broad field that studying the application of computers in storing, processing and analyzing chemical data. The field already has more than 30 years of development with focuses on subjects such as chemical representation, chemical descriptors analysis, library design, QSAR analysis and computer-aided drug design [3]. Along with these developments, several popular chemical data formats for data processing has been proposed. Intuitively, the chemical compound is best represented by graphs, also known as "chemical graph" or "molecular graph" where nodes represent atoms and edges represent bonds. The molecular graph is useful for distinguishing different structural isomers but does not contain 3D conformation of the molecules. To store 2D or 3D coordinates of compounds, chemical file formats such as Structure Data Format (SDF), MDL (Molfile), and Protein Data Bank (PDB) formats can be used. In contrast to the PDB file that simply store structural data, the SDF format provides additional advantages of recording descriptors and other chemical properties thus offers better functionality for cheminformatics analysis. Due to the limited memory capacity for handling large compound database, several chemical line notations have also been introduced. One such format is the simplified molecular-input line-entry system (SMILES) format pioneered by Weininger et al [4]. Other linear notations include Wiswesser line notation (WLN), ROSDAL, and SYBYL Line Notation (SLN). Instead of recording compound coordinates directly, the SMILES format store compound structure using simpler ASCII codes. While memory-efficient, there is no unique strings for representing chemical compound particularly for large and structurally complex molecules. To address this, canonical SMILES was proposed that applied the Morgan algorithm for consistent labeling and ordering of chemical structures [5]. Another limitation is the loss of coordinate information and necessitate structural generation programs like PRODRG to predict native molecular geometry [6]. Recently, the need to exchange chemical data over the world wide web (WWW) also saw the development of chemical markup language (CML) similar to the XML format. Despite the development of multiple chemical file formats, many commercial and open source packages have allowed convenient file format conversion using Obabel and RDKit softwares [7, 8].

2.2 Chemical representations

The ability to represent chemical compounds by machine-learning features that fully captured wide ranges of chemical and physical properties of the target molecule has been an active area of research in chemoinformatics and chemical biology

[9, 10]. These chemical features, also known as chemical descriptors, provide the ability to extract essential characteristic of the compound and offer the possibility of developing predictor that can classify novel structures with similar properties. Broadly speaking, the chemical descriptors can be classified as 0D, 1D, 2D, 3D, and 4D [11]. 0D and 1D descriptors like molecular mass, atom number counts can be easily extracted from the molecular formula but does not provide much discriminatory power for compound classification. In practice, 2D and 3D chemical descriptors are the most commonly used molecular features for cheminformatics analysis [12]. Since chemical compound can be viewed as different arrangements of atoms and chemical bond, 2D descriptors can be generated from the molecular graph based on different connectivity of the molecules. Notable 2D descriptors include Weiner index, Balaban index, Randic index and others [1]. Beyond 2D descriptors, 3D descriptors leverage information from molecular surfaces, volumes, and shapes to provide a higher level of chemical representation. The dependency of ligand conformations also prompts the development of 4D descriptors, which accounts for different conformations of the molecules generated over a trajectory from the molecular dynamics simulation [13]. However, the requirement of correct 3D conformation makes 3D and 4D descriptors limited in several aspects. Another type of high dimensional descriptors is molecular interaction field (MIF) developed by Goodford and colleagues [14]. The MIF aims to capture the molecular environment of the ligand based on several properties by placing probes in a rectangular grid surround the target compound. At each grid point, hypothetical probes corresponding to different types of energetic interactions (hydrophobic, electrostatic) were evaluated. The comparison of MIF of compounds enables the identification of critical functional groups for kinase drug-target interactions and drug design [15]. Furthermore, correlating these field values to compound activity enable comparative molecular field analysis (CoMFA), an extended form of 3D-QSAR [16]. Altman's group at Stanford University took a different approach by inspecting ligand environment using amino acid microenvironment. This Feature-based approach lead to direct applications in pocket similarity comparison for identifying novel microtubule binding activity of several anti-estrogenic compounds as well as kinase off-target binding activity [17, 18]. Chemical descriptors can likewise be generated based on the biological phenotypes. For example, drug-induced cell cycle profile changes of compound have been recently utilized to identify DNA-targeting properties of several microtubule destabilizing agents [19].

Besides chemical descriptors, the chemical fingerprint is another important chemical representation where the compounds are represented by a binary vector indicating the presence or absence of chemical features [20]. Common 2D chemical fingerprints include path-based fingerprint which detected all possible linear paths consisting of bonds and atoms of a structure given certain bond lengths. For a given pattern, several bits in a bit string is set. While path-based fingerprints like ECFP (Extended Connectivity Fingerprint) have a higher specificity, the potential limitation is "bit collision" where the number of possible patterns exceeds the bit capacity resulting in multiple patterns mapped to the same set of bits. Another type of fingerprint is substructure fingerprints. In the substructure fingerprint like (Molecular ACCess System) MACCS keys, the substructures are predefined and each bit in a bit string is set for specific chemical patterns. Although bit collision is less of an issue, the requirement to encompass all fragment space within a bit string often demands a larger memory size. Recently, the proposal of circular fingerprints represents the state-of-the-art in chemical fingerprint development [21]. In the circular fingerprint, each layer's feature is constructed by applying a fixed hash function to the concatenated features of the neighborhood in the previous layer and the results from the hashed function were mapped to bit string representing

specific substructures. A modified version of the circular fingerprint, known as graph convolution fingerprint, has recently been proposed where the hashed function is replaced by a differential neural network and a local filter is applied to each atom and neighborhoods similar to that of a convolution neural network. Many of the mentioned fingerprints has been implemented by several open source chemoinformatics package such as Chemoinformatics Development Kit (CDK) and RDKit and saw wide applications in compound database search and other computer-aided drug discovery tasks [22].

3. Artificial intelligence in drug discovery

The rise of artificial intelligence and, in particular, machine learning and deep learning has given rise to a tsunami of applications in drug discovery and design [23, 24]. Here, we provide an overview of machine learning concepts and techniques commonly applied for chemoinformatics analysis. In a nutshell, machine learning aims to build predictive models based on several features derived from the chemical data, many of which are measured experimentally, such as lipophilicity, water solubility while others are purely theoretical, such as chemical descriptors and molecular fields derived from the chemical graph or 3D structure data. With chemical features on one hand, on the other hand of the equation is the properties that the model intended to learn, which can take on categorical or continuous values and usually pertaining to compound activity in question. Given every pair of features and labels, the model can be trained by identifying an optimal set of parameters that minimizes certain objective functions. Following the training phase, the best model can then be applied to predict the properties of new compounds (**Figure 1**).

Although machine learning has just recently gained in popularity, its application in chemistry is not new. The pioneering work of Alexander Crum-Brown and Thomas Fraser in elucidating the effects of different alkaloids on muscle paralysis results in the proposal of the first general equation for a structure–activity relationship, which intended to bridge biological activity as a function of chemical structure [25]. Early QSAR models such as Hansch analysis were mostly linear or quadratic model of physicochemical parameters that required extensive experimental measurement. This model was succeeded by the Free-Wilson model, which considers the parameters generated from the chemical structure and is more closely resemble the QSAR model in use today. Machine learning techniques in cheminformatics analysis can be broadly classified as supervised learning, unsupervised learning, and reinforcement learning. However, new learning algorithms through a combination of these approaches are continuing being developed. Many of these approaches have already found wide application in QSAR/QSPR prediction, de novo drug design, drug repurposing, and retrosynthetic planning [26–28].

3.1 Supervised learning

3.1.1 Linear regression analysis

Supervised learning has a long history of development in QSAR analysis [29]. The supervised learning task can include classification, to determine whether a compound class belong to a certain class label, or regression, to predict the bioactivity of a compound over a continuous range of values. A well-known supervised learning approach is the linear regression model, and often the first-line method for exploratory data analysis among statistician. The goal of linear regression is to find

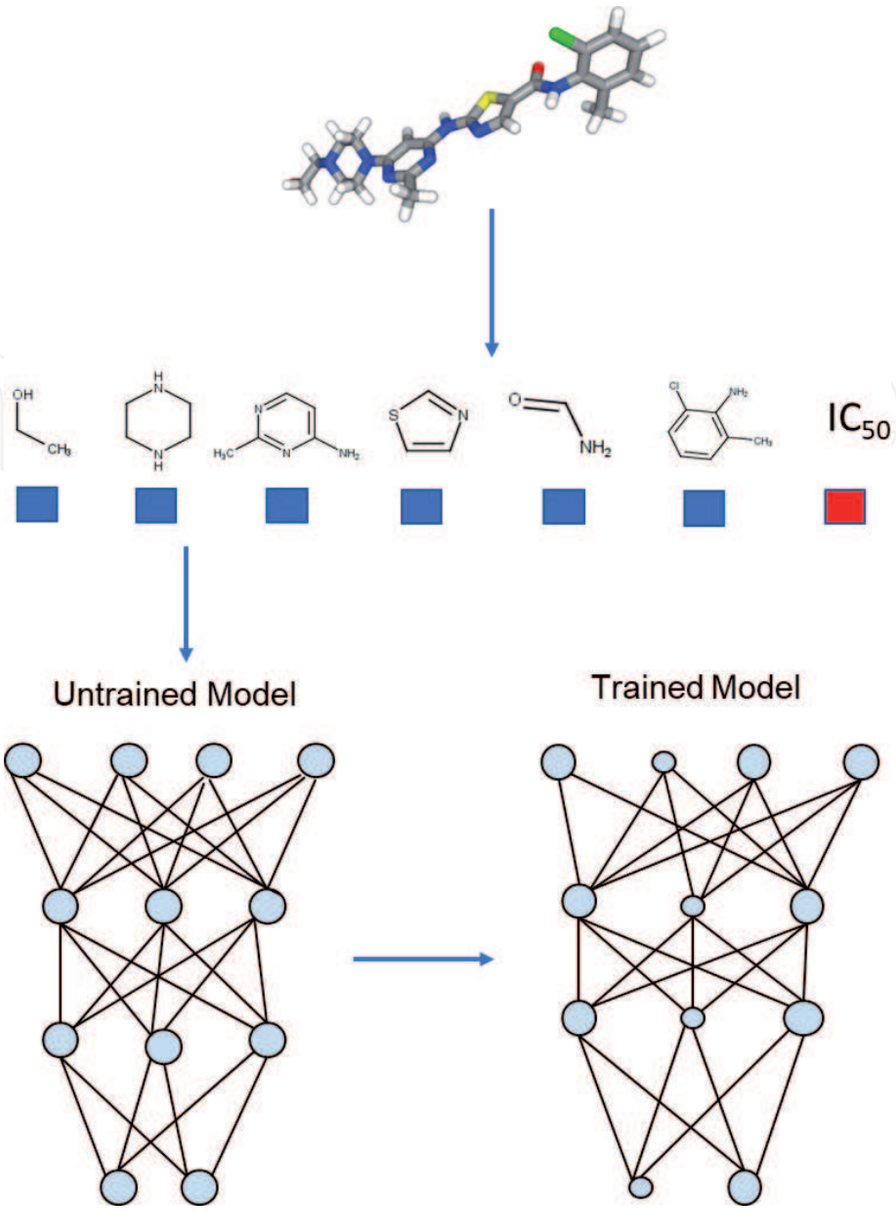

Figure 1.
Chemoinformatics prediction using artificial intelligence. Starting with a compound, the chemical feature is extracted from the compound 2D graph. The chemical features then serve as input for the machine learning model and trained based on the compound activity. The trained model with fitted parameters can then be used to predict activity of new compounds.

a linear function such that a fitted line that minimizes the distance to the outcome variables. When the logistic function is applied to the linear model, the model can also be applicable for binary classification. A direct extension of linear regression is polynomial regression that model relationships between independent and independent variable as high-degree polynomial of the same or different combination of chemical features. In the case of model underfitting, polynomial regression provides a useful alternative for feature augmentation for the linear model. Both linear and polynomial regression formed the basis of classical Hansch and Free-Wilson analysis [30]. Interestingly, today's situation is completely reversed. With the rapid explosion of chemical descriptors and fingerprints available at chemoinformatician's disposal, twin curse of dimensionality and collinearity has now become a significant issue.

Several approaches have been developed to tackle high dimensional data. One potential solution is to exhaustively explore all the possible combination of features to identify the best subset of predictors. However, this approach is inevitably

computationally infeasible for large feature space. To solve this, heuristic approach like forward and backward feature selection were developed where each feature was added to the predictors in a stepwise manner and only features that contribute greatest to the fit are kept [31]. An alternative approach for feature selection is dimensional reduction where a smaller set of uncorrelated features can be created as a combination of a larger set of correlated variables. One commonly used dimensional reduction technique is principal component analysis (PCA) that identifies new variables with the largest variances in the dataset [32]. Recently, variable shrinkage method like regularization and evolutionary algorithm has allowed feature selection during the model fitting phase. In the model regularization step, a penalty term is introduced to the objective function to control model complexity. The lasso regularization is one such approach that used an L1 penalty term to constraint objective function along the parameter axis, thus enable effective elimination of redundant features [33]. The evolutionary algorithm is another feature selection approach that encodes features as genes and through successive combination, the algorithm identifies the best set of features measured by a fitness score. Recently, elastic net combines penalties of the lasso and ridge regression and shows promise in variable selection when the number of predictors (p) is much bigger than the number of observations (n) [34]. Although linear regression analysis formed the backbone of early QSAR analysis, the simple linear assumption of feature vector space is a major limitation for modeling more complex system.

3.1.2 Artificial neural network and deep learning

The requirement to parameterize the QSAR model in a non-linear way saw the widespread application of artificial neural network (ANN) in the chemoinformatic analysis. The ANN, first developed by Bernard Widrow of Stanford University in the 1950s, is inspired by the architecture of a human brain, which consisting of multiple layers of interconnecting nodes analogous to biological neurons. The early neural network model is called "perceptron" that consists of a single layer of inputs and a single layer of output neurons connected by different weights and activation functions [35]. However, it was soon recognized that the one-layer perceptron cannot correctly solve the XOR logical relationship [36]. This limitation prompts the development of multi-layer perceptron, where additional hidden layers were introduced into the model and the weights were estimated using the backpropagation algorithm [37]. As a direct extension of ANN, several deep learning techniques like deep neural network (DNN) has been introduced to process high dimensional data as well as unstructured data for machine vision and natural language processing (NLP). In multiple studies, DNN outperformed several classical machine learning methods in predicting biological activity, solubility, ADMET properties and compound toxicity [38, 39].

To handle high-dimensional data, several feature extraction and dimension reduction mechanisms has been integrated into diverse deep learning frameworks (**Figure 2**). In particular, the convolution neural network is a popular deep learning framework for imaging analysis [40]. A convolution neural network consists of convolution layers, max-pooling layers, and fully connected multilayer perceptron. The purpose of the convolution and max-pooling layer is to extracted local recurring patterns from the image data to fit the input dimension of the fully connected layers. This utility has recently been extended for protein structure analysis in the 3D-CNN approach where protein structures are treated as 3D images [41]. Other deep learning approaches include autoencoder and embedding representation. Autoencoder (AE) is a data-driven approach to obtain a latent presentation of high dimensional data using a smaller set of hidden neurons [42, 43]. An autoencoder

consists of encoder and decoder. In the encoding step, the input signal is forward propagated to smaller and smaller sets of hidden layers thus effective map the data to low dimensional space. The training is achieved so that the hidden layers can propagate back to a larger set of output nodes to recover the original signal. A specific form of AE called variational AE (VAE) has recently been applied to de-novo drug design application where latent space was first constructed from the ZINC database from which novel compounds can be recovered by sampling such subspace [44]. In the context of NLP, word embedding such as word2vec implementation is a dimensional reduction technique to learn word presentation that preserves the similarity between data in low-dimension. This formulation has been extended to identify chemical representation in the analogous mol2vec program [45]. The requirement to model sequential data also prompted the development of recurrent neural networks (RNN). The RNN is a variant of artificial neural network where the output from the previous state is used as input for the current state. Therefore, this formulation has a classical analogy to the hidden Markov model (HMM), a type of belief network. RNN has been applied for de novo molecule design by "memorizing" from SMILES string in sequential order and generated novel SMILES by sampling from the underlying probability distribution [46]. By tuning the sampling parameters, it is found that RNN can oftentimes generated valid SMILES string not found in the original training set.

3.1.3 Instance-based learning

In contrast to parametrized learning that required extensive efforts in model tuning and parameter estimation, instance-based learning, also known as memory-based learning, is a different type of machine learning strategy that generates hypothesis from the training data directly [47]. Therefore, the model complexity

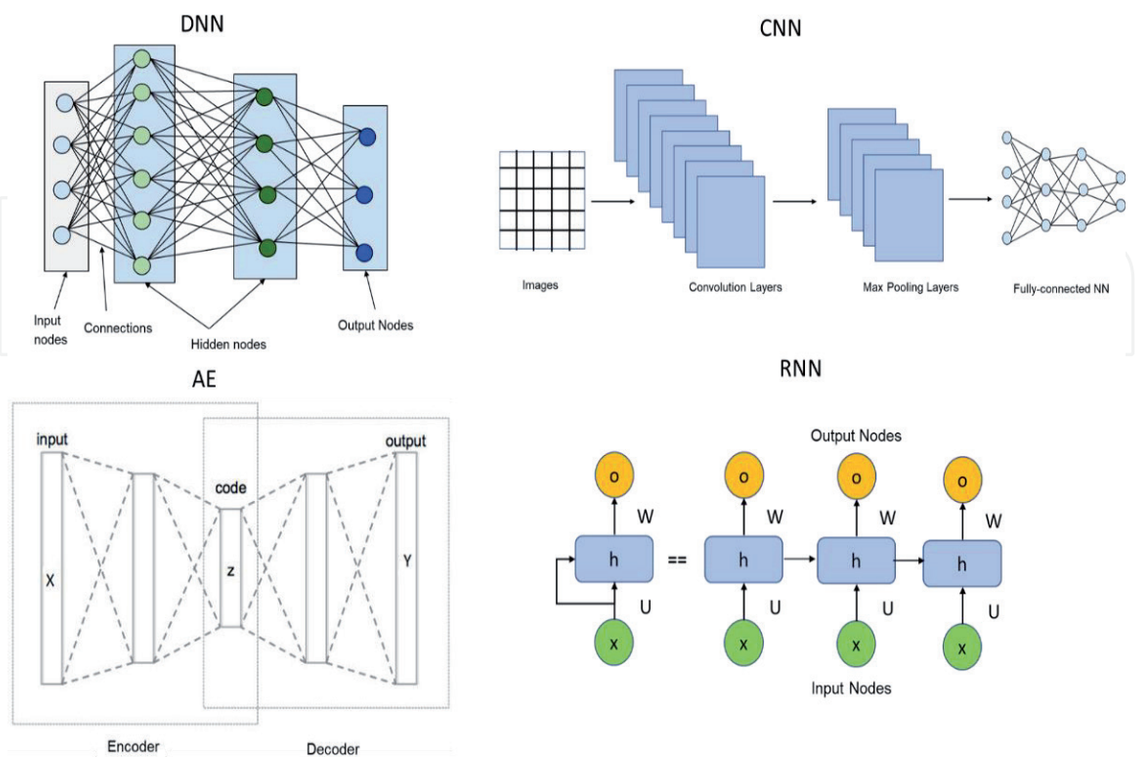

Figure 2.
Deep learning architectures for drug discovery. Four common types of deep learning network for supervised and supervised learning including deep neural network (DNN), convolutional neural network (CNN), autoencoder (AE) and recurrent neural network (RNN).

is highly dependent on the size and quality of the dataset. Notable instance-based learning method includes the k-Nearest Neighbor (kNN) prediction, commonly known as "guilt-by-association" or "like-predicts-like". In the kNN algorithm, a majority voting rule is applied to predict the properties of a given data, based on the k nearest neighbor within certain metric distance [48]. Using this approach, the properties of the data can be inferred from the dominant properties shared among its nearest neighbors. In the field cheminformatics, chemical similarity principle is a direct application of kNN where the similarity between chemical structures can be used to infer similar biological activity [49]. For analyzing large compound set, chemical similarity networks, or chemical space networks, can be used to identify chemical subtypes and estimate chemical diversity [50, 51]. Furthermore, the similarity concept is commonly applied in computational chemical database search to identify similar compounds from a lead series [52]. A major limitation of kNN is the correct determination of the number of nearest neighbors since that too high or low of such parameter can lead to either high false positive and false negative rates.

In the case of binary classification, such as compound activity discrimination, support vector machine (SVM) is a popular non-parametrized machine learning model [53]. For given binary data labels, SVM intended to find a hyperplane such that it has the largest distance (margin) to the nearest training data point of two classes. Furthermore, kernel trick allows mapping data points to high dimensional feature space that are linearly inseparable. For multilabel classification problems, other instance-learning models such as radial basis neural network (RBNN), decision trees and Bayesian learning are generally applicable [54]. In RBNN, several radial basis functions, which often depict as bell shape regions over the feature space, are used to approximate the distribution of the data set. Other approaches like decision tree, such as the Classification And Regression Tree (CART) algorithm, can also be applied for multi-variable classification and regression and has been used to differentiate active estrogen compound from inactives [55]. In the decision tree model, the algorithm provides explanations for the observed pattern by identifying predictors that maximize the homogeneity of the dataset through successive binary partitions (splits). The Bayesian classifier is yet another powerful supervised learning approach that predicts future events based on past observations known as prior. In essence, Bayes' theorem allows the incorporation of prior probability distributions to generate posterior probabilities. In the case of multi-variable classification, a special form of Bayesian learner known as the naïve Bayes learner greatly simplify the computational complexity with independence assumption between features. PASS Online is an example of a Bayesian approach to predict over 4000 kinds of biological activity, including pharmacological effects, mechanisms of action, toxic and adverse effects [56]. In another study, DRABAL, a novel multiple label classification method that incorporates structure learning of a Bayesian network, was developed for processing more than 1.4 million interactions of over 400,000 compounds and analyze the existing relationships between five large HTS assays from the PubChem BioAssay Database [57].

While instance-based learning encompasses a diverse set of methodology and present unique advantages in constantly adapting to new data, this approach is nevertheless limited by the memory storage requirement and, as the dataset grows, data navigation becomes increasingly inefficient. To address this, data pre-segmentation technique such as KD tree is a common approach for instance reduction and memory complexity improvement [58]. In another aspect, the ability to assemble different classifiers into a meta-classifier that will potentially have superior generalization performance than individual classifier also led to the development of ensemble learning. The ensemble learning algorithm can include models that combine multiple types of classifier or sub-sample data from a single

model. A notable example of ensemble learning is the random forest algorithm, which combines multiple decision trees and makes predictions via a majority voting rule for compound activity classification and QSAR modeling [59].

3.2 Unsupervised learning

Given a compound dataset, unsupervised learning can include tasks such as detecting subpopulation to determine the number of chemotypes to estimate chemical diversity and chemical space visualization. Putting in a broader perspective, the purpose of unsupervised learning is to understand the underlying pattern of the datasets. Another important problem stem from unsupervised learning is the ability to define appropriate metrics that can be used to quantify the similarity of data distributed over feature space. These metrics can be useful for chemometrics application including measuring the similarity between pairs of compounds.

3.2.1 Clustering

For unsupervised clustering, one popular approach is K-means clustering [60]. K-means clustering aims to partition the dataset into K-centroid. This is achieved by constantly minimizing the within-cluster distances and updating new centroids until the location of the K-centroids converges. K-means clustering has the advantage of operating at linear time but does not guarantee convergence to a global minimum. Another limitation is the requirement of a pre-determined number of clusters, which may not correspond to the optimal clusters for the data. To identify the optimal k values, one solution is called the "elbow method", which determine a k value with the largest change in the sum of distances as the k value increases. One study applied K-means clustering to estimate the diversity of compounds that inhibit cytochrome 3A4 activity [61]. Besides K-mean clustering, conventional clustering like hierarchical clustering is also commonly used. Hierarchical clustering can include agglomerative clustering, which merges smaller data objects to form larger clusters or divisive clustering, which generate smaller clusters by splitting from a large cluster. The hierarchical clustering has been demonstrated for their ability to classify large compound and enrich ICE inhibitors from specific clusters as well as for virtual screening application [62, 63].

Although hierarchical clustering is suitable for initial exploratory analysis, it is limited by several shortcomings such as high space and time complexity and lack of robustness to noise. Supervised clustering using artificial networks include the self-organization map (SOM), also known as Kohonen network [64]. The purpose of SOM is to transform the input signal into a two-dimensional map (topological map) where input features that are similar to each other are mapped to similar regions of the map. The learning algorithm is achieved by competitive learning through a discriminant function that determines the closest (winning) neuron. During each training iteration, the winning neuron has its weight updated such that it moves closer to the corresponding input vector until the position of each neuron converges. The advantages of SOM are the ability to directly visualize the high-dimensional data on low dimensional grid. Furthermore, the neural network makes SOM more robust to the noisy data and reduces the time complexity to the linear range. SOMs cover such diverse fields of drug discovery as screening library design, scaffold-hopping, and repurposing [65].

Recently, manifold learning has gained tremendous traction due to the ability to perform dimensional reduction while preserving inter-point distances in lower dimension space for large-scale data visualization. Manifold learning algorithm includes ISOMAP, which build a sparse graph for high dimensional data and

identify the shortest distance that best preserves the original distance matrix in low dimensional space [66]. While ISOMAP requires very few parameters, the approach is nevertheless computational expensive due to an expensive dense matrix eigen-reduction process. More efficient approaches such as Locally Linear Embedding (LLE) has been proposed for QSAR analysis [67]. LLE assumes that the high dimensional structure can be approximated by a linear structure that preserves the local relationship with neighbors. A related approach is t-distributed stochastic neighbor embedding (tSNE), which relies on the pair-wise probability distribution of data points to preserve local distance [68].

3.2.2 Similarity

The ability to measure data similarity is as important as the ability to discern the number of categories from a dataset. One approach for measuring data similarity is by determining the distance of two data points in the high-dimensional feature space. Intuitively, the similarity between two data points is inversely related to the measured distance between them. Commonly used distance metrics include Euclidean distance, Manhattan distance, Chebyshev distance [60]. All of these metrics is a specialized form of Minkowski distance, a generalized distance metrics defined in the norm space. Other important similarity measures such as the cosine similarity and Pearson's correlation coefficient, are commonly used to measure gene expression data or word embedding vector, when the magnitude of the vector is not essential. For binary features, metrics that measured shared bits between vectors can be used. For example, Tanimoto index, also known as the Jaccard coefficient, is one of the most commonly used metrics to measuring the similarity between two fingerprints in many cheminformatics applications. Tanimoto index has been extended to measure the similarity of 3D molecular volume and pharmacophore, such as those generated from the ligand structural alignment [69]. A generalized form of similarity metric is the kernel such as RBF or Gaussian kernel, which is a function that maps a pair of input vectors to high dimensional space and is an effective approach to tackle non-linearly separable case for discriminating analysis. The selection of an optimal similarity metrics can be achieved by clustering analysis, including comparing the clustering result and assess the quality of the clusters by different similarity measures.

3.3 Reinforcement learning

Reinforcement Learning came into the spotlight from the famous chess competition between professional chess player and AlphaGo that demonstrated the ability of AI to outcompete human intelligence [70]. Differ from supervised and unsupervised learning, the reinforcement learning focused on optimization of rewards and the output is dependent on the sequence of input. A basic reinforcement learning is modeled based on the Markov decision process and consists of a set of environment and agent state, a set of actions and transitional probability between states. At each time step, the agent interacts with the environment with a chosen action and a given reward. Several learning strategies have been developed to guide the action in each state. The most well-known algorithm is called the Q-learning algorithm [71]. The Q-learning predicts an expected reward of an action in a given state and as the agent interacts with the environment, the Q value function becomes progressively better at approximate the value of an action in a given state. Another approach for guiding the action for reinforcement learning is called policy learning, which aims to create a map that suggests the best action for a given state. The policy can be constructed using a deep neural network. Recently, deep Q-network (DQN) has been

constructed that approximate the Q value-functions using a deep neural network [72]. One recent example of using deep reinforcement learning in de novo design is demonstrated by the ReLeaSE (Reinforcement Learning for Structural Evolution), which integrates both predictive and generative model for targeted library design based on SMILES string. The generative model is used to generate chemically feasible compound while the predictive model is then used to forecast the desired properties. The ReLeaSE method can be used to design chemical libraries with a bias toward structural complexity or toward compounds with a specific range of physical properties as well as inhibitory activity against Janus protein kinase 2 [73].

4. Conclusion

The path of drug discovery from small molecule ligand to drug that can be utilized clinically is a long and arduous process. The fundamental concept of artificial intelligence and the application in drug design and discovery presented will facilitate this process. In particular, the machine learning and deep learning, which demonstrated great utility in many branches of computer-aided drug discovery like de novo drug design, QSAR analysis, chemical space visualization.

In this chapter, we presented the fundamental concept of artificial intelligence and their application in drug design and discovery. We first focused on chemoinformatics, a broad field that studying the application of computers in storing, processing, and analyzing chemical data. This field already has more than 30 years of development with focuses on subjects ranging from chemical representation, chemical descriptors analysis, library design, QSAR analysis, and retrosynthetic planning. We then discussed how artificial intelligence techniques can be leveraged for developing more effective chemoinformatics pipelines and presented with real-world case studies. From the algorithmic aspects, we mentioned three major class of machine learning algorithms including supervised learning, unsupervised learning, and reinforcement learning, each with their own strength and weakness as well as cover different areas of chemoinformatic applications.

As AI techniques gradually become indispensable tools for drug designer to solve their day-to-day problems, an emerging trend is to learn how to flexibly integrate these algorithms in the computational pipelines suitable for the problem at hand. For example, the process can start with an unsupervised learning to discerning the number of chemotypes followed by a supervised learning approach to predict multi-target activities. Furthermore, with the increasing computational power, deep learning network with increasing number layers and complexity will be also developed. Another potential development is the marriage between chemical big data and AI to mine the chemical "universe" for drug screening applications. The potential extensibility of AI in drug discovery and design is virtually boundless and awaits drug designer to further explore this exciting field.

Author details

Yu-Chen Lo[1,3*], Gui Ren[2], Hiroshi Honda[2] and Kara L. Davis[3]

1 Bioengineering, Stanford University, Stanford, CA, USA

2 Bioengineering, Northwestern Polytechnic University, Fremont, CA, USA

3 Pediatrics, Bass Center for Childhood Cancer, Stanford School of Medicine, Stanford, CA, USA

*Address all correspondence to: bennylo@stanford.edu

References

[1] Lo YC, Rensi SE, Torng W, Altman RB. Machine learning in chemoinformatics and drug discovery. Drug Discovery Today. 2018;**23**(8):1538-1546. DOI: 10.1016/j.drudis.2018.05.010

[2] Idakwo G, Luttrell J, Chen M, Hong H, Zhou Z, Gong P, et al. A review on machine learning methods for in silico toxicity prediction. Journal of Environmental Science and Health. Part C, Environmental Carcinogenesis & Ecotoxicology Reviews. 2018;**36**(4):169-191. DOI: 10.1080/10590501.2018.1537118

[3] Gasteiger J. Chemoinformatics: A new field with a long tradition. Analytical and Bioanalytical Chemistry. 2006;**384**(1):57-64. DOI: 10.1007/s00216-005-0065-y

[4] Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. Journal of Chemical Information and Computer Sciences. 1988;**28**(1):31-36. DOI: 10.1021/ci00057a005

[5] O'Boyle NM. Towards a universal SMILES representation—A standard method to generate canonical SMILES based on the InChI. Journal of Cheminformatics. 2012;**4**(1):22. DOI: 10.1186/1758-2946-4-22

[6] Schuttelkopf AW, van Aalten DM. PRODRG: A tool for high-throughput crystallography of protein-ligand complexes. Acta Crystallographica. Section D, Biological Crystallography. 2004;**60**(Pt 8):1355-1363. DOI: 10.1107/S0907444904011679

[7] O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open babel: An open chemical toolbox. Journal of Cheminformatics. 2011;**3**:33. DOI: 10.1186/1758-2946-3-33

[8] Lovric M, Molero JM, Kern R. PySpark and RDKit: Moving towards big data in cheminformatics. Molecular Informatics. 2019;**38**(6):e1800082. DOI: 10.1002/minf.201800082

[9] Gupta A, Kumar V, Aparoy P. Role of topological, electronic, geometrical, constitutional and quantum chemical based descriptors in QSAR: mPGES-1 as a case study. Current Topics in Medicinal Chemistry. 2018;**18**(13):1075-1090. DOI: 10.2174/1568026618666180719164149

[10] Haggarty SJ, Clemons PA, Wong JC, Schreiber SL. Mapping chemical space using molecular descriptors and chemical genetics: Deacetylase inhibitors. Combinatorial Chemistry & High Throughput Screening. 2004;**7**(7):669-676

[11] Sykora VJ, Leahy DE. Chemical descriptors library (CDL): A generic, open source software library for chemical informatics. Journal of Chemical Information and Modeling. 2008;**48**(10):1931-1942. DOI: 10.1021/ci800135h

[12] Nettles JH, Jenkins JL, Bender A, Deng Z, Davies JW, Glick M. Bridging chemical and biological space: "Target fishing" using 2D and 3D molecular descriptors. Journal of Medicinal Chemistry. 2006;**49**(23):6802-6810. DOI: 10.1021/jm060902w

[13] Pan D, Tseng Y, Hopfinger AJ. Quantitative structure-based design: Formalism and application of receptor-dependent RD-4D-QSAR analysis to a set of glucose analogue inhibitors of glycogen phosphorylase. Journal of Chemical Information and Computer Sciences. 2003;**43**(5):1591-1607. DOI: 10.1021/ci0340714

[14] Goodford PJ. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. Journal of Medicinal Chemistry. 1985;**28**(7):849-857. DOI: 10.1021/jm00145a002

[15] Naumann T, Matter H. Structural classification of protein kinases using 3D molecular interaction field analysis of their ligand binding sites: Target family landscapes. Journal of Medicinal Chemistry. 2002;**45**(12):2366-2378. DOI: 10.1021/jm011002c

[16] Cramer RD, Patterson DE, Bunce JD. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. Journal of the American Chemical Society. 1988;**110**(18):5959-5967. DOI: 10.1021/ja00226a005

[17] Lo YC, Liu T, Morrissey KM, Kakiuchi-Kiyota S, Johnson AR, Broccatelli F, et al. Computational analysis of kinase inhibitor selectivity using structural knowledge. Bioinformatics. 2019;**35**(2):235-242. DOI: 10.1093/bioinformatics/bty582

[18] Lo YC, Cormier O, Liu T, Nettles KW, Katzenellenbogen JA, Stearns T, et al. Pocket similarity identifies selective estrogen receptor modulators as microtubule modulators at the taxane site. Nature Communications. 2019;**10**(1):1033. DOI: 10.1038/s41467-019-08965-w

[19] Lo YC, Senese S, France B, Gholkar AA, Damoiseaux R, Torres JZ. Computational cell cycle profiling of cancer cells for prioritizing FDA-approved drugs with repurposing potential. Scientific Reports. 2017;**7**(1):11261. DOI: 10.1038/s41598-017-11508-2

[20] Durant JL, Leland BA, Henry DR, Nourse JG. Reoptimization of MDL keys for use in drug discovery. Journal of Chemical Information and Computer Sciences. 2002;**42**(6):1273-1280

[21] Kearnes S, McCloskey K, Berndl M, Pande V, Riley P. Molecular graph convolutions: Moving beyond fingerprints. Journal of Computer-Aided Molecular Design. 2016;**30**(8):595-608. DOI: 10.1007/s10822-016-9938-8

[22] Willighagen EL, Mayfield JW, Alvarsson J, Berg A, Carlsson L, Jeliazkova N, et al. The chemistry development kit (CDK) v2.0: Atom typing, depiction, molecular formulas, and substructure searching. Journal of Cheminformatics. 2017;**9**(1):33. DOI: 10.1186/s13321-017-0220-4

[23] Mater AC, Coote ML. Deep learning in chemistry. Journal of Chemical Information and Modeling. 2019;**59**(6):2545-2559. DOI: 10.1021/acs.jcim.9b00266

[24] Hessler G, Baringhaus KH. Artificial intelligence in drug design. Molecules. 2018;**23**(10):E2520. DOI: 10.3390/molecules23102520

[25] Klebe G. Drug Design. New York: Springer; 2013

[26] Jordan AM. Artificial intelligence in drug design-the storm before the calm? ACS Medicinal Chemistry Letters. 2018;**9**(12):1150-1152. DOI: 10.1021/acsmedchemlett.8b00500

[27] Jing Y, Bian Y, Hu Z, Wang L, Xie XQ. Deep learning for drug design: An artificial intelligence paradigm for drug discovery in the big data era. The AAPS Journal. 2018;**20**(3):58. DOI: 10.1208/s12248-018-0210-0

[28] Roy K. In Silico Drug Design. Waltham, MA: Elsevier; 2019. p. 886

[29] Gasteiger J. Handbook of Chemoinformatics: From Data to Knowledge. Weinheim: Wiley-VCH; 2003

[30] Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M, et al. QSAR modeling: Where have you been? Where are you going to? Journal of Medicinal Chemistry. 2014;**57**(12):4977-5010. DOI: 10.1021/jm4004285

[31] Hastie T, Tibshirani R, Friedman JH. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. New York, NY: Springer; 2009. xxii. p. 745

[32] Akella LB, DeCaprio D. Cheminformatics approaches to analyze diversity in compound screening libraries. Current Opinion in Chemical Biology. 2010;**14**(3):325-330. DOI: 10.1016/j.cbpa.2010.03.017

[33] Varnek A, Baskin I. Machine learning methods for property prediction in chemoinformatics: Quo Vadis? Journal of Chemical Information and Modeling. 2012;**52**(6):1413-1437. DOI: 10.1021/ci200409x

[34] Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Series B: Statistical Methodology. 2005;**67**(2):301-320. DOI: 10.1111/j.1467-9868.2005.00503.x

[35] Widrow B, Lehr MA. 30 Years of adaptive neural networks: Perceptron, Madaline, and backpropagation. Proceedings of the IEEE. 1990;**78**(9):1415-1442. DOI: 10.1109/5.58323

[36] Minsky M, Papert S. Perceptrons; an Introduction to Computational Geometry. Cambridge, Mass: MIT Press; 1969. p. 258

[37] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986;**323**(6088):533-536. DOI: 10.1038/323533a0

[38] Korotcov A, Tkachenko V, Russo DP, Ekins S. Comparison of deep learning with multiple machine learning methods and metrics using diverse drug discovery data sets. Molecular Pharmaceutics. 2017;**14**(12):4462-4475. DOI: 10.1021/acs.molpharmaceut.7b00578

[39] Whitehead TM, Irwin BWJ, Hunt P, Segall MD, Conduit GJ. Imputation of assay bioactivity data using deep learning. Journal of Chemical Information and Modeling. 2019;**59**(3):1197-1204. DOI: 10.1021/acs.jcim.8b00768

[40] Goodfellow I, Bengio Y, Courville A. Deep Learning. Cambridge, Massachusetts: The MIT Press; 2016. xxii. p. 775

[41] Torng W, Altman RB. 3D deep convolutional neural networks for amino acid environment similarity analysis. BMC Bioinformatics. 2017;**18**(1):302. DOI: 10.1186/s12859-017-1702-0

[42] Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science. 2006;**313**(5786):504-507. DOI: 10.1126/science.1127647

[43] Xu Y, Lin K, Wang S, Wang L, Cai C, Song C, et al. Deep learning for molecular generation. Future Medicinal Chemistry. 2019;**11**(6):567-597. DOI: 10.4155/fmc-2018-0358

[44] Gomez-Bombarelli R, Wei JN, Duvenaud D, Hernandez-Lobato JM, Sanchez-Lengeling B, Sheberla D, et al. Automatic chemical design using a data-driven continuous representation of molecules. ACS Central Science. 2018;**4**(2):268-276. DOI: 10.1021/acscentsci.7b00572

[45] Jaeger S, Fulle S, Turk S. Mol2vec: Unsupervised machine learning approach with chemical intuition.

Journal of Chemical Information and Modeling. 2018;**58**(1):27-35. DOI: 10.1021/acs.jcim.7b00616

[46] Segler MHS, Kogej T, Tyrchan C, Waller MP. Generating focused molecule libraries for drug discovery with recurrent neural networks. ACS Central Science. 2018;**4**(1):120-131. DOI: 10.1021/acscentsci.7b00512

[47] Gagliardi F. Instance-based classifiers applied to medical databases: Diagnosis and knowledge extraction. Artificial Intelligence in Medicine. 2011;**52**(3):123-139. DOI: 10.1016/j.artmed.2011.04.002

[48] Asikainen AH, Ruuskanen J, Tuppurainen KA. Consensus kNN QSAR: A versatile method for predicting the estrogenic activity of organic compounds in silico. A comparative study with five estrogen receptors and a large, diverse set of ligands. Environmental Science & Technology. 2004;**38**(24):6724-6729. DOI: 10.1021/es049665h

[49] Bajorath J. Molecular similarity concepts for informatics applications. Methods in Molecular Biology. 2017;**1526**:231-245. DOI: 10.1007/978-1-4939-6613-4_13

[50] Lo YC, Senese S, Li CM, Hu Q, Huang Y, Damoiseaux R, et al. Large-scale chemical similarity networks for target profiling of compounds identified in cell-based chemical screens. PLoS Computational Biology. 2015;**11**(3):e1004153. DOI: 10.1371/journal.pcbi.1004153

[51] Kunkel C, Schober C, Oberhofer H, Reuter K. Knowledge discovery through chemical space networks: The case of organic electronics. Journal of Molecular Modeling. 2019;**25**(4):87. DOI: 10.1007/s00894-019-3950-6

[52] Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK. Relating protein pharmacology by ligand chemistry. Nature Biotechnology. 2007;**25**(2):197-206. DOI: 10.1038/nbt1284

[53] Louis B, Agrawal VK, Khadikar PV. Prediction of intrinsic solubility of generic drugs using MLR, ANN and SVM analyses. European Journal of Medicinal Chemistry. 2010;**45**(9):4018-4025. DOI: 10.1016/j.ejmech.2010.05.059

[54] Schneider G. Neural networks are useful tools for drug design. Neural Networks. 2000;**13**(1):15-16

[55] Asikainen A, Kolehmainen M, Ruuskanen J, Tuppurainen K. Structure-based classification of active and inactive estrogenic compounds by decision tree, LVQ and kNN methods. Chemosphere. 2006;**62**(4):658-673. DOI: 10.1016/j.chemosphere.2005.04.115

[56] Lagunin A, Zakharov A, Filimonov D, Poroikov V. QSAR Modelling of rat acute toxicity on the basis of PASS prediction. Molecular Informatics. 2011;**30**(2-3):241-250. DOI: 10.1002/minf.201000151

[57] Soufan O, Ba-Alawi W, Afeef M, Essack M, Kalnis P, Bajic VB. DRABAL: Novel method to mine large high-throughput screening assays using Bayesian active learning. Journal of Cheminformatics. 2016;**8**:64. DOI: 10.1186/s13321-016-0177-8

[58] Wilson DR, Martinez TR. Reduction techniques for instance-based learning algorithms. Machine Learning. 2000;**38**(3):257-286. DOI: 10.1023/A:1007626913721

[59] Svetnik V, Wang T, Tong C, Liaw A, Sheridan RP, Song Q. Boosting: An ensemble learning tool for compound classification and QSAR modeling. Journal of Chemical Information and Modeling. 2005;**45**(3):786-799. DOI: 10.1021/ci0500379

[60] Odziomek K, Rybinska A, Puzyn T. Unsupervised learning methods and similarity analysis in chemoinformatics. In: Leszczynski J, Kaczmarek-Kedziera A, Puzyn TG, Papadopoulos M, Reis HK, Shukla M, editors. Handbook of Computational Chemistry. Cham: Springer International Publishing; 2017. pp. 2095-2132

[61] Roy K, Pratim RP. Comparative chemometric modeling of cytochrome 3A4 inhibitory activity of structurally diverse compounds using stepwise MLR, FA-MLR, PLS, GFA, G/PLS and ANN techniques. European Journal of Medicinal Chemistry. 2009;**44**(7):2913-2922. DOI: 10.1016/j.ejmech.2008.12.004

[62] Bocker A, Derksen S, Schmidt E, Teckentrup A, Schneider G. A hierarchical clustering approach for large compound libraries. Journal of Chemical Information and Modeling. 2005;**45**(4):807-815. DOI: 10.1021/ci0500029

[63] Bocker A, Schneider G, Teckentrup A. NIPALSTREE: A new hierarchical clustering approach for large compound libraries and its application to virtual screening. Journal of Chemical Information and Modeling. 2006;**46**(6):2220-2229. DOI: 10.1021/ci050541d

[64] Zupan J, Gasteiger J, Zupan J. Neural Networks in Chemistry and Drug Design. 2nd ed. Weinheim; New York: Wiley-VCH; 1999. xxii. p. 380

[65] Schneider P, Tanrikulu Y, Schneider G. Self-organizing maps in drug discovery: Compound library design, scaffold-hopping, repurposing. Current Medicinal Chemistry. 2009;**16**(3):258-266

[66] Balasubramanian M, Schwartz EL. The isomap algorithm and topological stability. Science. 2002;**295**(5552):7. DOI: 10.1126/science.295.5552.7a

[67] L'Heureux PJ, Carreau J, Bengio Y, Delalleau O, Yue SY. Locally linear embedding for dimensionality reduction in QSAR. Journal of Computer-Aided Molecular Design. 2004;**18**(7-9):475-482

[68] Wallach I, Lilien R. The protein-small-molecule database, a non-redundant structural resource for the analysis of protein-ligand binding. Bioinformatics. 2009;**25**(5):615-620. DOI: 10.1093/bioinformatics/btp035

[69] Lo YC, Senese S, Damoiseaux R, Torres JZ. 3D chemical similarity networks for structure-based target prediction and scaffold hopping. ACS Chemical Biology. 2016;**11**(8):2244-2253. DOI: 10.1021/acschembio.6b00253

[70] Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, et al. Mastering the game of go with deep neural networks and tree search. Nature. 2016;**529**:484. DOI: 10.1038/nature16961. Available from: https://www.nature.com/articles/nature16961#supplementary-information

[71] Watkins CJCH, Dayan P. Q-learning. Machine Learning. 1992;**8**(3):279-292. DOI: 10.1007/BF00992698

[72] Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, et al. Human-level control through deep reinforcement learning. Nature. 2015;**518**(7540):529-533. DOI: 10.1038/nature14236

[73] Popova M, Isayev O, Tropsha A. Deep reinforcement learning for de novo drug design. Science Advances. 2018;**4**(7):eaap7885. DOI: 10.1126/sciadv.aap7885