Chapter 2
Application of Network Pharmacology Based on Artificial Intelligence Algorithms in Drug Development

Wenxia Zhou, Xuejun Li, Lu Han, and Shengjun Fan

Guide to This Chapter

The continuous development and progress of biotechnology and information technology provides data for pharmaceutical research and application. It is difficult to fully utilize large-scale data with simple statistical analysis methods. In order to improve data utilization, pharmaceutical research must be promoted using advanced information analysis. Artificial intelligence has experienced half a century of development since its inception and has been successfully applied to many industrial and technological fields. Recently, breakthroughs in machine learning represented by deep learning have made artificial intelligence one of the most popular research directions. Artificial intelligence algorithms use different types of data based on various strategies to do multiple tasks such as search and discrimination, and are suitable for solving massive data analysis problems faced in network pharmacological research. This chapter briefly introduces artificial intelligence algorithms and their applications in network pharmacology research, and provides references for researchers to better understand and apply artificial intelligence.

W. Zhou (✉) · L. Han
Beijing Institute of Pharmacology and Toxicology, Beijing, China

X. Li · S. Fan
Peking University, Beijing, China
e-mail: xjli@bjmu.edu.cn

S. Li (ed.), *Network Pharmacology*, https://doi.org/10.1007/978-981-16-0753-0_2

2.1 Introduction to Artificial Intelligence Methods in Network Pharmacology

Network pharmacology [1] is a research method based on systems biology. The concept includes recognizing and discovering drugs based on the overall relationship between an organism and drugs. In recent years, the growth of high-throughput omics data and the accumulation of pharmacological knowledge have promoted the rapid development of network pharmacology. With the accumulation of different types of data resources and knowledge bases, mining effective information like drug targets, mechanism of action, and drug and organism interaction from massive, heterogeneous data has become increasingly important in network pharmacology research. Therefore, the demand for more accurate and efficient analysis algorithms has also increased [2].

There are three common problems that may be encountered in network pharmacology research: ① Optimal solution search; ② prediction and classification; ③ automatic construction of networks and pathways. Artificial intelligence can effectively perform feature extraction and potential relationship mining from complex big data, and is beneficial for solving common problems in network pharmacology. Combining artificial intelligence and network pharmacology has great potential to overcome the problems faced in the latter field.

Since the emergence of network pharmacology research, artificial intelligence has been closely integrated with it and widely applied. For example, when the drug–target interaction is evaluated using simulation, it is necessary to perform optimal solution search operation, such as genetic algorithm [3] or simulated annealing algorithm [4], as the core of molecular docking and molecular dynamics simulation technology to implement the conformation search strategy. During network analysis and prediction, classification and prediction are required, hence unsupervised learning clustering algorithms (Affinity propagation clustering algorithm, K-means clustering algorithm) and supervised learning are widely used. In mechanism research, it is necessary to construct the network and path automatically, hence various network construction-related artificial intelligence algorithms such as the Bayesian network algorithm are often applied.

This chapter briefly reviews the development history of artificial intelligence, and the classification and characteristics of the main algorithms applied in network pharmacology, in order to promote the better understanding of the applications and evaluation methods for researchers.

2.1.1 Introduction to Artificial Intelligence Algorithms

Artificial intelligence is an important branch of computer science. The definition of artificial intelligence has not yet been unified, but it can be summarized as studying the laws of human intelligence activities and constructing artificial systems with

certain intelligent behavior [5]. Thanks to high-performance scale computing equipment, big data accumulation, and algorithm innovation, artificial intelligence has been widely applied in image recognition [6, 7], speech recognition [8, 9], medical diagnosis [10], drug R&D [11], and many other fields, and its achievements cover all aspects of human life. Artificial intelligence algorithms that widely used in network pharmacology can be divided into three types: heuristic algorithms, machine learning, and network construction algorithms according to their problem-solving scope and application characteristics.

1. Introduction to Heuristic Algorithms

Heuristic algorithms are based on intuitive or empirically constructed algorithms that give feasible solutions to problems in acceptable time and space. Its classic algorithms include: simulated annealing algorithm [4], genetic algorithm [3], etc. Heuristic algorithms perform optimal solution search with limited computational cost and time. The optimal solution search often be applicable to specific problems such as sub-network, optimal conformation, and specific sequence search.

Network pharmacology problems using heuristic algorithms usually have two basic characteristics. First, the search results can be measured by quantitative index; second, the search target can be constructed in a certain way. Taking optimal conformation as an example, the change in binding free energy is used as the quantitative index, and new binding conformations can be constructed through operations such as translation and rotation of chemical bonds and atoms in molecules.

2. Introduction to Machine Learning

Machine learning is currently the most rapidly developing artificial intelligence algorithm. For large and high-dimensional complex data, machine learning method can effectively perform data classification, data fitting, prediction model establishment, feature selection, and other tasks.

Supervised machine learning methods mainly include two categories: regression and classification [12], by which the mapping relationship could be established from input X_i to output Y_i from a large amount of input data, to construct a prediction model or analyze the weight of input features. The commonly used regression algorithms include: LASSO (Least Absolute Shrinkage and Selection Operator) regression, ridge regression, and elastic net. Classification algorithms include logistic regression, Bayesian classifier algorithm, support vector machine, K-nearest neighbor, random forest, and artificial neural network. Additionally, deep learning [13] is a rapidly developing supervised learning method in recent years, which is an improvement of the artificial neural network structure. It is characterized by more hidden layer structures between the input and output layers. Its classic structure includes: Convolutional neural networks (CNN) and recurrent neural network (RNN).

Unsupervised machine learning methods include clustering [14] and dimension reduction [15], which do not rely on input data labels to establish the feature-to-label mapping, but focus on the characteristics and interrelationships of a large amount of

data. Based on various measurement relations, the input data is divided into different categories (clustering), or the dimension of input feature vector is reduced, to remove noise and reduce redundant features (dimension reduction). Commonly used clustering algorithms include K-means clustering algorithm, hierarchical clustering, and affinity propagation clustering algorithm. Commonly used dimension reduction algorithms include principal component analysis (PCA) and factor analysis.

3. Introduction to Network Generation Method

Network generation method can be divided into network construction and sub-net extraction methods based on new network connection relationships and their generation.

In network pharmacology research, the network nodes are composed of elements related to Drug property such as compounds, targets, genes, and diseases. Networks related to biological processes are usually the most complex. For example, gene expression regulation is a dynamic process involving time and space factors. Static networks often cannot effectively reflect the temporal and spatial specificity of biological processes [16]. However, to achieve a relatively accurate characterization of the dynamic regulation of biological networks, a large amount of data with temporal and spatial differences is required. Therefore, limited data volume, and uncertain knowledge expression and reasoning can be used to make predictions and generate new network connection relations. Network construction methods include association, Boolean model, dynamic Bayesian network, and differential equation.

The sub-net extraction method does not aim at discovering new network relationships, but can extract the most relevant sub-nets from the known background network, and is often used to explain the effects of drugs or disease mechanisms. Extracting key sub-networks and identifying overlapping networks from complex relationships are important components of network analysis. Identifying key sub-networks is often closely related to the discovery of drug targets, and identification of pathways and key regulatory factors. Heuristic algorithms such as simulated annealing algorithm, genetic algorithm, and Steiner's forest algorithm [17] are often used to find the sub-net with the highest score.

2.1.2 Performance Evaluation Method for Artificial Intelligence Algorithms

Although artificial intelligence algorithms solve specific problems in network pharmacology research through a reasonable computational model, blindly trusting the computational results of artificial intelligence algorithms is detrimental. The performance of artificial intelligence algorithms to solve problems needs to be systematically evaluated by scientific metric or measures in order to effectively reduce errors caused by various risks such as low data quality or overfittings.

Table 2.1 Evaluation indicators of typical artificial intelligence methods

Artificial intelligence methods	Introduction to the methods	Performance evaluation method and evaluation index
Heuristic algorithm	Based on the specific construction algorithm, artificial intelligence is used to search an optimal solution within a certain calculation consumption. The representative algorithms include annealing algorithm, genetic algorithm, etc.	Number of iterations, convergence time, etc.
Machine learning algorithm	A class of algorithms for knowledge learning and acquisition by simulating human learning behavior is usually used for prediction and classification in pharmacological research. Representative algorithms include deep learning algorithm and clustering algorithm.	Precision rate, recall rate, ROC curve, mutual information, contour coefficient, etc.
Network generation algorithm	The method of comprehensively generating the network using multidisciplinary analysis methods such as probability theory and graph theory is mostly used in molecular network construction and drug mechanism analysis. Representative algorithms include Bayesian network algorithm and shortest path method.	Precision rate, recall rate, etc.

To evaluate the performance and generalization ability of artificial intelligence algorithms, some performance evaluation methods are required. The most well-known "Turing test" [18] is the first evaluation method proposed to gauge whether a machine is intelligent. However, it has limitations and a smaller application scope. There are several different methods or metrics could be adopted according to the algorithm and data characteristics.

Different artificial intelligence methods need to use different performance evaluation metrics and approach to evaluate the performance of the methods. General evaluation indexes include loss value, accuracy, etc., and there are also commonly used evaluation indexes for different algorithms and data characteristics. Relevant evaluation indexes are briefly summarized in Table 2.1.

1. Heuristic Algorithm Evaluation

Multiple solutions may be obtained by heuristic algorithm due to its characteristics, hence evaluation metrics could be set according to different purposes. For example, in order to save time in large-scale calculations, genetic algorithm can involve relatively few iterations and use shorter convergence time as indexes while searching for feasible solutions, whereas higher global search ability can be used as the evaluation index to get better solutions.

2. Machine Learning Algorithms Evaluation

The evaluation metrics of machine learning are applied to different algorithms, purposes, and data characteristics. The essence is to evaluate the gap between

predicted and actual values through biased functional loss, and later by optimizing the parameters. This part mainly introduces the evaluation metrics of supervised classification algorithm, regression algorithm, and clustering algorithm.

Supervised classification algorithm can divide a given object X into a predefined category Y. In supervised classification, all samples can be divided into a training set, validation set, and test set. The training and validation set data are used to train the prediction model. The trained model then uses the test set to test its accuracy and generalization ability. Additionally, k-fold cross-validation method can be used to divide the training data into two parts based on the ratio of $(k-1)/K$ and $1/K$. The former is used for model training, and the latter is used to evaluate model performance and generalization ability. The most common evaluation index in supervised classification algorithm is accuracy, to predict the proportion of accurate classification in all samples. However, due to "imbalanced data" [19] problems, evaluation indicators with characteristics such as precision and recall indicators are often used. The former is focused on the correct proportion of positive samples predicted by the classifier, while the latter is more concerned with whether it is possible to predict more positive samples. The two evaluation indexes are applicable to various scenarios. For example, when predicting effective drugs from large amount of unrelated molecules, less false positive predictions are better for researchers in order to avoid subsequent invalid biological experiments. Therefore, the accuracy rate is often used as the classification index. However, when constructing a global network regulation relationship, it is more important to cover all targets nodes, so it has greater tolerance for false positive results, and the recall rate can be used as a classification index. In addition, there are also evaluation metrics that consider both accuracy rate and recall rate, such as F1 score, receiver operating characteristic curve (ROC curve), precision-recall curve, and confusion matrix.

The regression algorithm is a statistical analysis method to determine the interdependent quantitative relationship between two or more variables. The commonly used evaluation indicators of regression algorithm include: Mean absolute deviation (MAE), root mean squared error (RMSE), mean-square error (MSE), Huber loss, log-cosh loss, etc. Using different evaluation indicators may have a greater impact on constructing prediction models. For example, "mean absolute deviation" (also known as L1 loss) is less sensitive to the output error and is relatively more stable when an abnormal point exists. At the same time, the regression model is not unique, and there may be multiple optimal solutions. Whereas, the mean-square error (also known as L2 loss) squares the output error, so the error can be optimized to a greater extent, and it is easier to obtain a stable regression model. Also, it may be more sensitive to the response of abnormal points with lower robustness.

Clustering is an important representative of unsupervised learning. They can divide samples into different categories according to similarity measures. When the sample data has a given label, a matching degree of the real label and clustering can be calculated. Mutual information, Rand index, and other indicators are commonly used. When the sample data does not have a given label, a silhouette coefficient can be used to evaluate the rationality of the clustering division.

3. Network Generation Algorithm Evaluation

In case the complete regulatory network is known, the constructed network can be compared with the complete network to calculate the precision rate, recall rate, and other indicators. The evaluation method is the same as that of the classification algorithm in machine learning.

Several network pharmacology studies using artificial intelligence use individual case verification, such as comparing model results with literature or conducting experimental verification, instead of the above evaluation indicators. This approach is usually feasible, and combined with systematic validation can be persuasive and the result can be more reliable.

2.1.3 Applications of Artificial Intelligence

Network pharmacology research involves several application requirements such as optimal solution search, target and drug prediction, and regulatory network construction. Artificial intelligence can play a key role in solving various application needs of network pharmacology. Different artificial intelligence methods can solve problems and satisfy different needs. Therefore, it is important to determine whether the algorithms suit for research problem. The following helps classify and introduce the applied fields of artificial intelligence methods.

1. Applications of Heuristic Algorithm

The main application of the commonly used heuristic algorithm is optimal solution search, which is widely used in biology and pharmacy. For example, the heuristic algorithm based tool Blast (Basic Local Alignment Search Tool) [20] is used for protein or gene sequence matching, and Open Babel [21] uses the genetic algorithm to generate small molecule conformations that are used for searching in molecular docking [22] and molecular dynamics simulation [23], heuristic algorithm is also the core algorithms in the sub-net extraction process. If the problems in network pharmacology research have the following characteristics, heuristic algorithm can be applied: ① Quantifiable scoring system: The generated results of the heuristic algorithm can judge whether calculations meet the requirements of certain scoring indicators. ② New scheme generation based on current optimal solutions: Based on the known optimal solution, a new feasible solution is generated by evaluating the distance between the calculation and the optimal solution. ③ There are corresponding convergence or termination conditions. Taking Open Babel as an example, when generating small molecule conformations, to determine whether the conformation is stable, it can either use a quantitative scoring system such as the energy of the generated conformation, or by evaluating the RMSD (Root Mean Square Deviation) coordinate deviation between the generated conformation and natural conformation.

2. Applications of Machine Learning

There are differences in the application scope and analysis between unsupervised and supervised learning methods. The purpose of unsupervised learning is to explore the relationship between input data, while supervised learning establishes mapping from input to output data from the training data, to achieve the learning purpose.

Unsupervised learning can be divided into clustering, dimension reduction, association, and other types, in which clustering and dimension reduction algorithms are widely used in network pharmacology research. Commonly used clustering algorithms include K-means clustering algorithm, AP clustering algorithm, and hierarchical clustering. Input data can be divided into various categories according to the measurement relationships. For example, Iorio et al. [24] evaluated the similarity of gene expression profiles between pairs of 1309 drugs, and used the AP clustering algorithm to construct a drug–drug similarity network for drug repurposing.

Commonly used linear dimension reduction algorithms include principal component analysis (PCA), factor analysis, etc. In the analysis of high-dimensional data, the problem of "dimension disaster" is often encountered, hence the dimension reduction algorithm is often needed to reduce the dimensionality of feature vectors, so as to reduce noise and redundant features. For example, Subramanian et al. [25] used PCA and clustering algorithm to reduce the dimension of the transcriptome data, and compressed the expression data of more than 12,000 genes to 978 landmark genes. Moreover, the 978 landmark genes can be used to infer 80% of the network regulatory relationship at the transcription level, thereby greatly reducing the cost of transcriptome data measurement.

Since linear dimension reduction algorithm often cannot meet the analytical needs when processing complex data, nonlinear dimension reduction algorithm is also widely used. For example, the t-SNE [26] algorithm, which is often used for data visualization, can retain the proximity characteristics of high-dimensional data and reduce it to two-dimensional or three-dimensional space, which plays an intuitive role in the systematic research of complex omics data [27, 28].

The supervised learning method commonly used in network pharmacology [29] includes two main types: regression and classification, both of which are used to establish the mapping relationship between input X_i and output Y_i. The output Y of regression is continuous quantitative data, such as blood pressure, blood drug concentration, while the output of classification is often qualitative data, such as negative/positive diagnosis results, tumor classification. This indicates that different types of functional losses need to be used in the calculations; however, regression and classification problems can often occur simultaneously. Supervised learning helps establish a reliable prediction model, and the model is used to predict new potential relationships.

Regression algorithms can quantitatively describe the mapping relationship between variables, so they are widely used in omics analysis and network pathway inference. For example, Gamazon et al. [30] used linear regression to infer gene expression from single nucleotide polymorphisms and predicted biological

phenotypes. Xiong and Zhou [31] used linear regression to infer the regulatory network relationship of genes from the biological experimental data level. The classification algorithm is often used in the qualitative prediction of drug–target interactions. For example, Yamanishi et al. [32] integrated multiple types of biological data (such as chemical structures, drug side effects, amino acid sequences, and protein domains), and used machine learning to train user-submitted data and to predict unknown drug–target interaction network.

Deep learning [13], as an extension of artificial neural networks, is the most rapidly developing and applied artificial intelligence algorithm in recent years. It has similar functions to traditional machine learning methods, but also has new characteristics: ① Deep neural network structure is conducive to expressing complex mapping relationships: Traditional machine learning algorithms are mostly shallow structures, hence it is difficult to display highly complex functions, whereas deep learning introduces multiple hidden layers between the input and output ends to achieve a nonlinear network structure, thus, it has the ability to express complex functions. ② Multi-hidden layer structure is capable of autonomously extracting features: Traditional machine learning algorithms rely on humans to manually extract features, while deep learning can autonomously extract features. Due to the emergence of deep structures, the input features may be transformed into new feature space, whereas the hidden layers and irrelevant features are suppressed. The above two points ensure that deep learning has better performance in processing complex big data.

3. Applications of Network Generation

Network construction is the first step in the study of network pharmacology. The commonly used methods are association, Boolean model, Bayesian network, differential equation. Artificial intelligence algorithms in network construction lay more emphasis on logical reasoning and relationship discovery, which is different from deep learning and other predictive models.

High false positive rate often occurs in the process of network construction, the complex and huge networks are not conducive to further identification of key components in the network. Therefore, it is important to extract key sub-networks from complex relationships and identify overlapping networks [34]. For example, Steiner's forest algorithm can be used to extract protein and gene–gene interaction networks from complex networks and quickly identify key interaction pathways and factors.

2.1.4 Frontiers and Prospects of Artificial Intelligence

Artificial intelligence technology has penetrated all aspects of network pharmacology research. From molecular docking, function, and target prediction, to network construction and analysis, artificial intelligence is playing an increasingly important role. On the other hand, the molecular structure of drugs, therapeutic uses, clinical

response, and multi-latitude omics data obtained from laboratory measurements constitute big data in the research field, which also brings opportunities for the application of new artificial intelligence technologies [34].

Among all types of artificial intelligence algorithms, the one with the most noticeable development in recent years is undoubtedly the deep learning algorithm [6]. Its outstanding performance in large-scale data analysis and in solving a variety of computing problems has rendered it the research frontier of artificial intelligence. In the performance evaluation and comparison of large-scale training of pharmaceutical data, deep learning surpasses traditional machine learning algorithms [35, 36]. The feature extraction ability of deep learning is convenient for analyzing complex high-dimensional data. Although it has become an emerging research direction in various industries, its application in many specific directions is still a question worth exploring.

However, the application of artificial intelligence in network pharmacology research also has corresponding technical and application problems. The most common one is over-fitting problems in the training process [37]; it is usually necessary to ensure sufficient sample amount of training data, and adopt appropriate training parameters and reliable performance evaluation methods to reduce the over-fitting problem. In addition, the deep learning algorithm also brings about the interpretability of predictive models and the computational efficiency of the big data fitting process. In order to solve these potential problems, possible future research directions include studying and understanding the function of each layer of the neural network in deep learning, optimizing deep neural network training methods to ensure efficiency and speed, introducing time and space information to achieve complex data as input, and carrying out application research.

2.2 Application of Artificial Intelligence in Network Pharmacology Research

Network pharmacology aims to promote research by using network tools. Artificial intelligence in network pharmacology plays an important role in solving drug target discovery, Drug property mechanism determination, discovery of new uses of compounds, and research on Traditional Chinese Medicine. Artificial intelligence technology is used in target discovery based on analysis methods such as structural docking, structural comparison, network simulation, and machine learning. Artificial intelligence is also used in mechanism research such as pathway and molecular function prediction and Drug property pattern analysis. In terms of discovery of new uses, artificial intelligence is used in the prediction of new uses based on multiple phenotypes and molecular data after drug perturbation. In terms of TCM research, artificial intelligence is used in the research of Chinese medicine targets, mechanisms, and syndrome theories. The following sections introduce the application status of artificial intelligence in these aspects.

2.2.1 Prediction and Discovery of Drug Targets

The discovery of drug targets is a long-standing topic in network pharmacology research. According to the strategy and data differences in the discovery of drug targets using artificial intelligence, the analysis can be divided based on ligand structure similarity and quantitative structure–activity relationship, reverse molecular docking, action network simulation, and machine learning.

1. Analysis Based on Ligand Structure Similarity and Quantitative Structure–Activity Relationship

Structural data of drugs/compounds is easily available and not only fully reflects the basic characteristics of molecules, but can also be easily counted and compared. It was used earlier in network pharmacology research. Many artificial intelligence algorithms such as intelligent search and classification are used in structural comparison analysis methods. According to the research characteristics, the analysis can be divided into structural similarity comparison method, quantitative structure–activity relationship analysis method, and docking method. They are as follows:

The importance of structural similarity mainly comes from the similar property principle [38]: molecules with similar structures may bind to the same target and have similar biological functions. By comparing the chemical similarity of ligands, it can be inferred that they may have similar targets and pharmacological effects. New pharmacological effects can be found through this method. Also, biomacromolecules (targets) with different functions may have similar drug binding domains. Therefore, the similarity between the chemical characteristics of a drug that binds to a target and the structure of the target molecule can be used to predict the unknown target of drugs [39]. Similarity measurement includes three parts: structural characterization, weight calculation, and similarity coefficient [40]. Vilart et al. [41] proposed a method to identify new DDI (Drug–Drug Interactions) based on the similarity of molecular structures of drugs involved in the established DDI. The basic assumption is that if drug A and drug B interact to produce a specific biological effect, a drug similar to drug A (or drug B) may interact with drug B (or drug A) to produce the same effect. This study collected 9454 pairs of known DDI resources, and identified DDI candidates [41] by calculating the structural similarity of all drug pairs in DrugBank. Yan et al. [42] proposed a SDTRLS (substructure-drug-target Kronecker product kernel regularized least squares) method based on sub-structure similarity, Gaussian interaction profile (GIP), similarity network fusion (SNF), RLS-Kron classifier, and other technologies. In the independent verification of G protein-coupled receptors (GPCRs), the predictions are better than in the SDTNBI algorithm (substructure-drug-target network-based inference) [42]. Keiser et al. [43] compared 3665 drugs approved by the US FDA (Food and Drug Administration) and drugs that still in the research stage, with hundreds of drug targets. By comparing the chemical similarity between the drug and the ligand set, they predicted thousands of new associations. Thirty of these associations were experimentally verified, and 23 new drug–target associations were confirmed, of which 5 have

higher binding strength with the predicted target. In addition, there is a compound N, and the physiological significance of the interaction between n-dimethyltryptamine and 5-hydroxytryptamine receptor has been verified in gene knockout mice [43].

In addition to structural similarities, quantitative structure–activity relationship (QSAR) is another commonly used research method based on structural data. It refers to a quantitative relationship that links the structural parameters of a compound with its biological activity data through a corresponding algorithm. The basic idea is that similar molecules usually bind to similar proteins. The interaction is predicted by comparing new ligands with known protein ligands [40, 44]. The predictive ability of the QSAR model depends largely on the structural similarity between the training set and the test set molecules [45]. Zhang et al. [46] used a data set of 3133 compounds to build a QSAR model. The model was built using dragon descriptors (0D, 1D, and 2D), ISIDA-2D fragment descriptors, and support vector machine (SVM) method. In the QSAR modeling and verification process, the data set is randomly divided into modeling and external evaluation sets; and the sphere exclusion algorithm is used in the training set and the test set to divide the modeling set multiple times. Then, using the consensus approach, the QSAR model is applied to the VS (virtual screening) of the ChemBridge database. The 42 inactive compounds predicted by the model have been experimentally verified [46]. Melo-Filho et al. [47] developed a continuous combi-QSAR model for the oxadiazole inhibitor data set of smTGR, and further evaluated the top 10 compounds in vitro on *Schistosoma japonicum* and adult worms, and found that two compounds containing new chemical scaffolds had high activity in various life stages of parasites at low molecular concentrations [47]. Marcelo et al. [48] combined QSAR to develop SAR rules and a binary QSAR model of antituberculosis compounds based on chalcone. Then, these models were used to conduct synthesis and biological evaluation of 33 compounds, and candidate drugs [48] with low activity to symbiotic bacteria, good selectivity to mycobacterium tuberculosis, and low cytotoxicity to Vero cells were found.

Comparison of structural similarity and QSAR is based on the hypothesis that similar structures correspond to similar activities, and molecular docking is the most intuitive application of receptor–ligand hypothesis. Molecular docking is a traditional method for evaluating the chemical complementarity of small molecules and target molecules based on the three-dimensional (3D) structure of the target. DTIs (drug target interactions) were evaluated by using a scoring function to provide a quantitative docking score associated with binding affinity [49]. Molecular docking has a wide range of applications in DTI prediction. Starting from known target proteins, screening ligands with the best affinity from many known three-dimensional structure molecules are suitable for large-scale screening of candidate ligand compounds after obtaining disease targets. Ordinarily, for one or several given targets, such as estrogen receptor [50], HIV-1 integrase [51], potential active compounds can be prioritized by molecular docking. Web applications based on molecular docking, such as TarFisDock [52], DRAR-CPI [53], rDock [54], are all built for target search based on docking. Although molecular docking is widely used, it still has its limitations, such as not being suitable for situations where the number

of proteins is large and the three-dimensional structure is not available, it cannot be applied to membrane proteins with complex structures, such as ion channels and G protein-coupled receptors (GPCRs), and the extremely low efficiency of docking computing due to the huge consumption of computing resources [45].

The key assumption of drug target analysis based on calculation of similarities is that similar drugs tend to share similar targets [29]. Thus, internationally, Yamanishi et al. [30, 31] proposed a method to predict drug target relationship by combining chemical drug similarity and genetic similarity; Keiser et al. [32] compared the chemical structure of the drug with ligands known to regulate the function of protein receptors, and obtained indirect connections between the drug and the target through these ligands. In addition, there are methods to predict drug targets based on chemical similarity [33, 34] and side effect similarity [35].

Another type of method focuses on indirect drug–gene relationships and uses additional similarity measures to obtain drug-related genes. For example, Hansen et al. [36] used the similarity of protein–protein interaction networks to predict the drug–gene genetic association, and combined the gene expression data with the drug response data provided by Kutalik et al. [37] to infer the common module relationship between genes and drugs.

In China, Cheng et al. [38] developed three supervised inference models to predict the interactions between drugs and targets, namely drug similarity inference, target-based similarity inference, and network-based inference. Li et al. [39] developed the target prediction algorithm drugCIPHER based on the overall association of "drug network–molecular network." In this method, the authors developed a computational framework called drugCIPHER, based on the interrelationships observed in the fields of pharmacology and genomics, to infer drug target interactions on a genome-wide scale. Based on the protein–protein interaction network, three linear regression models are proposed, which connect the drug treatment similarity, chemical similarity, and the correlation between the combination of the two and the target, respectively. Experiments have shown that the model (drugCIPHER-MS) that combines drug treatment similarity and chemical similarity has achieved good results on the training set and test set. The model process is shown in Fig. 2.1.

2. Reverse Molecular Docking

In recent years, with the development of computer-aided drug design, a reverse molecular docking based on the "lock-key theory" has become a new means of drug target discovery in network pharmacology [55]. For a drug or new chemical entity, reverse molecular docking works opposite to molecular docking. Small molecular compounds are used as probes to search for biomacromolecules that may be combined with them in the database of candidate targets with known structures. Possible molecular complexes can be identified using space and energy matching and potential drug targets can then be predicted [56–58].

The concept of reverse molecular docking was proposed by researcher Chen Yuzong from the National University of Singapore. Chen connected a single small molecule with multiple biological targets by means of molecular docking and by downloading the protein structure of a biomolecule in the PDB database and the

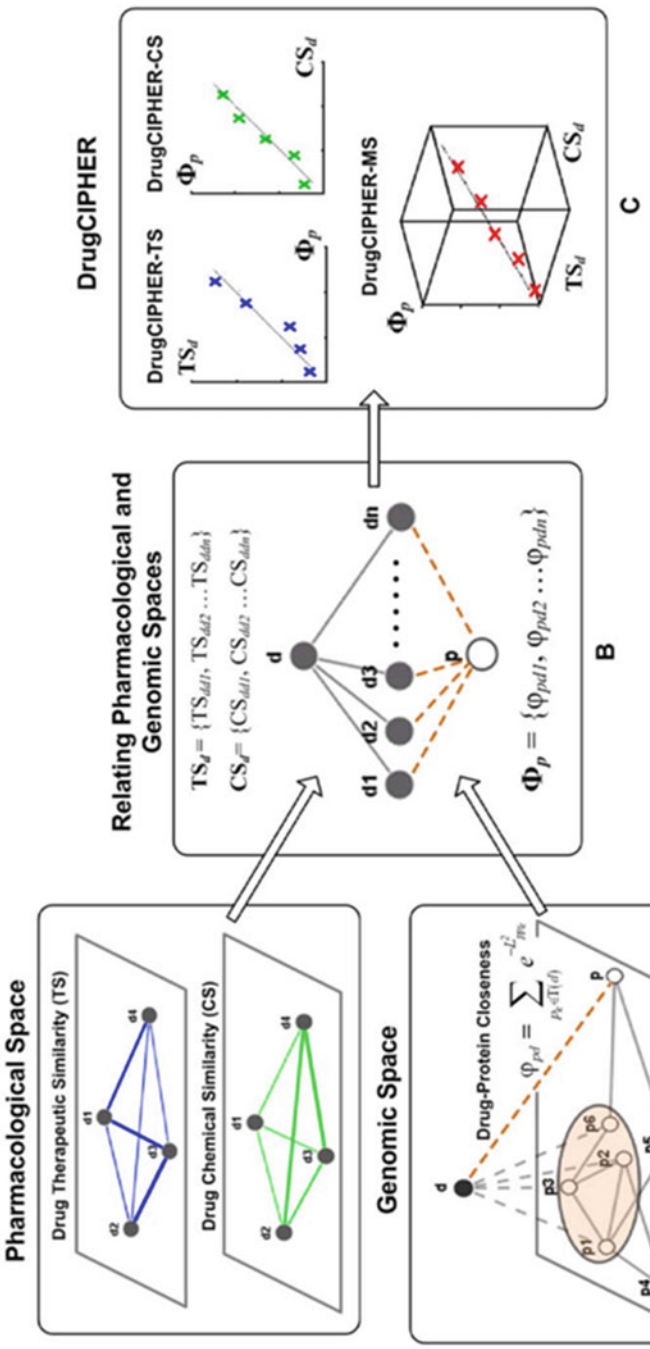

Fig. 2.1 Schematic diagram of target prediction algorithm drugCIPHER based on the overall association of "Drug Network–Molecular Network" [39]

INVDOCK platform. He then evaluated the binding energy of the ligand-compound, and preliminarily evaluated potential biological targets [59] of smaller active molecules. Subsequently, more convenient and rapidly reversing molecular docking network platforms have been developed, such as TarFisDock [60], PharmMapper [61], Reverse Screen 3D [62], and idTarget [63].

Guo et al. demonstrated that ganoderic acid D exerts an anti-cervical cancer effect [64] through the direct binding of 14-3-3 protein using bidirectional gel electrophoresis technology and INVDOCK. Subsequently, they used a similar method to clarify that the cardioprotective action of salvianolic acid B is through direct binding with human epidermal growth factor receptor (EGFR) [65]. Park et al. investigated potential biological targets of ginsenoside based on reverse molecular docking with their own protein target database, and found that dozens of biological targets such as MEK1 and EGFR could be directly regulated by ginsenoside [66].

3. Analysis Based on Action Network Simulation

The analysis based on interaction network simulation is different from the one based on structural comparison. It relies on an interaction database presented in the form of a network. Therefore, the advantage of this analysis is that it makes more extensive use of the observed interaction network to find targets. These network-based methods are usually based on algorithms in recommendation systems and relational algorithms in complex networks, which cover a larger target space and can predict potential DTIs by executing simple physical processes such as "resource diffusion," "collaborative filtering," and "random walk" on the network [67]. Topological similarity reasoning of drug target bipartite network and in vitro experiments have also been conducted. Cheng et al. [68] confirmed that five kinds of old drugs had multi-directional pharmacological properties on human estrogen receptor or dipeptidyl protease IV, and found that simvastatin and ketoconazole showed strong antiproliferative activity on human MDA-MB-231 breast cancer cell line [68]. The MD-Miner (Mechanism and Drug Miner) method proposed by Wu et al. [69] has found potentially effective drug candidates by constructing a patient-specific signal transduction network that integrates known disease-related genes with patient-derived gene expression profiles. This is based on the number of common genes between the patient-specific dysfunction signal transduction and the Drug property network, and also by a drug mechanism of action network, which integrates drug target and drug-induced expression profile data. This method has been evaluated on PC-3 prostate cancer cell line, which shows that compared to random selection, the success rate of finding effective drugs is significantly improved, and can provide in-depth understanding of potential mechanisms of action [69]. Isik et al. [70] studied whether biological responses and protein interaction networks of drug interference with cancer cells could reveal drug targets and key pathways. Through systematic analysis of more than 500 drugs in cMAP (connectivity map, gene expression profile database), it has been proven that drug interference usually has no significant effects on the expression of drug target genes, hence the changes in expression after drug treatment are insufficient to identify drug targets. However, network topology measurement and local radiance measurement that combine

perturbed gene and functional interaction network information are conducive to discovering cancer-specific pathways [70].

Link prediction in the network refers to predicting the possibility of a connection between two nodes in the network that have not yet been connected through information, such as known network nodes and structures [40]. This prediction includes both the prediction of unknown connections and the possibility of possible new connections.

Chen et al. [41] developed a rebooted random walk model—NRWRH, based on heterogeneous networks, to predict potential drug–target interactions by implementing random walks on heterogeneous networks. This work assumes that similar drugs often interact with similar targets and integrate the drug–drug similarity network, protein–protein similarity network, and known drug–target interaction network, into a heterogeneous network. In this work, NRWRH was used to predict potential drug–target interaction by integrating drug-related information. The originality of this method lies in the integration of three different networks (drug similarity network, target similarity network, and known drug–target interaction network) into a heterogeneous network. NRWRH is applied to four target proteins, including enzymes, ion channels, GPCR, and nuclear receptors, using cross-validation to predict potential drug–target interactions, and demonstrated superior performance of NRWRH over previous methods.

Abhik et al. [42] extended the experimental data set on the basis of NRWRH. This method also integrates the three networks of drug–drug similarity network, protein–protein similarity network, and known drug–target interaction network into a heterogeneous network, and expands relevant drug–target network data and uses external data sets for verification.

This section follows a brief demonstration of the link prediction analysis steps in the Python language.

(1) Description of Question

Let $G(V, E)$ be an undirected graph network, where V is a set of nodes and E is a set of edges. Given the link prediction method, assign a score value "S" to each pair of unconnected node pairs, and then sort all pairs according to the score value from the largest to smallest, with the first node pair having the highest probability of connecting edges [40].

(2) Link Prediction Method

Common link prediction methods are based on similarity, maximum likelihood estimation, and probability model [43, 44]. The similarity-based link prediction methods are divided into three main categories [45]—similarity based on nodes, pathways, and random walks. The concept of the method based on node similarity is: the greater the similarity between two nodes, greater the possibility of links between them. Therefore, there are many definitions of node similarity, including common neighbor index [46], Salton index [47], Jaccard index [48], HDI [49], etc. Based on the similarity index of pathways, there are mainly local path index [50], Katz index [51], and LHN-II index [56]. Similarity indexes based on random walk include

Table 2.2 Code implementation

Core codes	
import networkx as nx	# Import networkx toolkit
data = open("ppi.txt")	# Load PPI data
G = nx.Graph()	# Create empty graph, G network undirected graph
for i, line in enumerate(data):	
line = line.split("\t")	
G.add_edge(line[0], line[1])	# Add data to undirected graph
preds = nx.jaccard_coefficient (G,[(0,1),(2,3)])	# Calculate the Jaccard coefficients of all the unconnected nodes
for u, v, p in preds:	# Triple iterator in the form of (u,v,p), wherein print (% D,% d) -> %.8f' % (u, v, p)
P(u,v) = preds(u,v)	# (u, v) is a pair of nodes and P is their Jaccard coefficient.
>>>	# Program running results
(ATP6V1B1, ATP6V1A) -> 0.75000000	
(17, 1546) -> 0.75000000	

average commute time [56], restarted random walk [57], Cos+ index [58], and SimRank index [52].

(3) Algorithm Implementation Case

In this paper, the Jaccard coefficient in network topology similarity [48] and PPI network data have been used as inputs for predicting links to unconnected nodes in the PPI network.

Jaccard coefficient definition: Given two sets A and B, Jaccard coefficient is the ratio of the size of the intersection of A and B to the size of the union of A and B, which is defined as follows:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

The implementation of Python core code is shown in Table 2.2. The program input is PPI network data (the node number represents the number corresponding to the protein), and the results are shown in Fig. 2.2.

The number of nodes in the above figure represents the number of nodes in the PPI network in this program. We retained the mapping relationship between the numbers and protein molecules. As observed, using Jaccard coefficient, we calculated the relationship index between ATP6V1B1 (node 17) and ATP6V1A (node 1546) as 0.75.

(4) Application of Link Prediction in Network Pharmacology

Link prediction is not limited to social networks, but also has great application value in the biomedical field. With the development of network medicine, researchers have begun to analyze and predict the interaction between proteins,

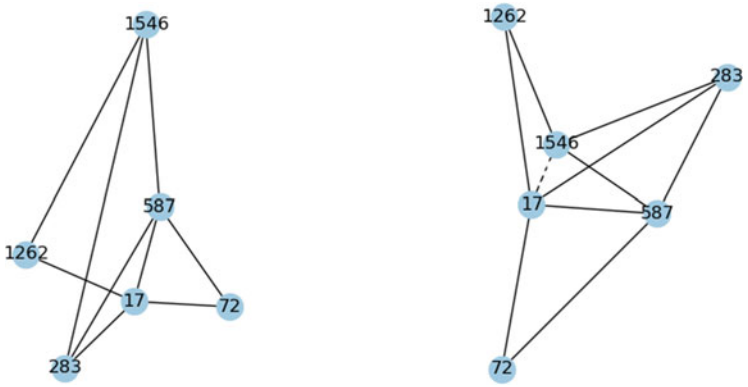

Fig. 2.2 Effect diagrams before and after operation of link prediction program

drugs, and targets at the molecular level [53]. There are links between nodes of protein interaction and the metabolic network [54, 71], which indicates that there is an interaction between them. The prediction of missing drug–target network and suspicious link is helpful to explore the mechanism of action of different drugs and to predict and evaluate drug efficacy. However, revealing the hidden interaction in such networks increases the cost of biological experiments, while the results of link prediction guide these experiments, improving the success rate of experiments, and thus reducing experimental costs. In addition, link prediction can also be used to find similar drugs in the drug network, as well as to find new drug targets, opening up a new path for the research and development of new drugs [72].

4. Analysis Based on Machine Learning

This analysis is different from analytical strategies based on structural comparison or action network simulation. Machine learning-based analytical methods have a more flexible database. It can be a structure, a network, or any other detection index that can be quantified. Many machine learning-based methods have been used to identify relationships between drugs and targets. Machine learning is an analysis method that generates prediction models based on some underlying algorithms and given data sets. It can be divided into unsupervised learning methods (clustering, dimension reduction, association, etc.), supervised learning methods (regression, classification, etc.), and semi-supervised learning methods. In most machine learning-based approaches, biological data sets from multiple sources are integrated, such as chemical structures of drugs, target protein sequences, and known drug–target interactions.

In terms of supervised learning, Yamanishi et al. [31] proposed a nuclear-regression-based method to infer drug target interaction by integrating chemical structure information of compounds, sequence information of target proteins, and topology of known drug target interaction network, to study the interaction of four kinds of drug targets in humans. Bleakley and Yamanishi [30] developed a supervised learning approach based on a two-part local model (BLM) to predict unknown

drug–target interactions, by transforming the edge prediction problem into a binary classification problem. Further, Yamanishi et al. [67] believed that pharmacological action similarity was related more to drug–target interaction than chemical structure similarity, so they further proposed a correlation-based model to infer the unknown drug–target relationship based on chemical structure information, genome sequence information, and large-scale pharmacological action information.

In terms of semi-supervised learning, Xia et al. [73] developed NetLapRLS, a semi-supervised learning method that combines chemical space, genomic space, and known drug–protein interaction network information into a heterogeneous biosphere to predict potential drug–target interactions.

In terms of deep learning, Wang and Zeng [74] proposed a method based on restricted Boltzmann machine (RBM). This framework of multidimensional drug target network not only predicts the binary interaction between drugs and targets, but also predicts the interactions between different types of drugs (i.e., how drugs interact). Ramsundar et al. integrated millions of data points, representing both positive and negative examples of DTI with more than 200 specific goals [34]. They used a multi-tasking framework in which each target prediction is considered a separate task that requires its own (linear) classifier. The AUC (area under the receiver operation curve) of the maximum cross-validation achieved by the deep learning method is 0.87, and it is proven that the multi-tasking aspects of their method always provide slight improvement (AUC increases about 0.01) with the same amount of data compared with the same single task analysis. Wen et al. [59] proposed Deep DTIs, a drug target prediction algorithm framework based on Deep Learning. This method first uses unsupervised pre-training to extract the characterization from the original input descriptor, and then uses the known drug target relationship tags to construct a classification model. Compared to other methods, DeepDTIs perform better and can be further used to predict whether a new drug target is associated with other existing targets or whether a new target interacts with some existing drugs. In addition to improving the prediction performance of deep learning models, the analysis of key chemical characteristics learned by machine learning models for predicting drug activity is also important for understanding the performance of the model, screening models with better generalization ability, and for further protein-compound binding modes. Ding et al. proposed a method to analyze the chemical characteristics learned from the QSAR model based on the neural network hidden layer functions and backtracking gradients. They then developed an interactive tool to identify the molecular characteristics of the GPCR family protein targets binding to compounds, which can be verified by eutectic structural analysis.

In DTI prediction, the general machine learning process is divided into three steps. Firstly, the input data of drugs and targets are preprocessed. The underlying model is then trained based on a set of learning rules. Finally, the test data set is predicted by using the prediction model [45]. Kumari et al. [95] developed a sequence-based prediction method to identify and distinguish human non-drug and drug target proteins. Training features include amino acid sequence characteristics, composition, and dipeptide compositions used to produce prediction models.

Through 10-fold cross-validation and leave-one-out validation tests, the sensitivity, specificity, and accuracy of the model (above 80%), and the Matthews correlation coefficient (above 0.7), can help in evaluating the composition pattern of human drug targets [75]. Zhang et al. [76] proposed a clustering-based multi-view DTI prediction method to achieve more accurate DTI predictions by integrating drug and target data from different views and maximizing clustering consistency in each view, to predict 54 kinds of potential DTI [76]. Jamali et al. [77] used machine learning method to analyze 443 sequence-derived protein features to predict whether proteins had drug properties, and compared the properties of different machine learning methods and conducted feature selection. New drug targets have been identified in cell signaling pathways, gene expression, and signal transduction [77].

In addition, this section provides a description of HTINet [78], a TCM target prediction method based on representation learning. In recent years, with the continuous development of network medicine and pharmacology, multi-source biological network data and databases have been widely accumulated, providing adequate data support for researchers. Meanwhile, representation learning [79] is developing rapidly in the field of deep learning. It is a method that learns the feature representation of each node in the network through the network structure and makes the node feature representation fit the original network structure. This method has been applied in many fields (image, video, and natural language understanding) and achieved good results. The HTINet model integrates TCM and Western medicine data (including Traditional Chinese Medicine, disease, symptoms, Western medicine, and targets) based on symptoms, and integrates a multi-source heterogeneous data network. It also obtains feature representations of Chinese medicine and genes based on the network representation method and finally builds a supervised classification model obtained from previous learning to predict the interaction relationship between Chinese medicine targets. The method flow is shown in Fig. 2.3.

The HTINet model has achieved a maximum of 95% AUC and 94% AUPR on the test set, and its performance has been greatly improved compared with the baseline model, indicating its potential in the prediction of TCM targets. In addition, this work also carried out external validation on some experimental results, randomly selected three Traditional Chinese Medicines (Polygonum bistorta, flos farfarae, and *Rhododendron dauricum*), and predicted its targets through the HTINet model and effectively verified the predicted targets in external databases and literature.

2.2.2 Study on the Drug Property Mechanism

One of the central research objectives of network pharmacology is to completely characterize the biological process under Drug property, i.e. to clarify the mechanism of Drug property. The clarification of intracellular chemical reactions and pathways is the most challenging issue in this field. Common biological pathways are related to metabolism, gene expression regulation, and molecular signaling.

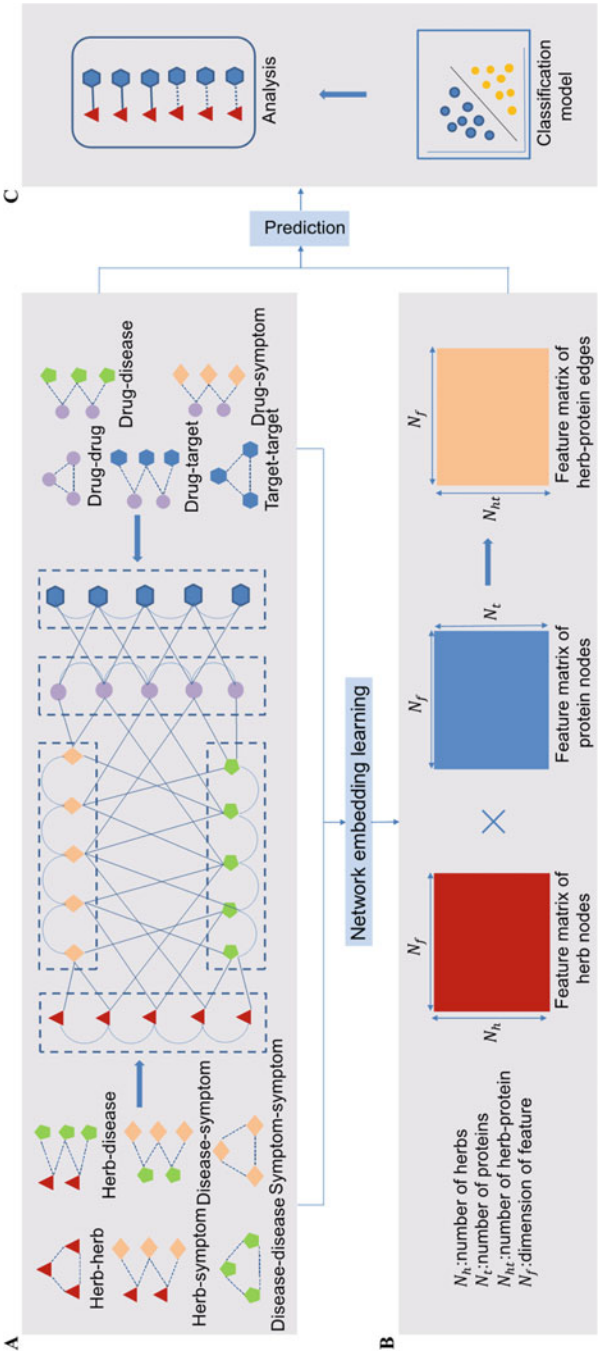

Fig. 2.3 HTINet model construction [78]

Pathways play a key role in the advanced research of functional genomics. For example, identifying disease-related pathways can lead to effective strategies for diagnosis, treatment, and prevention of disease. In addition, researchers can discover the root cause of diseases and use the information obtained from pathway analysis to develop new and better drugs by comparing the differences in some pathways between healthy people and patients. Mapping the dysfunctional pathways associated with various diseases is essential for a comprehensive understanding of these diseases.

One of the common problems in drug research is the accurate prediction of pathways and molecular functions. Pireddu et al. [80] proposed a model for predicting catalytic proteins in important reactions, and integrated these into a prototype system of previously proven metabolic pathways. Finally, 10 metabolic pathways were cross-validated for 13 organisms, and the results showed a 71.5% cross-validation accuracy and 91.5% recall rate [80] in the prediction of catalytic proteins of all reactions. In order to find a quantitative verification method for pathway prediction, Joseph et al. [81] developed a large gold standard data set that contained data on the presence or absence of 5610 metabolic pathways in various organisms. They also defined a set of 123 pathway characteristics and evaluated the information according to the gold standard. This data is used as input in various machine learning (ML) methods to achieve accurate prediction of metabolic pathways [81]. Boudellioua et al. [82] proposed a system that uses "rule mining techniques" to predict the metabolic pathways of prokaryotes. They used cross-validation technology to evaluate the performance of the system and achieved good results in identifying pathways [82]. Fan et al. used the Agilent LitSearch tool [67] to dig deeper into the Pubmed database from 1950 to 2014, for genes that regulate angiogenesis related to ischemia and lung cancer. They then constructed the disease target network for ischemia combined with lung cancer. Verification with molecular biology revealed that the mechanism of bidirectional vascular regulation in animal models of ischemia in lung cancer is related to the abnormal expression of elastase in centrioles [73].

Torcetrapib can inhibit the activity of cholesteryl ester transfer protein and increase high density lipoprotein in vivo. It could have been used as a new anti-lipid drug; however, phase III clinical trials have shown that torcetrapib can induce a fatal hypertensive response [51]. Understanding the molecular mechanisms that induce lethal reactions can help to avoid such situations in the future and clarify whether other CETP inhibitors, such as Anacetrapib and Dalcetrapib should continue to be used. Chang et al. constructed a specific renal metabolic network model through in-depth mining of GEO gene expression data [44]. Combined with the off-target effects of known drugs, CETP inhibitors and renal function were evaluated. At the same time, Fan et al. mapped the gene signaling network of human diseases by integrating the interactions of biomacromolecules in four databases including BioCarta, literature-mined network, Cancer Cell Map, and the HPRD database [45]. Torcetrapib-specific regulation network module was mined by analyzing the GEO database, and the abnormal gene set regulation of torcetrapib was

drawn. The possible explanation of torcetrapib-induced hypertension was thus clarified from a systematic view point.

Drug property model is the key to drug development. It usually involves a target through which a drug can induce pharmacological effects, including understanding the drug influence pathway and biological processes. This information can be used to support treatment hypotheses in animal models, clinical indications, and patient selection. It is also important to distinguish new drugs from current standards, treatments, and competing molecules. Although the mode of action of drugs is not necessary for FDA approval, most researchers hope to understand the function of drugs at the molecular level. There are already some examples of artificial intelligence usage to solve the discovery of Drug property patterns. Pang et al. [83] used "random forests" to analyze gene expression data and established a path based classification and regression method. This approach allows researchers to sequence important pathways from externally available databases, and identify important genes to take advantage of a continuous outcome variable in regression settings [83]. Hancock et al. [84] proposed a new classification model, HME3M. This probabilistic model is a combination of a mixed Markov model, which is used to identify frequently observed path clusters in a specific network structure, and proves that the HME3M algorithm is superior to the comparison method in the case of increasing network complexity and path noise. It is an accurate and reliable classification of metabolic pathways [84].

Carfilzomib is a conventional drug for treating multiple myeloma. However, clinical studies have found that long-term use of Carfilzomib can induce drug resistance in multiple myeloma. Zheng et al. analyzed KMS-11 cell lines that are resistant and sensitive to Carfilzomib in the GEO database, through a string biological macromolecule interaction platform [52]. This helped to model a gene regulatory network related to Carfilzomib resistance. The enrichment analysis results showed that abnormal changes in cytokine and receptor, autophagy, ErbB signaling, microRNA, and fatty acid metabolism pathways may be related to drug resistance exhibited in patients treated with Carfilzomib for multiple myeloma [53].

2.2.3 Discovery of New Drug Uses

Network pharmacology is not only used for drug target discovery and mechanism interpretation, but also for the discovery of new drug uses. Phenotypic and omics data generated in drug experiments and clinical applications provide important clues for the discovery of new drug uses. Artificial intelligence plays an important role in the use of this data.

1. Analysis of Drug-Phenotype Data

Drug phenotypic analysis is a method for analyzing the phenotypic changes in an organism after Drug property. It identifies the effects of a drug by analyzing cell and animal models in a disease state. Although drug discovery based on drug targets

once dominated the scene, several new disease targets determined by genomics and systems biology methods are categorized as non-usable [85]. Moreover, the function of these new targets is unclear. These issues have prompted researchers to refocus on the discovery of drug phenotypes as a complement to target-based drug discovery [86]. The phenotype of the drug includes characteristics of the drug's indications, side effects, etc., which are reflected at the individual level. Drug phenotypes can be attributed to many molecular interactions, including on-target or off-target binding, drug–drug interactions, dose-dependent pharmacokinetics, metabolic activity, downstream pathway interference, aggregation effects, and irreversible target binding. Although certain drug phenotypes such as side effects are unexpected results of drug intervention, they help in understanding the physiological changes caused by drugs. Phenotype-based methods for discovering new uses of drugs are being valued increasingly by researchers.

PubChem's bioassay function contains more than 740 million data points from biochemistry and phenotypic screening, covering more than 1 million biologically active molecules. Several compounds have hundreds or even thousands of analysis results [21, 22]. ChEMBL contains biometric data with more than 12 million data points. NPCPD29 contains a drug-phenotype matrix of nearly 35 clinically approved compounds, covering cardiovascular disease, diabetes, and cancer. In addition, the Center for Chemical Genomics of the National Institutes of Health has compiled a data set of approximately 2500 approved compounds that are screened in approximately 200 phenotypic and target-based tests, focusing on various cancers, malaria, nuclear receptors, and signal pathways [23].

Research on the sensitivity of cancer cell lines is the most important task in network pharmacology based on cell phenotype screening. The Cancer Therapeutic Response Portal assessed the sensitivity of 242 cancer cell lines with genetic characteristics to 354 types of small molecule probes and drugs [16]. The GDSC (Genomics of Drug Sensitivity in Cancer) database measured 138 anticancer drugs in 700 cell lines [18]. The Cancer Cell Line encyclopedia provides detailed genetic characterization of 1000 cancer cell lines and can be used to assess cell line similarity and predict drug perturbation growth rates in other cell lines [24].

SIDER (Side Effects Resources) is a public side effect database that contains compiled information from FDA package specifications, linking 888 drugs with 1450 side effects [27]. The OFFSIDES database analyzed more than 400,000 adverse reactions not listed on the official FDA drug labels, and determined that each drug had an average of 329 off-label ADEs [28]. Finally, the FDA Adverse Event Reporting System (FAERS) is the database of information on adverse event and drug error reports submitted to the FDA by manufacturers, health-care professionals, and the public [29, 30].

Relationships between drugs and phenotypes can be used to identify shared target proteins among chemically different drugs and to infer new indications using their phenotypic similarities [87]. One of the underlying principles behind this theory and related approaches is that drugs that share a large number of similar phenotypes may be associated with common mechanisms of action associated with the treatment of a disease, and may serve as phenotypic biomarkers for specific diseases

[88]. Currently, several new indications and targets have been found by using drug phenotypes using artificial intelligence methods. For example, Dimitri et al. developed DrugClust [89], a machine learning algorithm for drug side effects prediction. According to the Bayesian score, the first batch of drugs was clustered based on their characteristics, and then the side effects were predicted. Biological validation of the clustering can be completed using enrichment analysis. The process of drug discovery is realized by verifying obtained clusters and possible new interactions between some side effects and non-targeted pathways. Luo et al. [90] constructed a drug side effect network based on SIDER2 (Side Effect Resource 2) database, and introduced the link prediction method into the network to develop and evaluate the framework of drug side effect prediction. Ferrero et al. [91] developed the drug re-positioning hypothesis on the basis of disease genetics by mining the public repository and transcriptome profiles of GWAS (Genome-Wide Association Studies) data [91]. Yin et al. [92] used the drug indications in the Medicine Indications Resource (MEDI) as the gold standard to evaluate whether the drug indications found from GWAS and Phewa (Phenome-Wide Association Studies) have clinical indications [92]. Yang et al. [88] extracted the relationship between 3175 diseases and SEs (Side Effects). A naive Bayesian model was then established based on SEs' features to predict the indications of 145 diseases. In addition, the QSAR model of SEs was used to predict the indications of 4200 clinical molecules [88]. Ye et al. [93] constructed a drug–drug network based on the similarity of clinical side effects. The indication of a drug can be inferred by enriching the function of its neighboring FDA-approved drug in the network. It has high accuracy in drug prediction for diabetes, obesity, laxatives, and mycobacteria infection. A large number of predicted results were approved by the FDA or supported by preclinical/clinical studies [93]. Previous studies have shown that chemical structure, target protein, and side effects can provide rich information for drug similarity evaluation. However, each individual data source plays an important role on its own, and data integration is expected to reposition drugs more accurately. Wang et al. [94] established a new drug re-positioning method (predicted drug re-positioning) by integrating the molecular structure, molecular activity, and phenotypic data, and by characterizing drugs by analyzing their chemical structure, target proteins, and side effect data, and defining their disease-related core functions. Then, an SVM was trained to calculate and predict new drug–disease interactions, which has advantages over other methods in terms of accuracy and coverage rate [94].

Scheiber et al. used the known drug–ADE (adverse drug event) association, and the extension of NaïveBayes modeling to connect specific chemical characteristics of drugs with 4210 ADE terms [56]. Liu et al. used the causal relationship analysis based on Bayesian network structure to connect the chemical and biological characteristics of drugs with ADE, which can be interpreted as causality [57]. Vilar et al. used the GBA method in large insurance claims databases to estimate drug associations with four different ADE: acute kidney failure, acute liver failure, acute myocardial infarction, and upper gastrointestinal ulcer [58].

2. Transcriptome Data Analysis

The omics data generated from drug trials undoubtedly provides valuable information for the discovery of new uses of drugs. Compared with other omics data such as proteomics and metabolomics, transcriptome data have many advantages such as high throughput, low cost, precise quantification, and sufficient complexity. Therefore, the large-scale use of transcriptome data for drug discovery is the most rapidly developed and mature method.

(1) Integrated Library Project Based on Network Cellular Response Imprinting (LINCS) [95]

The CMap project [96] and LINCS (The Library of Integrated Network-Based Cellular Signatures) project [97] promoted the development of a comprehensive and large-scale transcriptome database with drug research as an important goal. Drugs and target perturbation data collected and recorded have been used to determine the connections, similarities, or differences between diseases, drugs, genes, and pathways, which provide great opportunities for computational pharmacogenomics and drug design. Unlike classic pharmacology that only focuses on one target at a time, the transcriptomics data provided by CMap and LINCS opens the door for systems biology methods at the pathway and network level [98]. The LINCS project highlights the potential of gene transcription analysis as a universal language for linking chemistry, biology, and clinical practice by inferring genome-wide similarities or differences [99]. In recent years, several studies have used various machine learning methods to analyze Cmap data and LINCS data for target discovery and drug re-positioning. For example, Xie et al. [100] systematically explored and predicted the re-positioning of 480 marketed drugs with other therapeutic attributes using LINCS drug-induced transcript level data, which was based on the machine learning algorithm Softmax for multiple classification problems. Young et al. [101] used the gene silencing perturbation data in LINCS, adopted the linear regression model, and combined the prior and posterior probability to infer the regulatory relationship in genes, thus verifying the relationship identified in the TRANSFAC (TRANScription FACtor database) and JASPAR. Lee et al. [102] used LINCS data to evaluate the ability to predict novel re-positioning of drugs based on several perturbations in four cancer types [102]. Sawada et al. [103] proposed a new computational method for predicting inhibition and activation targets of drug candidate compounds. Integrating chemical induction and gene interference with the gene expression profile of human cell lines helps avoid excessive dependence on the chemical structure of compounds or proteins. Based on the transcriptomic changes of the overall gene expression profile after chemical treatment, as well as the transcriptomic changes after gene knockout and overexpression, the combined learning algorithm was used to build a prediction model of a single target protein. This method can distinguish inhibition targets from activation targets, and can accurately identify therapeutic effects [103]. Liu et al. analyzed the CMap transcription profile and revealed its hidden factors by weighted gene co-expression network analysis (WGCNA). Simultaneously, seven common modules associated with protein binding, extracellular

matrix tissue, and translation were identified. Finally, the drugs were clustered by module expression, and the mechanism of action (MoA) was inferred according to their common activity profiles. Sirota et al. systematically compared the gene expression profiles of 164 small molecule compounds from CMAP with a set of expression profiles derived from the GEO database for 100 different diseases. Based on this model, more than 1000 drug repurposing predictions were generated, linking at least one of 164 compounds to each of the 53 diseases [48].

(2) Gene Expression Omnibus (GEO) [104, 105]

The Gene Expression Omnibus (GEO) is a public information storage platform managed and maintained by the National Center for Biotechnology Information (NCBI) of the United States. The database mainly provides gene expression data retrieval, browsing, query, and download services, and is an important source for obtaining high-throughput chip expression profiles data. GEO includes two sub-databases: Datasets and Profiles database. The Datasets database stores the data of gene chip centered on experiments. The Profiles database stores gene-centric chip data. Currently, GEO has more than 900 drug perturbation experiments and can be another direct source of drug–target perturbations in network pharmacology research.

(3) ArrayExpress Database [106]

The ArrayExpress database is a microarray common repository of gene expression data developed and operated by the European Bioinformatics Institute (EMBI). Its main purpose is to store and record annotated high-throughput data sets and original image sets from all over the world. The ArrayExpress interface is simple and supports multiple retrieval methods. So far, the database includes more than 6000 sets of high-throughput experimental data, including expression data such as RNA-seq, ChIP-seq, GRO-seq, epigenetic profiles, and FAIRE-seq.

3. Docking Profiles Data Analysis

The combination of listed drug targets that are not thoroughly studied with different targets leads to a wide range of side effects. Hence, the cost of screening all potential molecular targets in biological experiments is high. The "molecular docking profiles" using virtual large-scale molecular docking is helpful to study the drug–target relationship, and plays an important role in the development of new clinical indications of drugs. Yang et al. [107] used molecular docking and logistic regression to construct a real-time prediction server DPDR-CPI based on small molecular structures. When a user submits a molecule, the server docks it with 611 human proteins to generate predictive CPI (chemical–protein interactome) characteristic profiles. It shows the correlation between the input molecules and about 1000 human diseases, and gives the highest prediction results [107]. Chen et al. [108] proposed a new ligand-based pipeline: given a set of experimental data, first, use principal component analysis (PCA) and genetic algorithm (GA) to establish a segment descriptor with the signature of the SVM model, and then the pipeline

develops QSARs in the form of the SVM prediction model, and applies the model in virtually screen compound databases [108].

Chavali et al. used the metabolic model to generate lists of 15 genes and 8 dual-gene combinations that were predicted to be relevant targets for neglected tropical diseases (mainly Leishmaniasis) [70]. By associating these genes with 254 FDA-approved compounds based on drug–target interactions, it was found that 14% (10 of 71) of these compounds were validated in overlapping with high content screening data for leishmaniasis. In addition, Chen et al. integrated information such as drug–target interaction, disease–gene association, and protein–protein interaction networks into heterogeneous networks (DrugNet, linking drugs, targets, and diseases) [35]. Using the ProphNet network propagation algorithm, we can define the input query node, drug, or disease, and rank the remaining nodes of other types, that is, the drug for the disease query, and vice versa.

4. Web-Based Drug Indication Analysis

With advancement in the interaction group detection methods and the accumulation of data resources, the discovery of drug indications based on network analysis is widely used in network pharmacology. Relevant studies have shown that drug–target network, drug–drug, drug–disease, protein–protein interaction, transcriptional, and signal transduction networks can be used to identify the efficacy characteristics of drugs, thus providing new opportunities for drug discovery or indication discovery.

Li et al. [109] developed a binary drug–target network approach to identify potential new indications for existing drugs through their relationship with similar drugs. In the bipartite network model, drug pair similarity integrates chemical structure similarity, common drug targets, and protein interactions. The author established a causal network (CauseNet) [110] based on the previous work, which is based on a multi-layered approach to genes, diseases, and drug targets to determine new therapeutic uses of existing drugs. In the causal network, the transition probability of each chain is estimated based on the known drug–disease treatment association.

Wu et al. [111] used the known relationship between disease genes and drug targets in the KEGG database to construct a heterogeneous drug network. Nodes represent drugs or diseases, and edges represent shared genes, biological processes, pathways, phenotypes, or combinations of these characteristics. The network is then clustered to identify modules that can be used to extract potential drug–disease pairs for drug re-positioning. This method not only considers genes, but also other features of constructing disease drug networks.

Chen et al. [68] developed a method based on functional linkage network (FLN) to find modules negatively related to drugs. FLN is a network in which nodes (proteins or genes) are connected by weighted edges to measure the probability of sharing a common biological function. The network is constructed by using different biological information sources (such as mutation and transcription level). These information sources act as the features of a Bayesian classifier, and calculate the possibility of each edge. FLN's filtering method is to remove all genes that are not

within the user-specified genetic distance from the disease mutation and display differential expression below a certain threshold. Such networks are processed to determine the extent to which drugs and disease-related genes are associated with possible re-positioning of candidate genes.

Ali et al. [69] used centrality measurement commonly used in social network analysis to identify drugs with better positioning in the side effect and drug indication networks. The basic assumption of this work was that drugs with similar phenotype profiles (e.g., side effects) can share similar therapeutic properties based on relevant mechanisms of action and vice versa. The development of side effect resources includes unique drugs with side effects and indications. Drugs are ranked according to their centrality scores, thus identifying 18 major drugs from the drug side effect network and 15 major drugs from the drug indication network. Indications and side effects of prominent drugs were inferred from profiles of their network neighbors and compared with existing clinical studies, while seeking optimal similarity threshold values between drugs. Threshold values can then be used to predict indications and side effects for all drugs. The similarities are measured by the extent to which they share a phenotypic profiles and neighbors.

Campillos proposed in 2008 that drug–target interaction networks using the principle of side-effect similarity might be overlooked in new drug discovery. By analyzing the side effects of 746 drugs already in the market, his team constructed a drug-side-target network with 1018 nodes, and found some new activities and new indications of some drugs through biological verification [87].

5. Analysis of Drug Indication Based on Machine Learning

The prediction of drug indication is also a typical machine learning problem [70]. Specifically, the interaction between drugs and the human body can be gauged and predicted through a series of clinical and biological characteristics. In this section, we summarize the general principles and types of drug indication analysis algorithms based on machine learning.

An important advantage of machine learning algorithm is its richness and rapid development. Any existing or new algorithm can be applied to drug indication analysis with some modification. In this section, the drug expression profile data combined with the machine learning algorithm is taken as an example to predict its indications, i.e. drug expression profile is used as a predictor (i.e., feature) for the therapeutic potential of drugs. The resulting variable can be a drug, for example, cardiovascular or anticancer drug or a drug targeted at a specific disease like diabetes. In the former case, consideration may be given to the classification of a drug in a category other than its own indications, for re-positioning. In the latter case, a drug with a high predictive probability but not shown as a disease, may be a candidate for re-positioning. Existing indications for drugs are readily available from public web resources such as the Anatomical Therapeutic Chemistry (ATC) classification system. The following is a detailed introduction of different types of drug indication prediction methods:

In terms of the prediction of drug indications, a linear model has advantages of rapid calculation speed, intuitiveness, and can be easily realized by a variety of

programming languages and statistical software. For example, the glmnet package in R language supports rapid implementation of normalized linear models and has detailed documentation available online [112]. Linear models are also easy to explain, as the importance of features can be gauged from the size of the regression coefficient, and methods have recently been developed to assess statistical significance [113]. However, linear models capture only linear relationships between input characteristics and output variables, which may not be the case in many real-world scenarios, including biomedical applications. A recent study [114] identified transcriptional response as a multi-label classification problem, identified novel therapeutic properties of drugs, and pointed out that multi-label logistic regression is superior to other methods such as random forest and convolutional neural networks.

In terms of drug indication prediction methods based on classification and regression models, Napolitano et al. [115] integrated a variety of drug characteristics, including chemical structure and proximity of targets in the interaction network and expression profiles, and used support vector machine (SVM) to predict the treatment category. Menden et al. [116] developed a machine learning model to predict the response of cancer cell lines to drug treatment, which was quantified by a semi-inhibitory concentration (IC50) value. In this model, the feed-forward perceptron neural network model and random forest regression model were established using the oncogenome characteristics and chemical properties (such as structural fingerprints) of the cell line. The predicted IC50 value was further cross-validated and independent blind tests were done. Gottlieb et al. [117] integrated various disease-related characteristics (such as phenotype and genetic characteristics), calculated the similarity of drugs and diseases, constructed classification features and further used logistic regression classifiers to predict new drug indications.

In terms of predicting drug indications based on collaborative filtering technology, Zhang et al. [118] proposed a unified calculation framework for integrating the multidimensional features of drug similarity and disease similarity. Simply put, drug similarity matrix and disease similarity matrix are extracted by integrating genome (e.g., drug target protein, disease gene), phenotype (e.g., disease phenotype, drug side effect), and chemical structure (e.g., drug chemical structure). Based on this information, this author turns the drug–disease network analysis into an optimization problem. This computational framework shows the effectiveness of exploring new indications for drugs. Yang et al. [119] used causal inference probability matrix factorization to infer drug–disease correlation. In this model, they integrate multi-level relationships, construct causal networks linking drug–target–pathway–gene–disease and learn PMF patterns based on known interactions. This approach can predict new drug–disease associations and thus be of value for drug indication analysis.

2.2.4 Traditional Chinese Medicine and Its Therapeutic Theory

The composition of TCM prescriptions is complex, and research on its ingredients and treatment is more complex than that of chemical drugs. Network pharmacology plays an important role in revealing the material basis of TCM and the theoretical research of treatment with TCM. Researchers are increasingly using artificial intelligence to solve important problems such as prediction of Chinese medicine target in theoretical research of TCM, molecular mechanism of Chinese medicine prescriptions, and molecular mechanism of syndrome theory.

1. TCM Target Prediction

The determination of drug targets is the key to drug R & D. TCM usually needs to have a synergistic effect between different ingredients due to its complex compound composition, resulting in the complex TCM mechanism of action. In terms of actual target prediction, Zhang et al. [120] proposed a systematic pharmacology method to predict the complexity of compound components and related multiple targets. This was done by identifying bioactive compounds of TCM, to clarify its molecular mechanism of action. System pharmacology method also helps to understand the complex interactions between biological systems, drugs, and diseases from a network perspective. Modern technologies such as drug screening (high-throughput screening, high content screening, and virtual screening) and omics methods (proteomics, genomics, metabolomics) have also been widely used in the identification of bioactive ingredients and drug targets in TCM. Wang et al. [121] introduced high content screening technology and used the HCS instrument to screen TCM-derived compounds and promoted technology development. In order to promote research on the function and mechanism of TCM, ETCM [122] provides the predicted target genes of Chinese medicine ingredients, TCM, and prescriptions according to the similarity of chemical fingerprints between TCM ingredients and known drugs. In the ETCM system, researchers also explored the relationship between TCM, formula, ingredients, gene target, and related pathways or diseases, to finally establish a network structure.

With the development of artificial intelligence, especially the progress made in natural language processing, drug target prediction and discovery have been combined to greatly improve research efficiency. Biomedical literature information can be obtained from the network. Sometimes the abstracts of these literatures contain important frontier research information of drugs and targets. If we can capture the latest research trends of drug targets on time, it will help to advance the process of target prediction. Extracting valuable information from massive amount of literature is the main aim of natural language processing. Real-time literature is collected through web crawler technology, and then large-scale distributed storage is carried out, which can be cleaned by data extraction, exchange, and loading, to preprocess structured data. Then, by using methods such as part-of-speech analysis, grammatical analysis, and semantic analysis in natural language processing technology,

combined with the similarity analysis, cluster analysis, topic mining, and relationship extraction in machine learning, the relationship between drugs and targets is established. Combined with the database of known drug targets, the knowledge mining system is conversely applied to improve the accuracy of drug target knowledge, thus further improving the efficiency of drug target prediction and reducing costs.

2. Study on the Molecular Mechanism of TCM Prescriptions

TCM and its formulations contain many active molecules with complex ingredients, resulting in complex interactions and mechanisms of action. Only by further understanding the mechanism of action and clinical efficacy can we help users. The basic form of TCM for disease prevention and treatment is TCM compound prescription, which is a quantitative mixture of several specific Chinese herbal medicinal plants. There are a lot of chemical substances in TCM compound prescriptions, which may interact with multiple disease-related targets. Therefore, at the molecular level, the TCM compound mechanism used for disease treatment is like that of multi-directional pharmacology or network pharmacology. TCM has existed since ancient times in China. Molecular biology originated in modern times, and its effective combination with TCM is a topic that needs to be explored. If we can prove the rationality of TCM prescriptions and formulas at the molecular level, it will help to integrate modern science and technology with ancient Chinese medicine prescriptions, which will not only provide a more reasonable scientific basis for further optimization of TCM prescriptions, but also provide a solid backing for TCM's growth in the international market. At present, many pharmacological studies have been used to reveal the mechanism of action of TCM and its molecular mechanism. For example, research in the field of aging shows that hemopoietic stem cell autophagy has anti-aging effects, and there are many new discoveries in the field of plant extracts and Chinese herbal medicine [123]. Among them, Chinese herbal medicine extracts represented by curcumin and resveratrol, some single Chinese medicine extracts, and classical Chinese medicine prescriptions have partial anti-aging effects by regulating the molecular mechanism of aging in vivo and in vitro. Research on the molecular mechanism of TCM prescriptions can be carried out with the help of the TCM information database TCM-ID [124], which provides comprehensive information on TCM, including prescription ingredients, molecular structure, and functional characteristics of TCM ingredients and active ingredients, TCM formula, clinical indications, and application of each Chinese herbal medicine. Zhu et al. designed the framework of the TCM prescription analysis system based on existing TCM prescription data resources and TCM prescription analysis systems, using artificial intelligence and data mining technology. This system assists in various applications, such as knowledge extraction and knowledgebase construction, establishment and improvement of prescription database, medication experience sorting and mining, and new drug development [125].

3. Study on Biomolecular Network Mechanism of TCM Syndromes

The biological basis of syndromes is the key to modernizing TCM. Currently, research is being conducted on blood stasis, cold syndrome, heat syndrome, etc. Research on syndromes includes several aspects such as the nature and essence of the syndrome and micro-syndrome differentiation [126]. The syndrome usually refers to the overall physiological and pathological state of the human body and diagnosis based on it. The TCM theory of disease treatment has gradually developed on the basis of syndrome differentiation. Syndrome theory has accompanied the development of TCM and has been guiding clinical work as well. However, syndromes and their classification have not been effectively developed in recent years. The main reason lies in the lack of appropriate supportive scientific data, and what information exists, is often obtained through subjective inquiry from TCM doctors. In recent years, the basic research of syndrome biology has shifted from inquiry to theoretical research and has made a lot of progress. Some studies have tried to correlate the phenotype of the syndrome with the microbiological molecules, and then studied the syndrome. They have further combined it with modern scientific means to prove some of the already existing syndrome theories. Domestically, some scholars have studied the theory of syndrome biology from the perspective of biomolecular network [126], and have established a multi-layer architecture from phenotypic network, biomolecular network to drug network. Based on this network framework, some typical syndromes such as cold and heat syndromes were studied, which laid a good foundation for the scientific theoretical research of syndromes. At the same time, the characteristics of diseases and syndromes on the biological molecular network were studied, thus providing additional means of finding methods and drugs for systematic intervention of these disease syndromes. Therefore, the old topic of TCM syndrome differentiation and treatment has been extended to the modern field of molecules.

2.3 Application of Artificial Intelligence

This chapter briefly introduces the application of artificial intelligence technology in network pharmacology. With the rapid accumulation of effective data in the life science and pharmaceutical research fields, it has led to unique perspectives on the application of machine learning in new drug development or drug re-positioning. Information on the structure of small drug molecules is available on the PubChem [127] and drug bank [128] databases. These databases contain information of listed drugs, and QSAR is often used to study drugs with annotated information, to find potential new drugs [44]. The PDB (Protein Data Bank archive) [129] database reveals drug–target interaction relationship, based on ligand–target structure related data, information on side effects from Sider [130], and vector data for drug–target interaction relationships. These can be used to predict potential new targets for drugs [45]. In terms of omics data, there is a GEO (Gene Expression Omnibus) database

that stores high-throughput chip data [131], TCGA (The Cancer Genome Atlas) [132], etc. There are databases on expression profile based on cell response to drugs under different conditions, used to predict drug interactions/indications [100, 133] and side effects [134]. In summary, these expansive, high-dimensional databases provide relevant information on artificial intelligence, which plays an important role in drug research. The use of artificial intelligence to guide drug screening and discovery in future drug development may become the norm and bring revolutionary changes to the pharmaceutical industry.

References

1. Hopkins AL. Network pharmacology: the next paradigm in drug discovery. Nat Chem Biol. 2008;4(11):682–90.
2. Zhang Y. Progress in network pharmacology and modern research of traditional Chinese medicine. Chin J Pharmacol Toxicol. 2015;29(06):883–92 (in Chinese).
3. Maulik U, Bandyopadhyay S. Genetic algorithm-based clustering technique. 2000;33.
4. Selim S, Alsultan K. A simulated annealing algorithm for the clustering problem. 1991;24.
5. Zou L. Artificial intelligence and its development and application. Inf Network Secur. 2012 (02):11–3 (in Chinese).
6. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM. 2017;60(6):84–90.
7. Lecun Y, Boser B, Denker J, et al. Handwritten digit recognition with a back-propagation network. 1997;2.
8. Mikolov T, Deoras A, Povey D, et al. Strategies for training large scale neural network language models. 2011.
9. Mohamed A, Dahl GE, Hinton G. Acoustic modeling using deep belief networks. IEEE Trans Audio Speech Lang Process. 2012;20(1):14–22.
10. Kalinin AA, Higgins GA, Reamaroon N, et al. Deep learning in pharmacogenomics: from gene regulation to patient stratification. Pharmacogenomics. 2018;19(7):629–50.
11. Gawehn E, Hiss JA, Schneider G. Deep learning in drug discovery. Mol Inf. 2016;35(1):3–14.
12. Kotsiantis S. Supervised machine learning: a review of classification techniques. 2007;31.
13. Lecun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–44.
14. Jain AK, Murty MN, Flynn PJ. Data clustering: a review. ACM Comput Surv. 1999;31 (3):264–323.
15. Van Der Maaten L, Postma E, Herik H. Dimensionality reduction: a comparative review. 2007;10.
16. Zhou W. Network construction technology in network pharmacology research. Int J Pharm Res. 2016;43(05):797–812 (in Chinese).
17. Akhmedov M, Kedaigle A, Chong RE, et al. PCSF: an R-package for network-based interpretation of high-throughput data. PLoS Comput Biol. 2017;13(7):e1005694.
18. Turing AM. Computing machinery and intelligence// computers and thought. American Association for Artificial Intelligence; 1950.
19. He H, Garcia EA. Learning from imbalanced data. 2009;21.
20. Altschul S, Gish W, Miller W, et al. Basic local alignment search tool. 1990;215.
21. O'Boyle NM, Banck M, James CA, et al. Open Babel: an open chemical toolbox. J Cheminformatics. 2011;3(1):33.
22. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem. 2009:455–61.

23. Case DA, Cheatham TE, Darden T, et al. The Amber biomolecular simulation programs. J Comput Chem. 2005;26(16):1668–88.
24. Iorio F, Tagliaferri R, Bernardo DD. Identifying network of drug mode of action by gene expression profiling. J Comput Biol. 2009;16(2):241–51.
25. Subramanian A, Narayan R, Corsello SM, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. Cell. 2017;171(6):1437–52.e17.
26. Van Der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9:2579–605.
27. Shekhar K, Lapan SW, Whitney IE, et al. Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. Cell. 2016;166(5):1308–23.e30.
28. Grimes M, Hall B, Foltz L, et al. Integration of protein phosphorylation, acetylation, and methylation data sets to outline lung cancer signaling networks. Sci Signal. 2018;11(531): eaaq1087.
29. Vidyasagar M. Identifying predictive features in drug response using machine learning: opportunities and challenges. Annu Rev Pharmacol Toxicol. 2015;55(1):15–34.
30. Gamazon E, Wheeler H, Shah K. A gene-based association method for mapping traits using reference transcriptome data. 2015;47.
31. Xiong J, Zhou T. Gene regulatory network inference from multifactorial perturbation data. 2012.
32. Yamanishi Y, Kotera M, Moriya Y, et al. DINIES: drug-target interaction network inference engine based on supervised analysis. Nucleic Acids Res. 2014;42(W1):W39–45.
33. Gopalan PK, Blei DM. Efficient discovery of overlapping communities in massive networks. Proc Natl Acad Sci. 2013;110(36):14534–9.
34. Chen H, Engkvist O, Wang Y, et al. The rise of deep learning in drug discovery. Drug Discov Today. 2018;23(6):1241–50.
35. Mayr A, Klambauer G, Unterthiner T, et al. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. Chem Sci. 2018;9(24):5441–51.
36. Korotcov A, Tkachenko V, Russo DP, et al. Comparison of deep learning with multiple machine learning methods and metrics using diverse drug discovery data sets. Mol Pharm. 2017;14(12):4462–75.
37. Teschendorff AE. Avoiding common pitfalls in machine learning omic data science. Nat Mater. 2019;18(5):422–7.
38. Bero SA, Muda AK, Choo YH, et al. Similarity measure for molecular structure: a brief review. J Phys: Conf Ser. 2017;892:012015.
39. Zhang Y, Cheng X, Zhou W. Drug reorientation: an important application field of network pharmacology. Chin J Pharmacol Toxicol. 2012;26(6):779–85 (in Chinese).
40. Willett P. The calculation of molecular structural similarity: principles and practice. Mol Inf. 2014;33(6–7):403–13.
41. Vilar S, Harpaz R, Uriarte E, et al. Drug—drug interaction through molecular structure similarity analysis. J Am Med Inform Assoc. 2012;19(6):1066–74.
42. Yan C, Wang J, Lan W, et al. SDTRLS: predicting drug-target interactions for complex diseases based on chemical substructures. Complexity. 2017;2017:1–10.
43. Keiser MJ, Setola V, Irwin JJ, et al. Predicting new molecular targets for known drugs. Nature. 2009;462(7270):175–81.
44. Neves BJ, Braga RC, Melo Filho CC, et al. QSAR-based virtual screening: advances and applications in drug discovery. Front Pharmacol. 2018;9:1275.
45. Chen R, Liu X, Jin S, et al. Machine learning for drug-target interaction prediction. Molecules. 2018;23(9):2208.
46. Mitchell JBO. ChemInform abstract: the relationship between the sequence identities of helical proteins in the PDB and the molecular similarities of their ligands. ChemInform. 2002;33(10): no-no.
47. Bleakley K, Yamanishi Y. Supervised prediction of drug–target interactions using bipartite local models. Bioinformatics. 2009;25(18):2397–403.

48. Yamanishi Y, Araki M, Gutteridge A, et al. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. Bioinformatics. 2008;24(13):i232–40.
49. Michael J, et al. Predicting new molecular targets for known drugs. Nature. 2009;462 (7270):175–81.
50. Martin YC, Kofron JL, Traphagen LM. Do structurally similar molecules have similar biological activity? J Med Chem. 2002;45(19):4350–8.
51. Schuffenhauer A, Floersheim P, Acklin P, et al. Similarity metrics for ligands reflecting the similarity of the target proteins. J Chem Inf Comput Sci. 2003;43(2):391–405.
52. Cheng F, et al. Prediction of drug-target interactions and drug repositioning via network-based inference. PLoS Comput Biol. 2012;8(5):e1002503.
53. Zhao SW, Li S. Network-based relating pharmacological and genomic spaces for drug target identification. PLoS ONE. 2010;5:e11764.
54. Zhang L, Fourches D, Sedykh A, et al. Discovery of novel antimalarial compounds enabled by QSAR-based virtual screening. J Chem Inf Model. 2013;53(2):475–92.
55. Fan S, Li X. Reverse molecular docking: a new approach to discovery and identification of drug targets. Adv Physiol Sci. 2012;043(005):367–70 (in Chinese).
56. Kuhn M, Campillos M, González P, et al. Large-scale prediction of drug-target relationships. FEBS Lett. 2008;582(8):1283–90.
57. Hansen NT, Brunak S, Altman RB. Generating genome-scale candidate gene lists for pharmacogenomics. Clin Pharmacol Ther. 2009;86(2):183–9.
58. Kutalik Z, Beckmann JS, Bergmann S, et al. A modular approach for integrative analysis of large-scale gene-expression and drug-response data. Nat Biotechnol. 2008;26(5):531–9.
59. Chen YZ, Zhi DG. Ligand–protein inverse docking and its potential use in the computer search of protein targets of a small molecule. Proteins. 2001;43(2):217–26.
60. Li H, Gao Z, Kang L, et al. TarFisDock: a web server for identifying drug targets with docking approach. Nucleic Acids Res. 2006;34(Web Server):W219–W224.
61. Liu X, Ouyang S, Yu B, et al. PharmMapper server: a web server for potential drug target identification using pharmacophore mapping approach. Nucleic Acids Res. 2010;38(Web Server):W609–W614.
62. Kinnings SL, Jackson RM. ReverseScreen3D: a structure-based ligand matching method to identify protein targets. J Chem Inf Model. 2011;51(3):624–34.
63. Wang JC, Chu PY, Chen CM, et al. idTarget: a web server for identifying protein targets of small chemical molecules with robust scoring functions and a divide-and-conquer docking approach. Nucleic Acids Res. 2012;40(W1):W393–9.
64. Yue QX, Cao ZW, Guan SH, et al. Proteomics characterization of the cytotoxicity mechanism of ganoderic acid D and computer-automated estimation of the possible drug target network. Mol Cell Proteomics. 2008;7(5):949–61.
65. Feng LX, Jing CJ, Tang KL, et al. Clarifying the signal network of salvianolic acid B using proteomic assay and bioinformatic analysis†. PROTEOMICS. 2011;11(8):1473–85.
66. Gormley GJ, Stoner E, Bruskewitz RC, et al. The effect of finasteride in men with benign prostatic hyperplasia. J Urol. 2002;167(2, Part 2):1102–7.
67. Wu Z, Li W, Liu G, et al. Network-based methods for prediction of drug-target interactions. Front Pharmacol. 2018;9:1134.
68. Feixiong C, Chuang L, Jing J, et al. Prediction of drug-target interactions and drug repositioning via network-based inference. PLoS Comput Biol. 2012;8(5):e1002503.
69. Wu H, Miller E, Wijegunawardana D, et al. MD-Miner: a network-based approach for personalized drug repositioning. BMC Syst Biol. 2017;11(S5):86.
70. Isik Z, Baldow C, Cannistraci CV, et al. Drug target prioritization by perturbed gene expression and network information. Sci Rep. 2015;5(1):17417.
71. Melo Filho CC, Dantas RF, Braga RC, et al. QSAR-driven discovery of novel chemical scaffolds active against Schistosoma mansoni. J Chem Inf Model. 2016;56(7):1357–72.
72. Gomes MN, Braga RC, Grzelak EM, et al. QSAR-driven design, synthesis and discovery of potent chalcone derivatives with antitubercular activity. Eur J Med Chem. 2017;137:126–38.

73. Shen J, Tan C, Zhang Y, et al. Discovery of potent ligands for estrogen receptor β by structure-based virtual screening. J Med Chem. 2010;53(14):5361–5.
74. Hu G, Li X, Zhang X, et al. Discovery of inhibitors to block interactions of HIV-1 integrase with human LEDGF/p75 via structure-based virtual screening and bioassays. J Med Chem. 2012;55(22):10108–17.
75. Kumari P, Nath A, Chaube R. Identification of human drug targets using machine-learning algorithms. Comput Biol Med. 2015;56:175–81.
76. Zhang X, Li L, Ng MK, et al. Drug-target interaction prediction by integrating multiview network data. Comput Biol Chem. 2017;69:185–93.
77. Jamali AA, Ferdousi R, Razzaghi S, et al. DrugMiner: comparative analysis of machine learning algorithms for prediction of potential druggable proteins. Drug Discov Today. 2016;21(5):718–24.
78. Tang Y, Zhu W, Chen K, et al. New technologies in computer-aided drug design: toward target identification and new chemical entity discovery. Drug Discov Today: Technologies. 2006;3 (3):307–13.
79. Rognan D. Structure-based approaches to target fishing and ligand profiling. Mol Inf. 2010;29 (3):176–87.
80. Pireddu L, Poulin B, Szafron D, et al. Pathway analyst automated metabolic pathway prediction// IEEE Symposium on Computational Intelligence in Bioinformatics & Computational Biology. IEEE, 2005.
81. Dale JM, Popescu L, Karp PD. Machine learning methods for metabolic pathway prediction. BMC Bioinformatics. 2010;11(1):15.
82. Boudellioua I, Saidi R, Hoehndorf R, et al. Prediction of metabolic pathway involvement in prokaryotic UniProtKB data by association rule mining. PLoS ONE. 2016;11(7):e0158896.
83. Pang H, Lin A, Holford M, et al. Pathway analysis using random forests classification and regression. Bioinformatics. 2006;22(16):2028–36.
84. Hancock T, Mamitsuka H. A Markov classification model for metabolic pathways. Algorithms Mol Biol. 2010;5(1):10.
85. Pammolli F, Magazzini L, Riccaboni M. The productivity crisis in pharmaceutical R&D. Nat Rev Drug Discov. 2011;10(6):428–38.
86. Swinney DC, Anthony J. How were new medicines discovered? Nat Rev Drug Discov. 2011;10(7):507–19.
87. Campillos M, Kuhn M, Gavin AC, et al. Drug target identification using side-effect similarity. Science. 2008;321(5886):263–6.
88. Yang L, Agarwal P. Systematic drug repositioning based on clinical side-effects. P. Csermely. PLoS ONE. 2011;6(12):e28025.
89. Dimitri GM, Lió P. DrugClust: a machine learning approach for drugs side effects prediction. Comput Biol Chem. 2017;68:204–10.
90. Luo Y, Liu Q, Wu W, et al. Predicting drug side effects based on link prediction in bipartite network. Proceedings - 2014 7th International Conference on BioMedical Engineering and Informatics, BMEI 2014; 2015. p. 729–33.
91. Ferrero E, Agarwal P. Connecting genetics and gene expression data for target prioritisation and drug repositioning. BioData Mining. 2018;11(1):7.
92. Yin W, Gao C, Xu Y, et al. Learning opportunities for drug repositioning via GWAS and PheWAS findings. 2018.
93. Ye H, Liu Q, Wei J. Construction of drug network based on side effects and its application for drug repositioning. PLoS ONE. 2014;9(2):e87864.
94. Wang Y, Chen S, Deng N, et al. Drug repositioning by Kernel-based integration of molecular structure, molecular activity, and phenotype data. PLoS ONE. 2013;8(11):e78518.
95. Duan Q, Flynn C, Niepel M, et al. LINCS Canvas Browser: interactive web app to query, browse and interrogate LINCS L1000 gene expression signatures. Nucleic Acids Res. 2014;42 (W1):W449–60.

96. Lamb J. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. Science. 2006;313(5795):1929–35.
97. Keenan AB, Jenkins SL, Jagodnik KM, et al. The library of integrated network-based cellular signatures NIH program: system-level cataloging of human cells response to perturbations. Cell Syst. 2018;6(1):13–24.
98. Musa A, Ghoraie LS, Zhang SD, et al. A review of connectivity map and computational approaches in pharmacogenomics. Brief Bioinformatics. 2017:bbw112.
99. Iorio F, Rittman T, Ge H, et al. Transcriptional data: a new gateway to drug repositioning? Drug Discov Today. 2013;18(7–8):350–7.
100. Xie L, He S, Wen Y, et al. Discovery of novel therapeutic properties of drugs from transcriptional responses based on multi-label classification. Sci Rep. 2017;7(1)
101. Young WC, Raftery AE, Yeung KY. A posterior probability approach for gene regulatory network inference in genetic perturbation data. Math Biosci Eng. 2016;13:1241–51.
102. Lee H, Kang S, Kim W. Drug repositioning for cancer therapy based on large-scale drug-induced transcriptional signatures. E. PLoS ONE. 2016;11(3):e0150460.
103. Sawada R, Iwata M, Tabei Y, et al. Predicting inhibitory and activatory drug targets by chemically and genetically perturbed transcriptome signatures. Sci Rep. 2018;8(1):156.
104. Edgar R, lash A. 6. The Gene Expression Omnibus (GEO): a gene expression and hybridization repository. 2002.
105. Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res. 2013;41(Database issue):D991.
106. Parkinson H, Kapushesky M, Shojatalab M, et al. ArrayExpress--a public database of microarray experiments and gene expression profiles. Nucleic Acids Res. 2007;35(Database issue): D747–50.
107. Luo H, Zhang P, Cao XH, et al. DPDR-CPI, a server that predicts drug positioning and drug repositioning via chemical-protein interactome. Sci Rep. 2016;6(1):35996.
108. Chen JJF, Visco DP. Developing an in silico pipeline for faster drug candidate discovery: virtual high throughput screening with the signature molecular descriptor using support vector machine models. Chem Eng Sci. 2016:S0009250916300914.
109. Li H, Gao Z, Kang L, et al. TarFisDock: a web server for identifying drug targets with docking approach. Nucleic Acids Res. 2006;34(Web Server):W219–W224.
110. Luo H, Chen J, Shi L, et al. DRAR-CPI: a server for identifying drug repositioning potential and adverse drug reactions via the chemical-protein interactome. Nucleic Acids Res. 2011;39 (suppl_2):W492–W498.
111. Ruiz Carmona S, Alvarez-Garcia D, Foloppe N, et al. rDock: a fast, versatile and open source program for docking ligands to proteins and nucleic acids. PLoS Comput Biol. 2014;10: e1003571.
112. Lu L. Link prediction of complex networks. J UESTC. 2010;39(5):651–61 (in Chinese).
113. Chen X, Liu MX, Yan GY. Drug-target interaction prediction by random walk on the heterogeneous network. Mol Biosyst. 2012;8(7):1970.
114. Seal A, Ahn YY, Wild DJ. Optimizing drug-target interaction prediction based on random walk on heterogeneous networks. J Cheminform. 2015;7:40.
115. Zhang Y, Feng Y. Methods and development of link prediction. Measure Control Technol. 2019; 38(2):8–12 (in Chinese).
116. Liu W, Lü LY. Link prediction based on local random walk. EPL (Europhys Lett). 2010;89 (5):58007.
117. Chen B. Link prediction of complex networks and its application in recommendation. Nanjing University of Aeronautics and Astronautics, 2016;126 (in Chinese).
118. Lorrain F, White HC. Structural equivalence of individuals in social networks. Soc Networks. 1977;1(1):67–98.
119. Chowdhury G. Introduction to modern information retrieval. McGraw Hill; 1983.
120. Zhang W, Huai Y, Miao Z, et al. Systems pharmacology for investigation of the mechanisms of action of traditional chinese medicine in drug discovery. Front Pharmacol. 2019;10

121. Wang J, Wu MY, Tan JQ, et al. High content screening for drug discovery from traditional Chinese medicine. Chin Med. 2019;14(1):5.
122. Xu HY, Zhang YQ, Liu ZM, et al. ETCM: an encyclopaedia of traditional Chinese medicine. Nucleic Acids Res. 2019;47(Database issue):D976.
123. Liu BH, Gu YH, Tu Y, et al. Molecular regulative mechanisms of aging and interventional effects of Chinese herbal medicine. Zhongguo Zhong Yao Za Zhi. 2017;42(16):3065–71.
124. Chen X, Zhou H, Liu YB, et al. Database of traditional Chinese medicine and its application to studies of mechanism and to prescription validation. 2006.
125. Zhu Y, Gao B, Cui M. Design and implementation of TCM prescription analysis system framework. Chin J Tradit Chin Med. 2014;29(5):1543–46 (in Chinese).
126. Li S. conception and Research on biomarkers of TCM syndromes. J Tradit Chin Med. 2009 (9):7–10 (in Chinese).
127. Li Q, Cheng T, Wang Y, et al. PubChem as a public resource for drug discovery. Drug Discov Today. 2010;15(23–24):1052–7.
128. Wishart DS, Feunang YD, Guo AC, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2018;46(D1):D1074–82.
129. Burley SK, Berman HM, Bhikadiya C, et al. RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. Nucleic Acids Res. 2019;47(D1):D464–74.
130. Kuhn M, Letunic I, Jensen LJ, et al. The SIDER database of drugs and side effects. Nucleic Acids Res. 2016;44(D1):D1075–9.
131. Edgar R, lash A. 6. The Gene Expression Omnibus (GEO): a gene expression and hybridization repository. 2002.
132. Tomczak K, Czerwińska P, Wiznerowicz M. Review the Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Współczesna Onkologia. 2015;1A:68–77.
133. Aliper A, Plis S, Artemov A, et al. Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. Mol Pharm. 2016;13 (7):2524–30.
134. Wang Z, Clark NR, Ma'ayan A. Drug-induced adverse events prediction with the LINCS L1000 data. Bioinformatics. 2016;32(15):2338–45.