Genomics xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Genomics

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

Original Article

Predicting drug-target interactions using Lasso with random forest based on evolutionary information and chemical structure

Han Shi[a,b,c,1], Simin Liu[a,b,c,1], Junqi Chen[a,b,c,1], Xuan Li[c], Qin Ma[d], Bin Yu[a,b,e,*]

[a] College of Mathematics and Physics, Qingdao University of Science and Technology, Qingdao 266061, China
[b] Artificial Intelligence and Biomedical Big Data Research Center, Qingdao University of Science and Technology, Qingdao 266061, China
[c] Key Laboratory of Synthetic Biology, CAS Center for Excellence in Molecular Plant Sciences, Institute of Plant Physiology and Ecology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200032, China
[d] Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH 43210, USA
[e] School of Life Sciences, University of Science and Technology of China, Hefei 230027, China

ARTICLE INFO

Keywords:
Drug-target interactions
Pseudo-position specific scoring matrix
Molecular fingerprint
Lasso
SMOTE
Random forest

ABSTRACT

The identification of drug-target interactions has great significance for pharmaceutical scientific research. Since traditional experimental methods identifying drug-target interactions is costly and time-consuming, the use of machine learning methods to predict potential drug-target interactions has attracted widespread attention. This paper presents a novel drug-target interactions prediction method called LRF-DTIs. Firstly, the pseudo-position specific scoring matrix (PsePSSM) and FP2 molecular fingerprinting were used to extract the features of drug-target. Secondly, using Lasso to reduce the dimension of the extracted feature information and then the Synthetic Minority Oversampling Technique (SMOTE) method was used to deal with unbalanced data. Finally, the processed feature vectors were input into a random forest (RF) classifier to predict drug-target interactions. Through 10 trials of 5-fold cross-validation, the overall prediction accuracies on the enzyme, ion channel (IC), G-protein-coupled receptor (GPCR) and nuclear receptor (NR) datasets reached 98.09%, 97.32%, 95.69%, and 94.88%, respectively, and compared with other prediction methods. In addition, we have tested and verified that our method not only could be applied to predict the new interactions but also could obtain a satisfactory result on the new dataset. All the experimental results indicate that our method can significantly improve the prediction accuracy of drug-target interactions and play a vital role in the new drug research and target protein development. The source code and all datasets are available at https://github.com/QUST-AIBBDRC/LRF-DTIs/ for academic use.

1. Introduction

The identification of drug-target interactions is a hot area of research in the drug discovery process [1–3] because it plays an essential status in the development of finding new protein targeted drug and discovering new drug candidates [4,5]. Targets are specific protein molecules, such as receptors, enzymes, etc., that present in cells of human tissues and interact with drug molecules and confer drug effects [6]. The drug interacts with the target to influence the pharmacological effects of the target to achieve phenotypic effects [7,8]. The identification of drug-target interactions is crucial for understanding the mechanism of action of drugs, and the relationship between drug-targets is not yet fully understood [9].

Although the appearance of molecular medicine and the ongoing human genome project provide more chances for discovering new unknown target proteins. However, many drugs currently have incomplete target proteins. In the past few years, many efforts have been made to use experimental methods to verify drug-target interactions. Because drugs with satisfactory activity have unacceptable toxicity [10], few drug candidates can be marketed through the Food and Drug Administration (FDA) [11]. Furthermore, the use of experimental methods to determine drug-target interactions is time-consuming and expensive. Therefore, there is a pressing demand to develop new computational methods that can effectively identify these potential drug-target interactions.

Traditional calculation methods to predict drug-target interactions have been divided into the ligand-based methods and the target-based methods [12]. Ligand-based methods utilize ligand similarity to

* Corresponding author at: College of Mathematics and Physics, Qingdao University of Science and Technology, Qingdao 266061, China.
E-mail address: yubin@qust.edu.cn (B. Yu).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.ygeno.2018.12.007
Received 9 August 2018; Received in revised form 6 December 2018; Accepted 7 December 2018

organize pharmacological features and associations between target proteins, rather than based on sequence information or structural information of the target protein. Keiser et al. [13] proposed a method for predicting targets protein using the chemical two-dimensional structural similarity of ligands. However, this method has a significant disadvantage in that information on the protein domain is not used. Target-based methods are highly dependent on the integrity and accuracy of target structure information, and they can be obtained through experimental or computational simulations [14,15]. Molecular docking and scoring functions are a massive challenge in predicting drug-target interactions. Molecular docking is used to predict the optimal orientation of ligand-receptor recognition to form a stable complex, which in turn reflects the affinity between the ligand and the receptor by a scoring function. However, molecular docking methods cannot be applied to protein targets of unknown three-dimensional structure [16]. The above ligand-based prediction method and target-based prediction method respectively predict the ligand-target correlation using the structural similarity of the ligand and the stereostructure of the target protein. Although more and more potential drug targets and ligands have been discovered, screening for suitable active ligands from millions of small molecule compounds for different targets remains a considerable challenge. Therefore, it is necessary to develop a more efficient calculation method for drug-target interaction prediction based on machine learning.

The rapid development of machine learning techniques provides an effective and efficient method for predicting drug-target interactions [17]. How to extract effective feature-encoding protein sequences is an important part of the use of machine learning to predict drug-target interactions. In past studies, protein feature extraction methods were mainly based on amino acid sequences. Amino acid sequence-based feature extraction methods mainly include amino acid composition (AAC) [18], dipeptide composition (DC) [19], and pseudo-amino acid composition (PseAAC) [20,21]. In addition, in the past two decades, scholars have also proposed many characterization methods for pharmaceutical compounds. For example, Yap et al. [22] used CONCORD software to convert all 2D molecular structures of compounds into 3D structures and then used DRAGON Web to calculate for each compound molecule 18 categories including constitutional descriptors, geometric descriptors, and topological descriptors. Hammann et al. [23] used ChemAxon-related software to calculate 118 descriptors for each molecule, including elemental analysis, charge analysis, geometric features, and connectivity indicator. The National University of Singapore Yap et al. [24] wrote a JAVA program to develop a free description calculation software Padel-Descriptor. The software can calculate ten molecular fingerprints and more than 800 one-, two-, and three-dimensional descriptors.

In the process of predicting drug-target interactions, the processing of high-dimensional data is a complex issue. Excessive dimensions cause the model to become less efficient or even unable to work. Therefore, dimensionality reduction is an indispensable part of the data preprocessing process. The data dimensionality reduction methods are mainly classified into linear dimensionality reduction and nonlinear dimensionality reduction. The linear dimensionality reduction methods mainly include principal component analysis (PCA) proposed by Hotelling [25], linear discriminant analysis (LDA) proposed by Fisher [26], and linear discriminant analysis proposed by Michail Vlachos et al. [27]. Non-linear dimensionality reduction is mainly based on two methods. One is nuclear based, such as kernel principal component analysis (KPCA), kernel independent component analysis (KICA). The other is a manifold learning related method that recovers low-dimensional popular structures from high-dimensional sampling data. The main methods include isometric mapping (Isomap), local linear embedding (LLE), Laplacian eigenmap, etc.

According to these methods, in recent years, machine learning methods based on similarity and features have made significant progress in drug-target interactions prediction. Machine learning algorithms commonly used in prediction drug-target interactions, such as support vector machine [28,29], K-nearest neighbor [30] and random forest [31], etc. Furthermore, there are limitations in the process of feature extraction based on two types of methods. Considering the advantages of the two methods, some new calculation methods are also being proposed. Table 1 summarized some existing drug-target interactions prediction methods that are relevant to our proposed work.

In this study, we propose a novel method of drug-target interactions prediction based on Lasso dimensionality reduction and random forest classifier, which is called LRF-DTIs. First, we generate a pseudo-position specific scoring matrix (PsePSSM) to characterize the target protein from a position-specific scoring matrix (PSSM), while the FP2 fingerprint is used to represent each drug compound. Secondly, the positive and negative samples are constructed by using the drug-target interactions information on the extracted features. Then Lasso method is used to select features of the extracted features, and the SMOTE method is used to deal with the sample imbalance problem so that the characteristics of each drug-target pair are more significant and resolves the imbalance of the sample. Finally, the processed optimal feature vectors are input into a random forest classifier for the prediction of drug-target interactions. Through 5-fold cross-validation, the optimal parameters of the model are selected. According to the different dimensionality reduction algorithms and the selection of classifiers, a drug-target interactions prediction model is established. On the four datasets enzymes, ion channels, GPCRs, and nuclear receptors, satisfactory overall prediction accuracy are obtained. According to the comparison with other existing prediction methods, the experimental results show that the prediction method proposed in this paper can significantly improve the prediction effect of drug-target interactions.

2. Material and methods

2.1. Datasets

In this study, we use the gold standard dataset studied by Yamanishi et al. [32] as benchmark dataset, which can be downloaded from http://web.kuicr.kyoto-u.ac.jp/supp/yoshi/drugtarget/. In the gold standard dataset, drug-target interactions information has been obtained from KEGG BRITE [53], BRENDA [54], Super-Target [55] and DrugBank [56] databases. This dataset is split into four main datasets including enzymes, ion channels (IC), G-protein-coupled receptors (GPCR) and nuclear receptors (NR). The number of known drugs target in these classes is 445, 210, 223 and 54 respectively, and the number of target proteins in these classes is 664, 204, 95 and 26 respectively. After careful screening of these drugs and targets, a total of 5127 drug-target interaction pairs are obtained, and the number of known drug-target pairs interacting in each class is 2926, 1476, 635 and 90, respectively. Before further analysis, the detailed information on drugs and target proteins are obtained from the KEGG database. Each target protein is represented by an amino acid sequence and then saved in an .txt file. The chemical structure of each drug molecule is transformed as Mol file format, and then the file format is downloaded.

This paper also uses a dataset called DrugBank_approved [57], which contains all FDA-approved drugs and their corresponding protein targets in the DrugBank database [56]. The dataset contains 1556 drugs and 1610 targets. We treat the dataset targets and drugs in the same way as the gold standard dataset. Table 2 lists the information for all datasets.

2.2. Pseudo-position specific scoring matrix (PsePSSM)

Based on the extraction of amino acid sequences of proteins on enzymes, ion channels, GPCRs, and nuclear receptors, in order to express the characteristic information in the amino acid sequence as reasonably as possible, the pseudo-position specific scoring matrix (PsePSSM) features are used to encode the evolution and sequence

Table 1
The existing drug-target interactions prediction methods.

Author(s)	Datasets	Method	Performance parameters
Yamanishi et al. [32]	Gold standard dataset	a two-way network approach to integrate chemical and genomic spaces	SE, SP, PPV, AUC,
Bleakley et al. [33]	Gold standard dataset	bipartite local models to first predict target proteins of a given drug, then to predict drugs targeting a given protein	AUC, AUPR
Buza et al. [34]	Gold standard dataset, Kinase	represented drugs and targets in a similarly rich feature space, using ECkNN with center-aware regression techniques and constructed projection-based BLM sets.	AUC, AUPR
Shi et al. [35]	Gold standard dataset	some unstructured information used by enhancing similarity measures	AUC, AUPR
Wang et al. [30]	Gold standard dataset	stacked autoencoder of deep learning to fully extract raw data information from drug molecule structures and protein sequences	ACC, SE, Precision, MCC, AUC
Ezzat et al. [36]	DrugBank, Gold standard dataset	uses feature reduction and ensemble learning	AUC
Kuang et al. [37]	DrugBank	core matrix dimensionality reduction method based on the eigenvalue-based kernel matrix transformation method	AUC, AUPR
Mousavian et al. [38]	Gold standard dataset	protein feature: Bigram-PSSM drug feature: PubChem fingerprint classifier: SVM	AUC, AUPR, SE, SP, Precision
Yuan et al. [39]	DrugBank	combine feature-based with similarity-based methods	AUPR
He et al. [40]	Davis, Metz, KIBA	constructs features for each drug, each target and each drug–target pair	RMSE, AUC, AUPR, CI
Meng et al. [41]	Gold standard dataset	combined BIGP, PSSM, PCA, and RVM	Ac, SE, Precision, MCC
Li et al. [42]	Gold standard dataset	protein feature: PSSM drug feature: fingerprint developed discriminative vector machine (DVM) classifier	Pre, ACC, SE, MCC, AUC
Ezzat et al. [43]	Gold standard dataset	two methods of matrix decomposition using graph regularization	AUPR
Liu et al. [44]	Gold standard dataset	model the probability that a drug would interact with a target by logistic matrix factorization	AUC, AUPR
Zhang et al. [45]	KEGG	project the original space into a reduced space by randomly projecting the vectors of drug-target pairs with different dimensions	Recall, Precision, ACC
Luo et al. [46]	DrugBank, HPRD, Comparative Toxicogenomics, SIDER	uses vector space projection to predict new drug target interactions by integrating heterogeneous information	AUPR
Olayan et al. [47]	Gold standard dataset, DrugBank_FDA	use five trials of 10-fold cross-validation applies a random forest model on different graph-based features extracted from the DTI heterogeneous graph	Precision, Recall, SP, AUPR
Pahikkala et al. [48]	Gold standard dataset	made model formulation and evaluation settings better able to address the inherent complexity of predictive tasks in practical applications	AUC
He et al. [49]	Gold standard dataset	drug are encoded with functional groups and proteins encoded by biological features, feature selection: mRMR, nearest neighbor classification algorithm	Jackknife Cross-Validation Test
Laarhoven et al. [50]	Gold standard dataset	use the drug–target network define a Gaussian Interaction Profile (GIP) kernel	AUC, AUPR
Cheng et al. [51]	Gold standard dataset, DrugBank	three supervised inference named drug-based similarity inference, target-based similarity inference and network-based inference	ROC curves
Gönen et al. [52]	Gold standard dataset	joint bayesian formulation of projecting drug compounds and target proteins into a unified subspace using the similarities and estimating the interaction network	AUC

Table 2
The datasets used in our study.

Datasets	Drugs	Targets	Interactions
Enzyme	445	664	2926
IC	210	204	1476
GPCR	223	95	635
NR	54	26	90
DrugBank_approved	1556	1610	5877

information of the protein sequence.

For a target protein sequence P with L amino acid residues, we use the position-specific scoring matrix (PSSM) as its descriptor introduced by Jones [58]. The position-specific scoring matrix (PSSM) with a dimension of L × 20 can be expressed as:

$$\mathrm{P}_{\mathrm{PSSM}} = \begin{bmatrix} E_{1\to1} & E_{1\to2} & \cdots & E_{1\to j} & \cdots & E_{1\to20} \\ E_{2\to1} & E_{2\to2} & \cdots & E_{2\to j} & \cdots & E_{2\to20} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ E_{i\to1} & E_{i\to2} & \cdots & E_{i\to j} & \cdots & E_{i\to20} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ E_{L\to1} & E_{L\to2} & \cdots & E_{L\to j} & \cdots & E_{L\to20} \end{bmatrix} \quad (1)$$

where j is the 20 native amino acid types, $E_{i\to j}$ is the score of the residue of the i-th position in the amino acid sequence being mutated to the j-th amino acid residue, which can be searched using the PSI-BLAST [59] in the Swiss-Prot database. A positive score means that the corresponding residue is mutated more frequently than expected, and the negative score is just the opposite. In this process, the parameters of PSI-BLAST are set as: the threshold of E-value is 0.001, the maximum number of iterations for multiple searches is 3, and the rest of the parameters are set by default.

Each element in the original PSSM matrix is normalized to the interval (0, 1) using Eq. (2).

$$\overline{E}_{i\to j} = \frac{1}{1 + \exp(-E_{i\to j})} \quad (2)$$

However, according to the PSSM descriptors, proteins with different lengths will correspond to matrices with different numbers of rows. To make the PSSM descriptor a uniform representation, one possible method is to represent the protein sample P as

$$\overline{\mathrm{P}}_{\mathrm{PSSM}} = [\overline{E}_1, \overline{E}_2, \cdots, \overline{E}_{20}]^T \quad (3)$$

where T is the transpose operator,

$$\overline{E}_j = \frac{1}{L}\sum_{i=1}^{L} \overline{E}_{i\to j} \quad (4)$$

where $\overline{E}_{i\to j}$ is the score of the residue of the i-th position in the amino acid sequence changed to the j-th amino acid residue after normalization, $\overline{E}_j$ is the average score of the amino acid residue in protein P being mutated amino acid type j during the process of evolution. However, if Eq. (3) is used to represent protein P, all sequence information will be lost. To avoid complete loss of sequence information, in combination with the concept of the pseudo-amino acid composition (PseAAC) initially proposed by Chou [60], we use position-specific scoring matrix (PsePSSM) to represent the protein P

$$\mathrm{P}^{\lambda}_{\mathrm{PsePSSM}} = [\overline{\mathrm{E}}_1, \overline{\mathrm{E}}_2, \cdots, \overline{\mathrm{E}}_{20}, G_1^1, G_2^1, \cdots, G_{20}^1, \cdots, G_1^{\lambda}, G_2^{\lambda}, \cdots, G_{20}^{\lambda}]^T \quad (5)$$

where

$$G_j^{\lambda} = \frac{1}{L-\lambda}\sum_{i=1}^{L-\lambda}[\bar{E}_{i\to j} - \bar{E}_{(i+\lambda)\to j}]^2 (j = 1, 2, \cdots, 20; 0 \le \lambda \le L) \quad (6)$$

where G_j^{λ} is correlation factor of the j-th amino acid and the continuous distance along the protein sequence is λ. This means that G_j^1 represents the relevant factor coupled along the most continuous PSSM score on the protein chain of amino acid type j, G_j^2 represents the second closest PSSM score by coupling, and so on. Therefore, a protein sequence can be expressed as Eq. (5) using PsePSSM and generates a $20 + 20 \times \lambda$-dimensional feature vector.

The PsePSSM algorithm converts the protein sequences with different lengths in the dataset into vectors with the same dimension after feature extraction. In this paper, λ is set to 10 after executing the optimization program for the training sample by 5-fold cross-validation (see in section 3). So, to facilitate the implementation of the following algorithm, the characteristic dimension of each target protein is 220.

2.3. FP2 molecular fingerprint

To express drug compounds, different types of descriptors can be defined based on various types of drug properties. And in specific studies, it has been indicated that some descriptors are molecular structure fingerprints that effectively represent the drug [61–63].

The molecular fingerprint of the drug can be extracted through the following steps:

Step 1: For a given drug, a Mol file format containing detailed chemical structure information of the drug is obtained from the KEGG database (http://www.kegg.jp/kegg/) by using its drug code.

Step 2: Download OpenBabel Software [64]. The Open Babel software toolkit (available from http://openbabel.org/) can process chemical data between different formats, allowing users to search, convert, analyze and store different chemical data areas in the fields of molecular modeling, chemical analysis, materials science, biochemistry and so on. OpenBabel can generate four types of fingerprints: FP2, FP3, FP4, and MACCS. In the present study, FP2 format fingerprints were used as drug descriptors.

Step 3: The drug molecules with Mol file format are converted into the FP2 format molecular fingerprint file format using the OpenBabel software. The FP2 format molecular fingerprint sequence is a hexadecimal digit sequence of length 256 that can be converted to a decimal digit sequence between 0 and 15 as a drug molecule 256-dimensional vector.

2.4. Constructing positive and negative samples

The drug-target interactions network is generally regarded as a bipartite graph in which each node represents a target protein or drug molecule and each side (only target protein and drug can be linked by a side) indicates a drug-target interaction. All the known interacting drug-target pairs in this paper are regarded as positive samples, in order to isolate the positive and negative samples from the classification framework of a bipartite graph, while the residual sealed interacting drug-targets pairs are considered negative samples. Because the number of interacting drug-target pairs is far less than the non-interacting drug-target pair, the positive and negative samples constructed are fantastically unbalanced, and the number of positive samples is far less than the number of negative samples. To work out the deviation because of imbalanced samples, this paper uses the SMOTE method to make random oversampling on a small scale of positive samples and random undersampling on a large scale of negative samples so that the positive and negative samples are balanced.

2.5. Representing drug-target pair

Molecular representation and coding are one of the decisive steps that affect the results of the binary classification system. To encode each drug-target pair, representations of drug and target protein molecules have been described. Next, to extract more information about the mechanism of action between the drug and the target protein, this article takes into account the structural information from the drug and the target protein to effectively represent the interaction based on the constructed positive and negative samples. Finally, in this study, we used PsePSSM to represent proteins and FP2 fingerprints to represent drugs. Each target was composed of 220-dimensional, and each drug consisted of 256-dimensional. Hence, each interaction (positive or negative) information is represented as a 476-dimensional vector by a combination of the drug molecule fingerprint and the target protein structure information.

2.6. Lasso method

We know that the characteristics of the 476-dimensional data contained in each drug-target pair may contain some noise and redundant information, which may have a negative impact on the performance of the predictive model. To remove useless information and extract the most discriminating features from the descriptors of drug-target pairs, we use the Lasso method to reduce the dimensionality of the original data features.

The Lasso method is a compression estimation method proposed by Tibshirani [65]. The basic idea of Lasso is to minimize the penalty function under the constraint that the sum of the absolute values of the regression coefficients is less than a constant. The specific dimensionality reduction method based on Lasso is as follows:

For a given sample datasets

$$X = [x_1, x_2, \cdots, x_N]^T \in R^{N\times d} \quad (7)$$

where x_i is the eigenvector of the i-th sample, N is the number of samples, d is feature dimension.

$Y = [y_1, y_2, \cdots, y_N] \in R^N$ represents the corresponding response vectors for these samples. For the study of binary classification problem in this paper, $y_i \in \{0, 1\}$is the class label of the sample. And then the objective function of the Lasso dimensionality reduction method optimization is [66]:

$$\min_w \| Y - X^T w \|_2^2 + \lambda \| w \|_1 \quad (8)$$

where w is regression coefficients of eigenvectors. Regularization term $\|w\|_1$ adopt L_1-paradigm which will generate a sparse solution in the feature space, that is to say, the coefficients corresponding to the irrelevant and redundant features will be set to 0, and the features are corresponding to the non-zero coefficients will be retained for subsequent classification. $\lambda > 0$ is a regularization parameter in the model and its role is to balance the complexity of the model and the degree of data fitting.

2.7. SMOTE method

Because there is a high degree of imbalance in the samples used in this study and the number of negative samples is significantly higher than the number of positive samples, the imbalance of the samples used in the study is shown in Table 3.

From Table 3, we can see that for the four datasets used in this study, there are severe imbalance problems between the positive and

Table 3
The imbalance of the samples used in the dataset.

Datasets	Positive samples	Negative samples	Sample ratio
Enzyme	2926	292,554	99.98
IC	1476	41,364	28.02
GPCR	635	20,550	32.36
NR	90	1314	14.60

negative samples. In the four datasets, the imbalance of the enzyme dataset is the most serious and the ratio of negative samples to positive samples reach 99.98%, while the dataset with the least degree of imbalance in the four datasets is a nuclear receptor, and the proportion of negative and positive samples also reach 14.60%. Usually, random undersampling and random oversampling are the two most important methods for dealing with imbalanced datasets. To handle the problems brought by data resampling, this paper uses the Synthetic Minority Oversampling Technique (SMOTE) method proposed by Chawla et al. [67]. The SMOTE method is a method of randomly undersampling a large scale of samples while randomly oversampling a small scale of samples. Its detailed operation is as follows:

For each positive sample x, search its k (k value usually takes 5) nearest neighbors, if the sampling rate is N, then randomly select N samples from the k nearest neighbors, denoted as $y_1, y_2, \ldots, y_N$. Random linear interpolation is performed on the line segments formed by x and y_i $(i = 1, 2, \cdots, N)$ to obtain a new positive sample z_i, as shown in the following equation:

$$z_i = x + rand\,[0,1] \times (y_i - x) \tag{9}$$

where $rand(0,1)$ means to generate a random number between 0 and 1. Add these newly synthesized samples to the original positive sample and form a new, more balanced dataset. The visual display of SMOTE sample synthesis method is shown in supplementary materials Fig. S1.

2.8. Random forest

Random Forest (RF) is a new machine learning method proposed by Breiman [68] in 2001. It builds the Bagging integration based on the decision tree and further introduces the random attribute selection in the decision tree training process. As a classifier, RF has a strong classification ability. In recent years, the RF method has been widely used in various problems such as classification, prediction, the importance of variables, dimensionality reduction and abnormal point detection [69–72]. Specifically, the traditional decision tree selects an optimal attribute from the set of attributes of the current node (assuming that there are d attributes) when selecting the attribute. While in RF, for each node in the base decision tree, a subset containing attributes is first randomly selected from the attribute set of the node, and then an optimal attribute is selected from this subset for the division.

Assuming that there are N sample units and M variables in the training sample, the principle and solution algorithm of the random forest are as follows:

1) N sample units are randomly selected from the training sample to generate a large number of decision trees;
2) $m < M$ variables are randomly selected at each node of each tree as a candidate variable for dividing the node, and the number of variables at each node should be consistent;
3) Each tree grows maximally without any trimming;
4) The category of the terminal node is determined by the mode corresponding to the node.
5) The generated multiple classification trees constitute a random forest, and the new data is discriminated and classified by a random forest classifier. The classification result is determined by the number of votes cast by the tree classifier.

The flow chart of the random forest classification algorithm is shown in supplementary materials Fig. S2.

2.9. Performance evaluation

In this study, we used 5-fold cross-validation to evaluate the performance of the model. The basic method is to randomly divide the samples into five subsets, and the number of subsets is roughly the same. Then, one subset is taken as the test sample, and the remaining four subsets are used as the training sample until all the samples in each subset are tested. At the same time, the confusion matrix (also known as contingency table or error matrix [73]) will be used to represent the predictive performance of the predictor.

In confusion matrix, TP, FP, TN, and FN are abbreviations for true positives, false positives, true negatives, and false negatives, respectively. In this paper, Accuracy (ACC), Specificity (SP), and Sensitivity (SE) are used as evaluation criteria for the prediction model proposed in the experiment. These indicators are defined as follows:

- Accuracy

Shows the ratio of the number of samples correctly classified by the classifier to the total number of samples for a given set of test data. It is defined as

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

- Sensitivity

Shows the proportion of all positive examples that have been classified correctly and measures the classifier's ability to identify positive examples. It is defined as

$$SE = \frac{TP}{TP + FN}$$

- Specificity

Shows the proportion of all negative examples that have been classified correctly and measures the classifier's ability to identify negative examples. It is defined as

$$SP = \frac{TN}{TN + FP}$$

Moreover, AUC is another important tool that can be used to assess the accuracy of classification. AUC can be obtained by summing the areas under the ROC curve. The ROC curve is a plot of true positive rate (TPR) and false positive rate (FPR) at multiple threshold settings. Typically, by establishing a prediction model, a scalar of the predicted target tuple can be calculated to represent the probability that the tuple is true. By setting different conditions, different thresholds can be calculated to predict the TPR and FPR of the model. Finally, with TPR as the Y-axis and FPR as the X-axis, the desired pattern can be drawn.

However, it is well known that ROC curves often overestimate predictive performance, especially in the biomedical field, where PR curves are introduced to reevaluate their classifiers. According to the confusion matrix, the precision and recall ratio can be defined. These two indicators are the fundamental factors of the PR curve. The definition is as follows:

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

The precision and recall are a contradiction measure. In many cases, we can sort the samples according to the prediction results of the classifier, and predict the samples one by one as a positive example in this order, and then calculate the current recall rate and precision rate each time. The precision is the vertical axis, and the recall rate is plotted on the horizontal axis. The precision-recall curve is obtained, which is referred to as the “PR curve”. The PR curve visually shows the recall rate and precision of the classifier on the sample as a whole. When comparing, if the PR curve of one classifier is completely encased by another learner, it can be concluded that the performance of the latter is better than the former, that is, the surrounding area of the two

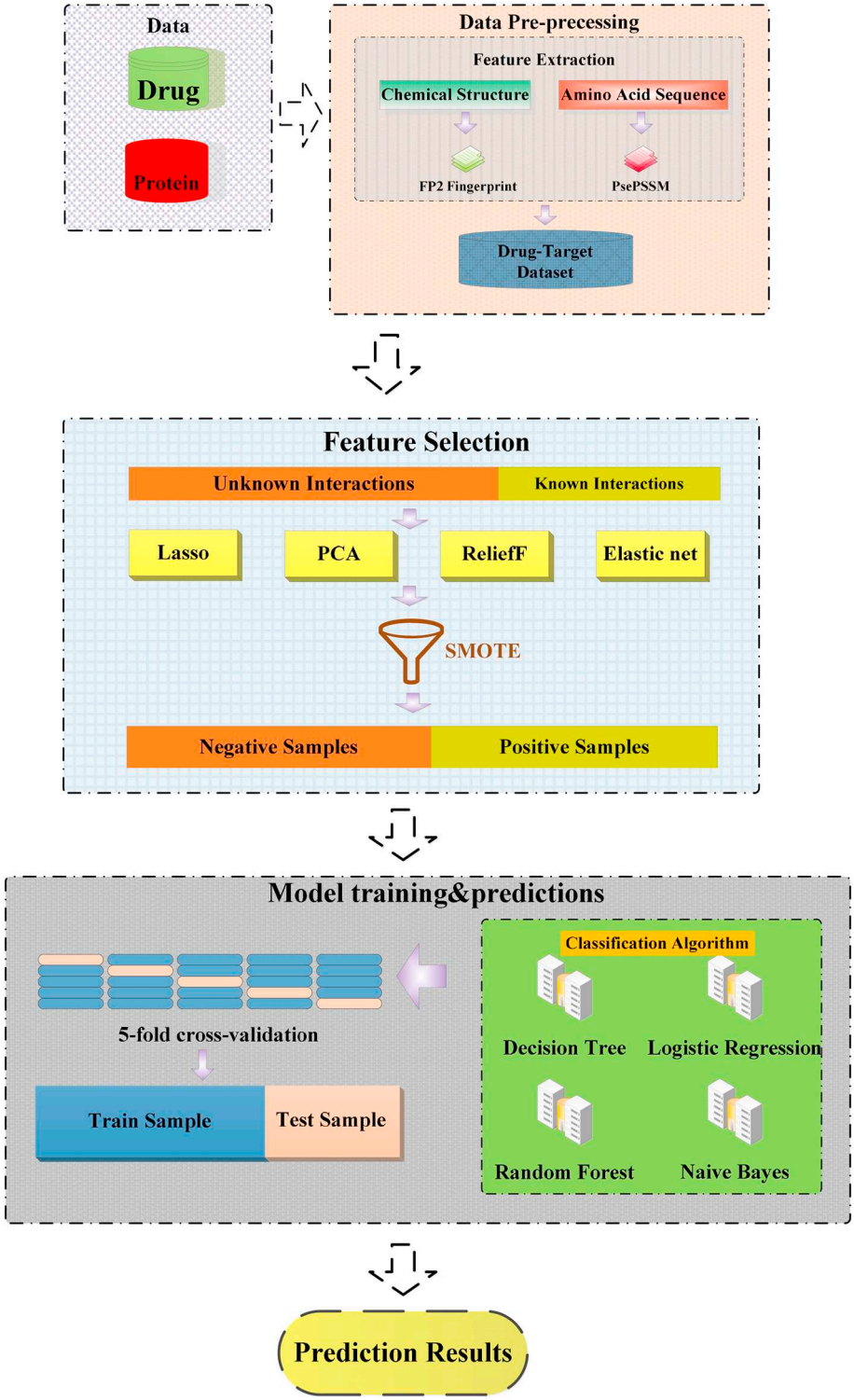

Fig. 1. The process of constructing a drug-target interactions prediction model.

curves is compared, and the size of the enclosed area is defined as AUPR value, the larger the value, the better the learner performance.

For convenience, the drug-target interactions prediction method proposed in this study is called LRF-DTIs. The general framework is shown in Fig. 1. Our experimental environment is Intel(R) Xeon(R) CPU E5–2650 v3 @ 2.30GHz 32.0GB of RAM, and R x64 3.4.3 programming implementation.

The specific steps of LRF-DTIs for drug-target interaction prediction method are described as:

1) Enter the protein sequence and drug's Mol file format chemical structure for each of the drug-target datasets;
2) Using Open Babel open-source chemical toolbox to extract FP2 fingerprint features of drug molecules, generating 256-dimensional feature vectors. And the PsePSSM is used to extract the $20 + 20 \times \lambda$-dimensional feature vector from the amino acid sequence of the target protein. Fusion of two features to represent drug-target pairs, generating a $(256) + (20 + 20 \times \lambda)$-dimensional feature vector;

3) According to the accuracy of the prediction, the optimal parameter λ in PsePSSM is selected;
4) Lasso, PCA, ReliefF, and Elastic net methods are used to reduce the dimension of the numerical sequence extracted after 2) to eliminate the redundant information in the sequence, and then select the optimal dimensionality reduction method according to the model evaluation indicator.
5) Apply the SMOTE method to balance drug-target datasets;
6) The optimal feature vectors after dimensionality reduction are input into four classification algorithms: random forest, Naïve Bayes, decision tree and logistic regression. The optimal algorithm is selected to perform drug-target interactions prediction;
7) In gold standard dataset, in the light of the optimal parameters of the models obtained in 3), 4) and 6), SE, SP, ACC, AUC are calculated, and the predicted performance of the model is compared with other methods;
8) Test the prediction effect of the new interaction sample in the model on NR dataset;
9) Validate the prediction model in this paper using the DrugBank_approved dataset.

3. Results and discussion

3.1. PsePSSM method selection of optimal parameter

In recent years, many scholars have used PsePSSM to extract protein sequences. The selection of parameters λ in evolutionary information can directly affect the prediction accuracy of the model. This article uses the PsePSSM method for feature extraction of target proteins. To find the optimal value of λ, we set values from 6 to 14 in order. For different λ values, datasets are processed by Lasso dimensionality reduction and SMOTE method and use random forest classifiers to separate enzymes, ion channels, GPCRs, and nuclear receptors. The prediction accuracy and average prediction accuracy of the four datasets under different parameters λ are obtained by using 5-fold cross-validation. The results are shown in Table 4.

As can be seen from Table 4, using PsePSSM method for feature extraction of target proteins, the prediction accuracy and the average prediction accuracy of the drug-target interaction for the four datasets also change with different parameters while other conditions remain unchanged. In the enzyme dataset, the forecast accuracy rate fluctuates between 97.92% and 98.03%, and the highest forecast accuracy rate is 96.71% when λ = 10. In the ion channel dataset, the forecast accuracy rate fluctuates between 96.92% and 97.22%, and the highest accuracy reached 97.22% when λ is 8. In the GPCR dataset, the forecast accuracy rate fluctuates between 95.49% and 96.08%, and the highest accuracy rate reached 96.08% when λ is 10. In the nuclear receptor dataset, the forecast accuracy rate fluctuates between 93.52% and 95.83%. The highest prediction accuracy rate is 95.83% when λ is 10. To find out the best λ value more intuitively, the prediction accuracy rate and average prediction accuracy rate when selecting different values for the four datasets of the enzyme, ion channel, GPCR, and nuclear receptor is shown in supplementary materials Fig. S3.

To maintain the consistency of the parameters in the model, we consider the λ value corresponding to the highest average accuracy as the optimal parameter, according to the average prediction accuracy rate of the four datasets. The average prediction accuracy of the four datasets is up to 96.71%, when λ = 10. In summary, the optimal parameter λ value of 10, using the PsePSSM method to extract features of each target protein can get 220-dimensional feature vectors.

Table 4
Prediction accuracy and average prediction accuracy of four datasets under different parameters.

λ	5-fold cross-validation (%)								
	6	7	8	9	10	11	12	13	14
Enzyme	98.01	98.02	97.99	98.02	**98.03**	97.98	97.97	97.99	97.92
IC	97.19	97.17	**97.22**	97.06	97.15	96.97	96.92	97.04	96.97
GPCR	95.84	95.49	95.59	**96.08**	95.83	95.85	95.55	95.80	95.97
NR	94.07	95.00	94.63	95.02	**95.83**	94.63	95.28	94.54	93.52
Average	96.28	96.42	96.36	96.55	**96.71**	96.36	96.43	96.34	96.10

Note: the corresponding accuracy presented after feature selection method, parameters in SMOTE and classification method of the pipeline has already been optimized.

3.2. Effect of dimensionality reduction on results

The dimensionality reduction algorithm can process the complex redundant information in the feature vector, which can improve the prediction accuracy to some extent. Based on the 220-dimensional feature vectors obtained in PsePSSM and the 256-dimensional feature vectors generated after extracting the FP2 molecular fingerprint, selecting an appropriate method to improve the accuracy rate is of great significance to the establishment of the prediction model. This paper compares the impact of PCA, ReliefF, Lasso, and Elastic net on the prediction results, and chooses the optimal dimensionality reduction method. Among them, after many trials, when the dimensionality reduction is performed using PCA, the feature with a contribution rate of more than 85% is selected. When selecting features by using ReliefF, the first 30% of the features are selected after sorting according to the weight size. When using Lasso and Elastic net for dimensionality reduction, features corresponding to non-zero feature coefficients are selected, and other parameters are default parameters. To ensure a fair comparison during the dimensionality reduction process, in this experiment, random forest classifiers and 5-fold cross-validation are used to predict and evaluate model performance. At the same time, to prevent the problem of over-fitting, when using the 5-fold cross-validation method in the following discussion of this paper, we repeat the prediction model 10 trials, and the average of 10 trials results are used as the final result of the performance of the evaluation model. The prediction results are shown in Table 5.

From Table 5, it can be seen that the Lasso and elastic net methods have great differences in the four-dimensionality reduction methods. The Lasso method has more obvious advantages in the four methods. For the dataset enzymes, using Lasso as the model's dimensionality

Table 5
Comparison of different parameter selection methods in four datasets.

Datasets	Methods	5-fold cross-validation			
		AUC	SE (%)	SP (%)	ACC (%)
Enzyme	Lasso	**0.9982**	**97.66**	**98.51**	**98.09**
	PCA	0.9417	87.30	87.95	87.62
	ReliefF	0.8700	86.02	80.94	83.48
	Elastic net	0.8605	80.98	82.36	81.67
IC	Lasso	**0.9965**	**96.71**	**97.93**	**97.32**
	PCA	0.9305	85.44	86.96	86.20
	ReliefF	0.9487	87.24	97.68	92.46
	Elastic net	0.8040	74.67	70.66	72.67
GPCR	Lasso	**0.9918**	**95.26**	**96.11**	**95.69**
	PCA	0.9122	83.56	86.31	84.93
	ReliefF	0.8358	79.06	78.52	78.79
	Elastic net	0.7785	67.77	78.19	72.93
NR	Lasso	**0.9559**	**93.85**	95.81	**94.88**
	PCA	0.9049	80.44	89.00	84.72
	ReliefF	0.9185	78.96	**98.46**	88.71
	Elastic net	0.8418	70.06	89.96	80.01

Note: the corresponding Performance Parameters (AUC, SE, SP, ACC) presented after Parameters λ, parameters in SMOTE and classification methods of the pipeline have already been optimized.

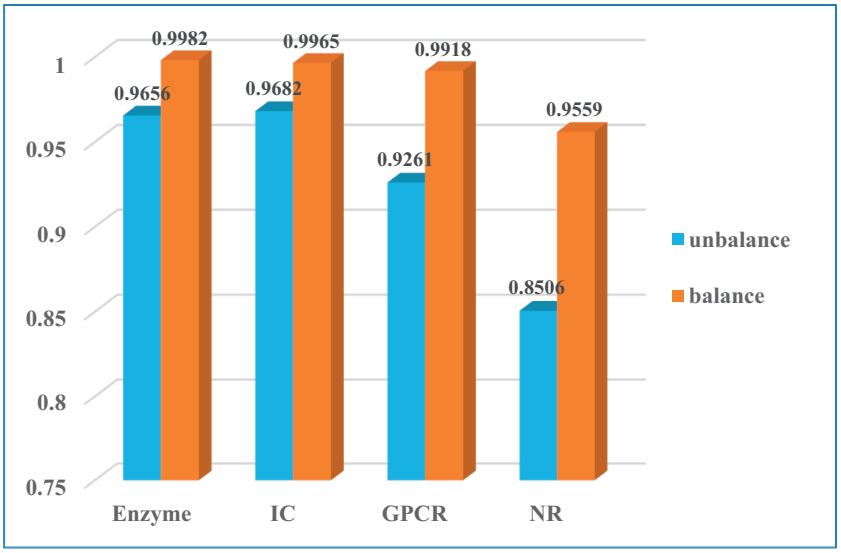

Fig. 2. Comparison of model AUC before and after SMOTE optimization for the four datasets.

reduction algorithm, the SE, SP, and ACC of the model reach 97.66%, 98.51%, and 98.09%, respectively. These evaluation indicators are superior to the other three dimensionality reduction algorithms. Similarly, for the dataset ion channel, the SE, SP, and ACC reach 96.71%, 97.93%, and 97.32%, respectively. These evaluation indicators are superior to the other three dimensionality reduction algorithms. For the dataset GPCR, using the Lasso algorithm as the model's dimensionality reduction algorithm, the SE, SP, and ACC reach 95.26%, 96.11%, and 95.69%, respectively. These evaluation indicators are better than the other three dimensionality reduction algorithms. For the dataset nuclear receptors, the Lasso algorithm is used as the model's dimensionality reduction algorithm. The SE, SP, and ACC of the model reach 93.85%, 95.81%, and 94.88%, respectively, the indicators are higher than the other three dimensionality reduction algorithms except that the SP results are 2.65% lower than ReliefF. The Lasso method in all four datasets can exceed 90% in terms of assessment. In terms of parameter indicator ACC, the Lasso method has more than 10% improvement over the elastic network. Although indicators SE and SP can exceed the Lasso dimensionality reduction method in some datasets, it still has certain advantages in integrating all data. Meanwhile, using the Lasso method for dimensionality reduction has obvious advantages over other methods on the indicator AUC. To compare the size of the AUC more intuitively, the ROC curves on the enzyme, IC, GPCR, and NR datasets for four different dimensionality reduction methods through 10 trials of 5-fold cross-validation are shown in supplementary materials Fig. S4.

Using the Lasso method to reduce dimension can effectively reduce information redundancy and delete some unimportant features, which helps to avoid overfitting of the prediction model and reduce the complexity of the training model. Therefore, in this study, we determine that Lasso is the best dimensionality reduction algorithm for this study.

3.3. Analysis of imbalanced and balanced datasets

In general, the classifier's prediction results are more biased towards a large number of categories, and there are serious imbalance problems between the datasets, which will lead to a decrease in the evaluation performance of the prediction model. However, in the available datasets researched in this paper, the number of drug-target pairs that their interactions are known is significantly smaller than the number of drug-target pairs that are not interacting with each other, which will lead to a decrease in the prediction accuracy of future prediction models. Therefore, to improve the generalization ability of the classifier and reduce the influence of the bias caused by the imbalance of the dataset, we attempt to use the SMOTE method to perform random undersampling on the majority of samples that do not interact with each other before selecting the appropriate classifier. At the same time, random oversampling is performed on a few samples of known interactions to overcome the problem of dataset imbalance.

To achieve data equalization by reasonably deleting or adding some samples, and then reducing the negative impact of data imbalance on the classifier, the SMOTE method is used to balance the positive and negative samples for the Lasso feature-selected data. In the SMOTE method, we set the oversampling parameter to 500, the undersampling parameter to 120, and the nearest neighbor algorithm parameter to 5. For the sake of fairness, this section uses the target evolutionary information obtained from PsePSSM and the drug molecule characteristics obtained from the FP2 molecular fingerprint. After Lasso's dimensionality reduction data, we conduct a comparative analysis of balanced data and unbalanced data. Based on the ten trials of 5-fold cross-validation, the evaluation indicators (i.e. AUC、SE、SP and ACC) of the model obtained by using the random forest classification algorithm are shown in supplementary materials Table S1.

Due to the imbalance of the dataset, the indicator ACC has no significance in measuring the merits of the algorithm. Meanwhile, since the number of positive examples is much less than the number of negative examples and the indicators SE and SP are proportional to the correct proportion of positive and negative examples in the sample. Therefore, among all the above-mentioned indicators, the indicator that can reasonably measure the evaluation performance of the prediction model is AUC. To more accurately reflect the effect of data balance on the prediction performance of the model, Fig. 2 depicts the comparison of the AUC of the enzyme, ion channel, GPCR, and nuclear receptor datasets on unbalanced datasets and balanced datasets.

Through the Fig. 2, we can see that the dataset after SMOTE processing has achieved a significant improvement in the indicator AUC. After the four datasets are balanced, the evaluation indicator AUC increases by 2.83%–10.53%. For the dataset enzyme, the indicator AUC increases by 3.26% after balancing. For dataset ion channels, the indicator AUC increases by 2.83% after balancing. For the dataset GPCR, the indicator AUC increases by 6.57% after balancing. For the dataset nuclear receptors, after balancing, the indicator AUC increases by 10.53%. Therefore, through the above comparative analysis, we can obtain that the model has a greater improvement in the prediction performance after the SMOTE processing.

3.4. Effect of classification algorithm on results

For a given prediction task, performance depends not only on characteristics but also on the choice of classifiers which play an essential role in predicting the quality of the model. In this study, four classification algorithms are investigated: random forest, decision tree, Naïve Bayes, and logistic regression. These four classification algorithms are used to compare the four datasets of the enzyme, ion channel, GPCR, and nuclear receptor. Under the 5-fold cross-validation test, after many experiments, the parameters selected by the classifier are as follows: the random forest creates an integration that generates 500 lessons, the decision tree uses the C5.0 algorithm to control the number of bootstrap cycles sets to 1, the Laplacian estimate of the Naïve Bayes is 0. Logistic regression classifiers are classified by constructing logistic regression models. And all other parameters use default parameters. To ensure fairness, the target protein sequences in the datasets are extracted by PsePSSM for the four datasets constructed in this paper. After feature extraction of FP2 molecular fingerprints for drugs, the Lasso dimensionality reduction method and SMOTE treatment unbalance method is used to further process. Under the 10 trials of 5-fold cross-validation, the main prediction results of the four different classification algorithms are shown in Table 6.

From Table 6, it can be seen that the prediction performance using the random forest model has obvious advantages on the four datasets. For the dataset enzyme, the random forest algorithm is used as the model's prediction algorithm. The SE, SP, and ACC of the model reach 97.66%, 98.51%, and 98.09%. These evaluation indicators are superior to the other three classification algorithms. However, using the Naïve Bayes algorithm, the SE, SP, and ACC are 73.13%, 67.58% and 70.36%, which are 24.53%, 30.93%, and 27.73% lower than using random forest algorithm. For the dataset ion channel, the SE, SP, and ACC of the model reach 96.71%, 97.93% and 97.32% in the random forest model. The results are better than the other three classification algorithms. Compared with the Naïve Bayes classifier, the model increases by 16.45%, 41.42% and 28.94% on SE, SP, and ACC. Similarly, for the dataset GPCR, the SE, SP, and ACC of the random forest model reach 95.26%, 96.11%, and 95.69%, respectively. The results of these indicators are better than the other three classification algorithms. It is 16.19%, 33.86%, and 25.03% higher than the Naïve Bayes evaluation indicator. For dataset nuclear receptors, the SE, SP, and ACC of the model reach 93.85%, 95.81%, and 94.88%, and the results are better than the others. For the indicator AUC, the results obtained by the random forest classifier on the four datasets are all above 95%, significantly exceeding the other three classification algorithms. The ROC curves for four different classification algorithms on the enzyme, ion channel, GPCR, and nuclear receptor datasets through 10 trials of 5-fold cross-validation are shown in supplementary materials Fig. S5.

In order to more accurately measure the predictive performance of the classification algorithm, Fig. 3 shows one of the PR curves of the four datasets of enzyme, ion channel, GPCR and nuclear receptor under four different classification algorithms in 10 trials of 5-fold cross-validation and the others PR curves of the four datasets of enzyme, ion channel, GPCR and nuclear receptor under four different classification algorithms in 10 trials of 5-fold cross-validation are shown in supplementary materials Fig. S6.

As can be seen from Fig. 3, when using the random forest as the classification algorithm, the PR curves of the four datasets of the enzyme, ion channel, GPCR, and nuclear receptor surround the PR curves of the other three classifiers. The corresponding AUPR value is also larger, and the performance of using a random forest as a classifier is better.

The AUPR of the random forest in the enzyme dataset reaches 0.9983, which are 2.83%, 13.55%, and 21.82% higher than that of the decision tree, logistic regression, and Naïve Bayes, respectively. Besides, in the ion channel dataset, the AUPR of the random forest model reaches 0.9964. It is 4.65%, 13.50%, and 27.76% higher than the decision tree, logistic regression model and Naïve Bayes. In the GPCR dataset, the AUPR of the model is 0.9913, which are 5.35%, 16.09%, and 27.60% higher than that of the decision tree, logistic regression and Naïve Bayes, respectively. For the nuclear receptor dataset, the AUPR of the model reaches 0.9867. It is 9.60%, 23.33% and 30.70% respectively higher than that of the other three models.

Because the random forest model is easy to use and it is an ensemble learning based on the decision tree classifier. It can get a more accurate prediction result with good comprehensive features, and it has less tendency for overfitting.

Table 6
Four classifiers under different classifier prediction models to evaluate performance results.

Datasets	Model	5-fold cross-validation				
		AUC	AUPR	SE (%)	SP (%)	ACC (%)
Enzyme	Random Forest	**0.9982**	**0.9983**	**97.66**	**98.51**	**98.09**
	Decision Tree	0.9754	0.9700	95.52	93.76	94.64
	Logistic Regression	0.9151	0.8628	83.45	87.47	85.46
	Naïve Bayes	0.7817	0.7801	73.13	67.58	70.36
IC	Random Forest	**0.9965**	**0.9964**	**96.71**	**97.93**	**97.32**
	Decision Tree	0.9555	0.9499	92.38	90.81	91.60
	Logistic Regression	0.9048	0.8614	81.27	87.30	84.29
	Naïve Bayes	0.7423	0.7188	80.26	56.51	68.38
GPCR	Random Forest	**0.9918**	**0.9913**	**95.26**	**96.11**	**95.69**
	Decision Tree	0.9464	0.9378	91.20	88.98	90.09
	Logistic Regression	0.8749	0.8304	78.22	83.54	80.88
	Naïve Bayes	0.7570	0.7153	79.07	62.25	70.66
NR	Random Forest	**0.9559**	**0.9867**	**93.85**	**95.81**	**94.88**
	Decision Tree	0.9207	0.8907	89.81	87.41	88.61
	Logistic Regression	0.8375	0.7534	76.00	83.31	79.66
	Naïve Bayes	0.7454	0.6797	83.13	50.22	66.68

3.5. Comparison with other methods

For the prediction of drug-target interactions, many prediction methods have been proposed. To more objectively evaluate the effectiveness of the proposed drug-target interaction prediction model proposed in this paper, we compared the prediction performance with the other prediction method using the same datasets based on 5-fold cross-validation repeat the prediction model 10 trials, and the average of results is used as the final result. Table 8 details the comparison between the methods presented in this paper with other prediction methods on enzymes, ion channels, GPCRs, and nuclear receptors datasets.

From Table 7, we can see that for the dataset of the enzyme, the accuracy rate of the prediction method of this method is 98.09%, which is 0.36%–7.78% higher than other prediction methods. The accuracy of our prediction model is 0.36% higher than that of Meng et al. [41], 7.78% higher than that proposed by Cao et al. [74] and the overall accuracy of the model is significantly improved. In addition, the AUC of the prediction model reaches 0.9982, which is 6.94% higher than the algorithm proposed by Li et al. [42] and 2.02% higher than the algorithm proposed by Laarhoven et al. [50]. In the ion channel dataset, the accuracy of prediction in this paper is 97.32%, which is 4,20%–8.41% higher than other prediction methods. In contrast, our method is 8.41% higher than the prediction method proposed by Cao et al. [74] and 5.59% higher than the method proposed by Li et al. [42]. Our prediction results are better than other methods. For the GPCR dataset, the accuracy rate is 95.69%, which is 6.32%–11.01% higher than other methods. The prediction accuracy of this paper is 8.92% higher than that of Meng et al. [41] and 6.32% higher than that of Li et al. [42]. The prediction accuracy of our model is higher than the others. At the same time, regarding parameter AUC, the method proposed in this paper reaches 0.9918, which is 10.62% and 10.16% higher than that of Li et al. [42] and Cao et al. [74]. Compared with the method of Laarhoven

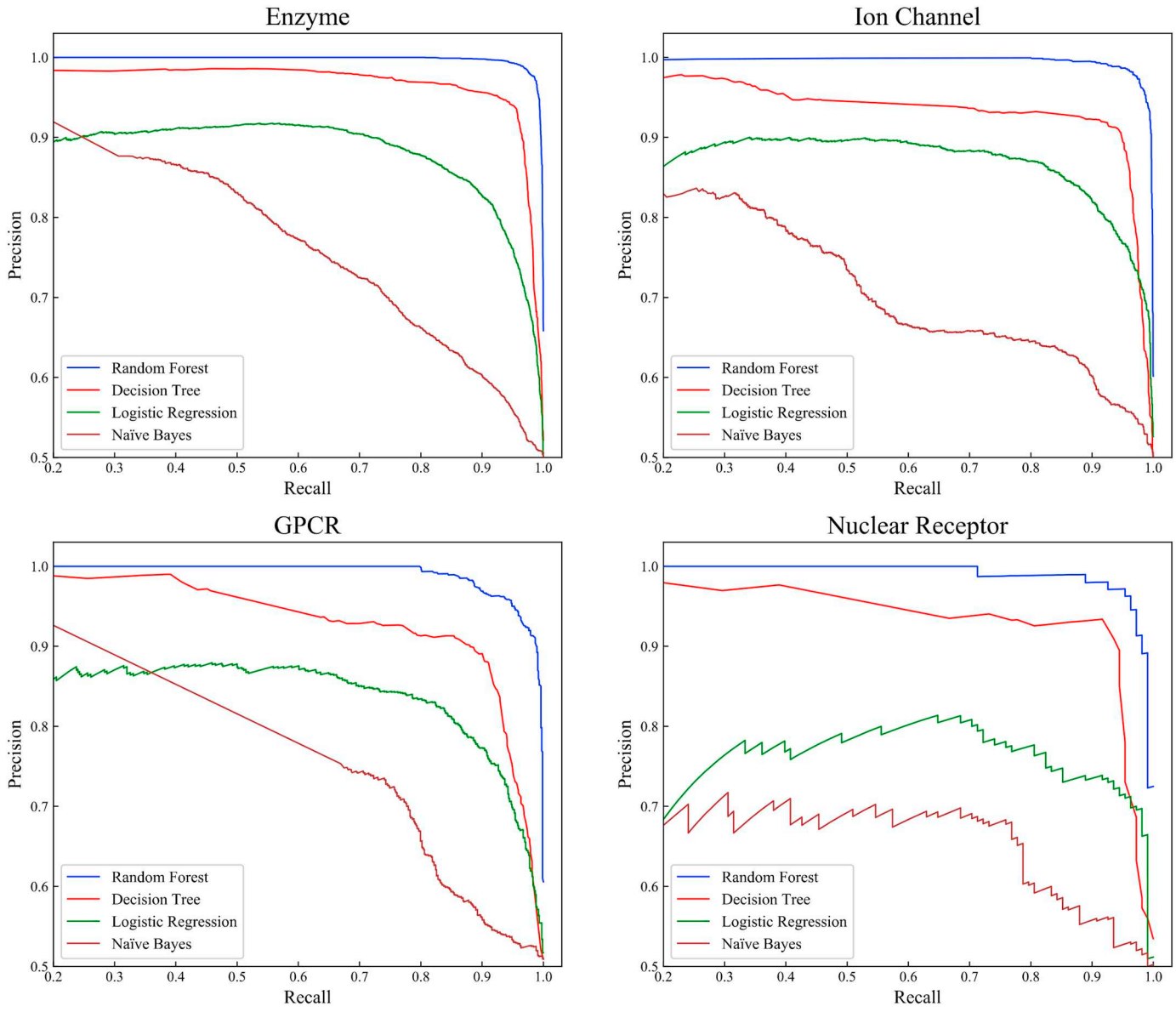

Fig. 3. One of the PR curves of four datasets under different classification algorithms in 10 trials of 5-fold cross-validation.

et al. [50] and Cheng et al. [51], the AUC has raised 3.78% and 4.58%. For the nuclear receptor dataset, our prediction method accuracy rate is 94.88%, which is 2.66%–11.14% higher than other prediction methods. It is 11.14% higher than the method proposed by Cao et al. [74] and 7.10% higher than the prediction method proposed by Meng et al. [41].

This paper combines the random forest classifier to classify the data, and the random forest can process very high-dimensional data, and the training speed is faster than other classifiers such as support vector machines. Random forests can balance errors when dealing with unbalanced data. Random forests provide an effective way to balance dataset errors when there is a classification imbalance. When the data is missing, the remaining classifiers do not work well, and random forests have a strong anti-interference ability. In addition, the ability of random forests to deal with overfitting is relatively strong, and most classifiers will have an overfitting phenomenon in the classification process. This random forest can be well avoided. Because of the above, the prediction accuracy of our model can be slightly better than the other three methods. In conclusion, the above results fully demonstrate that our prediction model can significantly improve the prediction accuracy of drug-target interactions, and the prediction results are satisfactory.

3.6. The discussion about false positive predictions

From a biological point of view, the most valuable aspect of drug-target prediction is to reveal possible new drug-target combinations. Aim at this goal, establishing a suitable prediction model can provide a valuable reference for experimental verification. Therefore, it may be beneficial to make the prediction method a bit too sensitive. In this paper, we use 2008 data to study some predictions that are considered to be false positives. Taking the benchmark nuclear receptor dataset as an example, the positive drug-target pairs are selected according to our model in one prediction results of 5-fold cross-validation, which with unknown interactions in 2008. Table 8 lists 23 pairs of dug-target pairs of false positive screened in once 5-fold cross-validation.

For the drug-target pairs screened in Table 8, the newly discovered drug-target pairs were found from the KEGG [53], DrugBank [56] database in the past decade. It can be found that among the 23 pairs of false positive drug-target pairs, 4 pairs have been verified as new interactions in the last decade, hsa2100 and D00067, hsa5915 and D00348, hsa2099 and D00951, hsa2908 and D00690, respectively. The newly discovered four drug-target pairs have important significance in reality. Among them, the drug-target pairs hsa2099 and D00951 can be

Table 7
Comparison of prediction results of different methods in gold standard datasets.

Datasets	Methods	AUC	SE (%)	SP (%)	ACC (%)
Enzymes	PDTPS[a]	N/A	97.44	98.02	97.73
	Li et al.[b]	0.9288	92.90	93.19	93.16
	Laarhoven et al.[c]	0.9780	88.00	**99.90**	N/A
	Cheng et al.[d]	0.9750	N/A	N/A	N/A
	SAR[e]	0.9486	90.10	90.64	90.31
	LRF-DTIs[f]	**0.9982**	**97.66**	98.51	**98.09**
IC	PDTPS[a]	N/A	93.32	92.95	93.12
	Li et al.[b]	0.9171	92.65	90.72	91.73
	Laarhoven et al.[c]	0.9840	29.00	**100.00**	N/A
	Cheng et al.[d]	0.9760	N/A	N/A	N/A
	SAR[e]	0.9428	89.38	88.20	88.91
	LRF-DTIs[f]	**0.9965**	**96.71**	97.93	**97.32**
GPCR	PDTPS[a]	N/A	84.89	88.35	86.77
	Li et al.[b]	0.8856	89.27	89.45	89.37
	Laarhoven et al.[c]	0.9540	33.10	**100.00**	N/A
	Cheng et al.[d]	0.9460	N/A	N/A	N/A
	SAR[e]	0.8902	82.54	85.49	84.68
	LRF-DTIs[f]	**0.9918**	**95.26**	96.11	**95.69**
NR	PDTPS[a]	N/A	92.63	84.44	87.78
	Li et al.[b]	0.9300	**96.62**	87.78	92.22
	Laarhoven et al.[c]	0.9220	15.60	**100.00**	N/A
	Cheng et al.[d]	0.8380	N/A	N/A	N/A
	SAR[e]	0.8822	82.35	84.72	83.74
	LRF-DTIs[f]	**0.9559**	93.85	95.81	**94.88**

N/A represents not available.
[a] is reported in [41], and uses 5-fold cross-validation to evaluate the performance of the model.
[b] is reported in [42], and uses 5-fold cross-validation to evaluate the performance of the model.
[c] is reported in [50], and uses leave-one-out cross-validation to evaluate the performance of the model.
[d] is reported in [51], and uses 30 trials of 10-fold cross validation to evaluate the performance of the model.
[e] is reported in [74], and uses 5-fold cross-validation to evaluate the performance of the model.
[f] is the method proposed in this paper, and uses 10 trials of 5-fold cross-validation to evaluate the performance of the model.

Table 8
Drug-target pairs of false positive screened in once 5-fold cross-validation.

Target	Drug	Target	Drug
hsa2099	D00506	hsa2908	D00690
hsa2099	D00956	hsa367	D00088
hsa2099	D00088	hsa5241	D00961
hsa2099	D00930	hsa5241	D00075
hsa2099	D01217	hsa5465	D00348
hsa2099	D00965	hsa5465	D00094
hsa2099	D00951	hsa5915	D00348
hsa2099	D00730	hsa6257	D00348
hsa2099	D00585	hsa7421	D00143
hsa2100	D00316	hsa7421	D01132
hsa2100	D00067	hsa9970	D01132
hsa2103	D01132		

used as a contraceptive and treat secondary amenorrhea, abnormal uterine bleeding, renal cell carcinomas and so on [75]. The drug-target pairs hsa2908 and D00690 can reduce inflammation and scar tissue formation by diminishing the release of leukocytic acid hydrolases, prevention of macrophage accumulation at inflamed sites, interference with leukocyte adhesion to the capillary wall, reduction of capillary membrane permeability, reduction of complement components, and inhibition of histamine and kinin release [76]. This shows that the prediction model established in this paper can provide more practicality for drug discovery.

3.7. Predictions on new drug-target interactions

To make the LRF-DTIs method more practical in reality, this paper applies the method to the effect of the prediction model built by the new interaction drug-target pairs. Taking the benchmark nuclear receptor dataset as an example, two nuclear receptor targets and 31 corresponding drugs are first collected from the currently published major databases (KEGG [53], DrugBank [56] and ChEMBL [77]). They produce the interaction of 33 new interactions. Then, the features corresponding to the new target and the new drug are extracted as described in 2.2 and 2.3. Finally, the samples generated by the new interaction are used as the testing set. All the samples in the nuclear receptor dataset are used as the training set. The drug-target pair features are input into the constructed model to test the prediction results. Fig. 4 shows the results of the new prediction interactions. The drug-target pairs in the figure have been verified to be interactive, and the solid line indicates that the prediction model constructed by this paper successfully predicts the interaction, and the dotted line indicates the drug that predicts the error.

As shown in Fig. 4, it can be seen that 29 pairs of 33 pairs new interactions in the nuclear receptor dataset are predicted successfully by experimental data using LRF-DTIs method. That means our prediction method could be applied to predict the drug-target interactions network, indicating that the proposed algorithm is more practical in reality.

According to material from KEGG [53] and DrugBank [56] database, it is found that among the above 33 drug-target pairs, some of the successful drug-target pairs play an important role in reality. For example, in hsa2099 and D04104, the target hsa2099 is the Estrogen receptor, and the drug D04104 is named Etonogestrel. Etonogestrel is used as a female contraceptive which is a progestin or a synthetic form of the naturally occurring female sex hormone, progesterone. Once bound to the receptor, progestins like etonogestrel will slow the frequency of release of gonadotropin releasing hormone (GnRH) from the hypothalamus and blunt the pre-ovulatory LH (luteinizing hormone) surge [78]. In hsa5241 and D03799, the target hsa5241 is named progesterone receptor, and the name D03799 is Dienogest. Dienogest acts as an agonist at the progesterone receptor (PR) with a weak affinity that is comparable to that of progesterone [79]. It promotes antiproliferative, immunologic and antiangiogenic effects on endometrial tissue. Once bound to the receptor, Dienogest reduces the level of endogenous production of oestradiol and thereby suppressing the trophic effects of oestradiol on both the eutopic and ectopic endometrium. It is an antagonist at androgen receptors, improving androgenic symptoms such as acne and hirsutism.

3.8. Validation of the proposed model using DrugBank_approved dataset

Although the model established in this paper has achieved satisfactory results in the gold standard dataset, it is also important to verify whether the model constructed in this paper produces better results on the new dataset. Therefore, this paper validates the model presented in this paper using a dataset called DrugBank_approved [57], which is a dataset constructed by Ba-Alawi et al. [57] in 2016. The data contains all FDA-approved drugs and their corresponding protein targets in the DrugBank dataset.

Similar to the gold standard dataset, this paper uses the PsePSSM and FP2 molecular fingerprints to extract the corresponding features from the target and drug in the DrugBank_approved dataset and then uses the Lasso and SMOTE algorithms to reduce the dimension and deal with the sample imbalance problem. Finally, using 5-fold cross-validation, input the processed data into the random forest classifier. To test the effect of the new dataset in the construction model, the results predicted by the LRF-DTIs method are compared with the evaluation indicators of the model in [44]. Table 9 shows the results of the performance using the LRF-DTIs in the DrugBank_approved dataset and the

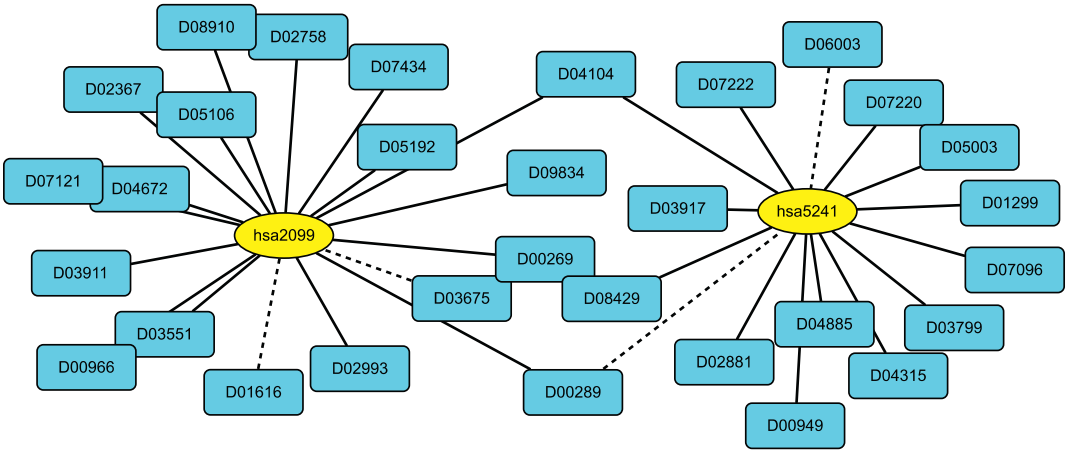

Fig. 4. Results of new predicted drug-target interactions.

Table 9
The result and comparison of validation DrugBank_approved dataset using LRF-DTIs.

Methods	AUC	SE (%)	SP (%)	ACC (%)
DASPfind	0.8884	N/A	N/A	N/A
LRF-DTIs	0.9978	97.48	99.20	98.34

Note: N/A represents not available.

comparison of the DASPfind method used in [57].

From Table 9, it can be seen that by comparing the parameter AUC, the method of this paper is 10.94% higher than the DASPfind method. The accuracy obtained by using the 5-fold cross-validation prediction in the DrugBank_approved dataset reaches 98.34%. This indicates that our proposed LRF-DTIs algorithm for predicting other drug-target interactions datasets are also valid.

4. Conclusion

The explosive growth of data has prompted the arrival of the era of big data, and a large number of drug molecular information and protein sequences are constantly flowing into the database. Traditional methods of biological experiments have been unable to meet the needs of life science research in today's society. Therefore, it is important to study a new method for drug-target interactions prediction based on machine learning. This paper proposes a new drug-target interactions prediction method. The four drug-target data sets are combined with the pseudo-position specific scoring matrix (PsePSSM) and medicinal chemical structure information to extract the features of the protein and drug sequences, and the redundant information in the dataset is processed based on the Lasso dimensionality reduction method. The SMOTE method is used to deal with the imbalance of positive and negative samples, and finally, the processed data is input into the random forest classification algorithm for drug-target interactions prediction. PsePSSM feature extraction can generate features that describe the original information of a protein, while the drug chemical structure information methods can extract FP2 molecular fingerprints to describe the molecular structure information. Lasso dimension reduction can effectively remove the redundant information in the drug-target dataset, making the characteristics of the drug-target pairs more obvious. The SMOTE method can effectively avoid overfitting of the model, improve the generalization performance of the model, and make full use of valid data. In addition, the selection of optimal parameters in PsePSSM can significantly improve the prediction accuracy rate. Random forest classification algorithms can handle multiple data types, have faster learning speeds, effectively handle noise data, and build highly accurate classifiers. The overall prediction accuracy of the drug-target interactions for the four datasets based on Lasso and the random forest model reach 98.09%, 97.32%, 95.69%, and 94.88%, respectively, and are compared with other prediction methods. Meanwhile, the proposed model has been tested to predict the new drug-target interaction and have been verified on a new dataset. The results show that the proposed method not only can effectively improve the prediction accuracy of drug-target interactions but also can make efficient predictions for novel drug-target interactions and new dataset. We expect that this method can provide powerful predictive tools in the study of drug-target interactions and new drug development.

However, the enzyme dataset in the drug-target interactions dataset is too large, the classifiers used in this study cannot rapidly predict drug-target interactions. Therefore, the improvement of the model's operating speed during prediction is of great significance to the promotion of the model. In the future work, we will make appropriate improvements to the classifier based on deep learning knowledge, to speed up the speed of the prediction model and improve the prediction accuracy of the model. In addition, the prediction and identification of drug-drug interactions are of paramount importance for patient safety and successful treatment [80,81]. Researching drug-drug interactions prediction based on machine learning is another work of us.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 61863010 and 11771188), the Natural Science Foundation of Shandong Province of China (No. ZR2018MC007), the Project of Shandong Province Higher Educational Science and Technology Program (No. J17KA159), and the College Students' Innovative Practice Training Program of Chinese Academy of Sciences. This work used the Extreme Science and Engineering Discovery Environment, which is supported by National Science Foundation (No. ACI-1548562).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2018.12.007.

References

[1] M.A. Yildirim, K.I. Goh, M.E. Cusick, A.L. Barabási, M. Vidal, Drug-target network, Nat. Biotechnol. 25 (10) (2007) 1119–1126.
[2] S.C. Janga, A. Tzakos, Structure and organization of drug-target networks: insights from genomic approaches for drug discovery, Mol. BioSyst. 5 (2009) 1536–1548.
[3] M. Kuhn, M. Campillos, P. Bork, I.P. Gonzã, L.J. Jensen, Large-scale prediction of drug-target relationships, FEBS Lett. 582 (8) (2008) 1283–1290.
[4] Y.C. Wang, Z.X. Yang, Y. Wang, N.Y. Deng, Computationally probing drug-protein interactions via support vector machine, Lett. Drug Des. Discov. 7 (2010) 370–378.
[5] Z. Xia, L.Y. Wu, X. Zhou, S.T.C. Wong, Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces, BMC Syst. Biol. 4 (2010) S6.
[6] M.R. Dreher, W. Liu, C.R. Michelich, M.W. Dewhirst, F. Yuan, A. Chilkoti, Tumor vascular permeability, accumulation, and penetration of macromolecular drug carriers, J. Natl. Cancer Inst. 98 (5) (2006) 335–344.
[7] Y. Tabei, E. Pauwels, V. Stoven, K. Takemoto, Y. Yamanishi, Identification of chemogenomic features from drug–target interaction networks using interpretable classifiers, Bioinformatics 28 (18) (2012) i487–i494.
[8] M. Hao, Y. Wang, S.H. Bryant, Improved prediction of drug-target interactions using regularized least squares integrating with kernel fusion technique, Anal. Chim. Acta 909 (2016) 41–50.
[9] M. Takarabe, M. Kotera, Y. Nishimura, S. Goto, Y. Yamanishi, Drug target prediction using adverse event report systems: a pharmacogenomic approach, Bioinformatics 28 (18) (2012) i611–i618.
[10] J.C. Dearden, In silico prediction of drug toxicity, J. Comput. Aided Mol. Des. 17 (2–4) (2003) 119–127.
[11] H. Van de Waterbeemd, E. Gifford, ADMET in silico modelling: towards prediction paradise? Nat. Rev. Drug Discov. 2 (3) (2003) 192–204.
[12] Y.F. Dai, X.M. Zhao, A survey on the computational approaches to identify drug targets in the postgenomic era, Biomed. Res. Int. 239654 (2015) 1–9.
[13] M.J. Keiser, B.L. Roth, B.N. Armbruster, P. Emsberger, J.J. Irwin, B.K. Shoichet, Relating protein pharmacology by ligand chemistry, Nat. Biotechnol. 25 (2) (2007) 197–206.
[14] B.K. Shoichet, Virtual screening of chemical libraries, Nature 432 (7019) (2014) 862–865.
[15] G. Klebe, Virtual ligand screening: strategies, perspectives and limitations, Drug Discov. Today 11 (13–14) (2006) 580–594.
[16] A.C. Cheng, R.G. Coleman, K.T. Smyth, Q. Cao, P. Soulard, D.R. Caffrey, A.C. Salzberg, E.S. Huang, Structure-based maximal affinity model predicts small-molecule druggability, Nat. Biotechnol. 25 (1) (2007) 71–75.
[17] X. Chen, C.C. Yan, X. Zhang, X. Zhang, F. Dai, J. Yin, Y. Zhang, Drug-target interaction prediction: databases, web servers and computational models, Brief. Bioinform. 17 (4) (2016) 696–712.
[18] H. Nakashima, K. Nishikawa, Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies, J. Mol. Biol. 238 (1) (1994) 54–61.
[19] M. Reczko, H. Bohr, The DEF data base of sequence based protein fold class predictions, Nucleic Acids Res. 22 (17) (1994) 3616–3619.
[20] Y.A. Huang, Z.H. You, X. Chen, A Systematic Prediction of drug-target interactions using molecular fingerprints and protein sequences, Current Protein Peptide Sci. 19 (5) (2018) 468–478.
[21] K.C. Chou, Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes, Bioinformatics 21 (1) (2005) 10–19.
[22] C.W. Yap, Y.Z. Chen, Prediction of cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines, J. Chem. Inf. Model. 45 (4) (2005) 982–992.
[23] F. Hammann, H. Gutmann, U. Baumann, C. Helma, J. Drewe, Classification of cytochrome P450 activities using machine learning methods, Mol. Pharm. 6 (6) (2009) 1920–1926.
[24] C.W. Yap, PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints, J. Comput. Chem. 32 (7) (2011) 1466–1474.
[25] H. Hotelling, Analysis of a complex of statistical variables into principal components, J. Educ. Psychol. 24 (6) (1933) 417–441.
[26] R.A. Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugenics 7 (2) (1936) 179–188.
[27] M. Vlachos, C. Domeniconi, D. Gunopuos, G. Kollios, N. Koudas, Non-linear dimensionality reduction techniques for classification and visualization, Proc. 8th ACM SIGKDD Internat. Conf. On Knowledge Discovery and Data Mining, Vol. 2 2002, pp. 645–651.
[28] Q. Kuang, X. Xu, R. Li, Y. Dong, Y. Li, Z. Huang, Y. Li, M. Li, An eigenvalue transformation technique for predicting drug-target interaction, Sci. Rep. 5 (2015) 13867.
[29] J. Hu, Y. Li, J.Y. Yang, H.B. Shen, D.J. Yu, GPCR–drug interactions prediction using random forest with drug-association-matrix-based post-processing procedure, Comput. Biol. Chem. 60 (2016) 59–71.
[30] L. Wang, Z.H. You, X. Chen, S.X. Xia, F. Liu, X. Yan, Y. Zhou, K.J. Song, A computational-based method for predicting drug–target interactions by using stacked autoencoder deep neural network, J. Comput. Biol. 25 (2017) 361–373.
[31] H. Yu, J. Chen, X. Xu, Y. Li, H. Zhao, Y. Fang, X. Li, W. Zhou, W. Wang, Y. Wang, A systematic prediction of multiple drug-target interactions from chemical, genomic, and pharmacological data, PLoS One 7 (5) (2012) e37608.
[32] Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda, M. Kanehisa, Prediction of drug-target interaction networks from the integration of chemical and genomic spaces, Bioinformatics 24 (2008) i232–i240.
[33] K. Bleakley, Y. Yamanishi, Supervised prediction of drug-target interactions using bipartite local models, Bioinformatics 25 (18) (2009) 2397–2403.
[34] K. Buza, L. Peška, Drug-target interaction prediction with bipartite local models and hubness-aware regression, Neurocomputing 260 (2017) 284–293.
[35] J.Y. Shi, S.M. Yiu, Y. Li, H.C.M. Leung, F.Y.L. Chin, Predicting drug–target interaction for new drugs using enhanced similarity measures and super-target clustering, Methods 83 (2015) 98–104.
[36] A. Ezzata, M. Wu, X. Li, C.K. Kwoh, Drug-target interaction prediction using ensemble learning and dimensionality reduction, Methods 129 (2017) 81–88.
[37] Q.F. Kuang, Y.Z. Li, Y.M. Wu, R. Li, Y.C. Dong, Q. Xiong, Z.V. Huang, M.L. Li, A kernel matrix dimension reduction method for predicting drug-target interaction, Chemom. Intell. Lab. 162 (2017) 104–110.
[38] Z. Mousavian, S. Khakabimamaghani, K. Kavousi, A. Masoudi-Nejad, Drug-target interaction prediction from PSSM based evolutionary information, J. Pharm. Toxicol. Methods 78 (2016) 42–51.
[39] Q. Yuan, J. Gao, D. Wu, S. Zhang, H. Mamitsuka, S. Zhu, DrugE-Rank: improving drug–target interaction prediction of new candidate drugs or targets by ensemble learning to rank, Bioinformatics 32 (12) (2016) i18–i27.
[40] T. He, M. Heidemeyer, F. Ban, A. Cherkasov, M. Ester, SimBoost: a read-across approach for predicting drug-target binding affinities using gradient boosting machines, J. Cheminform. 9 (1) (2017) 24.
[41] F.R. Meng, Z.H. You, X. Chen, Y. Zhou, J.Y. An, Prediction of drug-target interaction networks from the integration of protein sequences and drug chemical structures, Molecules 22 (2017) 1119.
[42] Z.W. Li, P.Y. Han, Z.H. You, X. Li, Y.S. Zhang, H.Q. Yu, R. Nie, X. Chen, In silico prediction of drug-target interaction networks based on drug chemical structure and protein sequences, Sci. Rep. 7 (1) (2017) 11174.
[43] A. Ezzat, P. Zhao, M. Wu, X.L. Li, C.K. Kwoh, Drug-target interaction prediction with graph regularized matrix factorization, IEEE ACM Trans. Comput. Bioinform. PP (99) (2017) 646–656.
[44] Y. Liu, M. Wu, C. Miao, P. Zhao, X.L. Li, Neighborhood regularized logistic matrix factorization for drug-target interaction prediction, PLoS Comput. Biol. 12 (2016) e1004760.
[45] J. Zhang, M.C. Zhu, P. Chen, B. Wang, DrugRPE: random projection ensemble approach to drug-target interaction prediction, Neurocomputing 228 (2017) 256–262.
[46] Y. Luo, X. Zhao, J. Zhou, J. Yang, Y. Zhang, W. Kuang, J. Peng, L. Chen, J. Zeng, A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information, Nat. Commun. 8 (1) (2017) 573.
[47] R.S. Olayan, H. Ashoor, V.B. Bajic, DDR: efficient computational method to predict drug-target interactions using graph mining and machine learning approaches, Bioinformatics 34 (7) (2018) 1164–1173.
[48] T. Pahikkala, A. Airola, S. Pietila, S. Shakyawar, A. Szwajda, J. Tang, T. Aittokallio, Toward more realistic drug-target interaction predictions, Brief. Bioinform. 16 (2) (2014) 325–337.
[49] Z. He, J. Zhang, X.H. Shi, L.L. Hu, X. Kong, Y.D. Cai, K.C. Chou, Predicting drug-target interaction networks based on functional groups and biological features, PLoS One 5 (3) (2010) e9603.
[50] T.V. Laarhoven, S.B. Nabuurs, E. Marchiori, Gaussian interaction profile kernels for predicting drug–target interaction, Bioinformatics 27 (21) (2011) 3036–3043.
[51] F. Cheng, C. Liu, J. Jiang, W. Lu, W. Li, G. Liu, W. Zhou, J. Huang, Y. Tang, Prediction of drug-target interactions and drug repositioning via network-based inference, PLoS Comput. Biol. 8 (5) (2012) e1002503.
[52] M. Gönen, Predicting drug–target interactions from chemical and genomic kernels using Bayesian matrix factorization, Bioinformatics 28 (18) (2012) 2304–2310.
[53] M. Kanehisa, S. Goto, M. Hattori, K.F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, M. Hirakawa, From genomics to chemical genomics: new developments in KEGG, Nucleic Acids Res. 34 (2006) D354–D357.
[54] I. Schomburg, A. Chang, C. Ebeling, M. Gremse, C. Heldt, G. Huhn, D. Schomburg, BRENDA, the enzyme database: updates and major new developments, Nucleic Acids Res. 32 (2004) D431–D433.
[55] S. Günther, M. Kuhn, M. Dunkel, M. Campillos, C. Senger, E. Petsalaki, J. Ahmed, E.G. Urdiales, A. Gewiess, L.J. Jensen, R. Schneider, R. Skoblo, R.B. Russell, P.E. Bourne, P. Bork, R. Preissner, SuperTarget and Matador: resources for exploring drug-target relationships, Nucleic Acids Res. 36 (2007) D919–D922.
[56] D.S. Wishart, C. Knox, A.C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, M. Hassanali, DrugBank: a knowledgebase for drugs, drug actions and drug targets, Nucleic Acids Res. 36 (2008) D901–D906.
[57] W. Ba-Alawi, O. Soufan, M. Essack, P. Kalnis, V.B. Bajic, DASPfind: new efficient method to predict drug–target interactions, J. Cheminform. 8 (1) (2016) 15.
[58] D.T. Jones, Protein secondary structure prediction based on position-specific scoring matrices, J. Mol. Biol. 292 (1999) 195–202.
[59] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.
[60] K.C. Chou, Prediction of protein cellular attributes using pseudo amino acid composition, Proteins 43 (2001) 246–255.
[61] Y. Ding, J. Tang, F. Guo, Identification of drug-target interactions via multiple information integration, Inf. Sci. 418 (2017) 546–560.
[62] Y. Yamanishi, E. Pauwels, H. Saigo, V. Stoven, Extracting sets of chemical substructures and protein domains governing drug-target interactions, J. Chem. Inform. Model. 51 (2011) 1183–1194.
[63] Y. Tabei, Y. Yamanishi, Scalable prediction of compound-protein interactions using minwise hashing, BMC Syst. Biol. 7 (2013) S3.
[64] N.M. O'Boyle, M. Banck, C.A. James, C. Morley, T. Vandermeersch, G.R. Hutchison, Open Babel: an open chemical toolbox, J. Cheminform. 3 (2011) 33.
[65] R. Tibshirani, Regression shrinkage and selection via the LASSO: a retrospective, J.

R. Statist. Soc. B 73 (2011) 273–282.
[66] D. Ghosh, A.M. Chinnaiyan, Classification and selection and biomarkers in genomic data using LASSO, J Biomed Biotechnol 2 (2005) 147–154.
[67] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.
[68] L. Breiman, Random forests, Mach. learning 45 (2001) 5–32.
[69] G. Nimrod, A. Szilágyi, C. Leslie, N. Ben-Tal, Identification of DNA-Binding proteins using structural, Electrostatic and Evolutionary Features, J. Mol. Biol. 387 (4) (2009) 1040–1053.
[70] A.G. Heidema, J.M. Boer, N. Nagelkerke, E.C. Mariman, E.J. Feskens, The challenge for genetic epidemiologists: how to analyze large numbers of SNPs in relation to complex diseases, BMC Genet. 7 (23) (2006) 1–15.
[71] P. Han, X. Zhang, R.S. Norton, Z.P. Feng, Large-scale prediction of long disordered regions in proteins using random forests, BMC Bioinf. 10 (2009) 8.
[72] J.D. Watts, R.L. Lawrence, P.R. Miller, C. Montagne, Monitoring of cropland practices for carbon sequestration purposes in north Central Montana using landsat imagery, Remote Sens. Environ. 113 (2009) 1843–1852.
[73] S.V. Stehman, Selecting and interpreting measures of thematic classification accuracy, Remote Sens. Environ. 62 (1997) 77–89.
[74] D.S. Cao, S. Liu, Q.S. Xu, H.M. Lu, J.H. Huang, Q.N. Hu, Y.Z. Liang, Large-scale prediction of drug-target interactions using protein sequences and drug topological structures, Anal. Chim. Acta 752 (2012) 1–10.
[75] A.S. Kumar, E. Cureton, V. Shim, T. Sakata, D.H. Moore, C.C. Benz, L.J. Esserman, E.S. Hwang, Type and duration of exogenous hormone use affects breast cancer histology, Ann. Surg. Oncol. 14 (2) (2007) 695–703.
[76] X. Chen, Z.L. Ji, Y.Z. Chen, TTD: Therapeutic Target Database, Nucleic Acids Res. 30 (1) (2002) 412–415.
[77] A.P. Bento, A. Gaulton, A. Hersey, L.J. Bellis, J. Chambers, M. Davies, F.A. Krüger, Y. Light, L. Mak, S. McGlinchey, M. Nowotka, G. Papadatos, R. Santos, J.P. Overington, The ChEMBL bioactivity database: an update, Nucleic Acids Res. 42 (2014) 1083–1090.
[78] P. Imming, C. Sinning, A. Meyer, Drugs, their targets and the nature and number of drug targets, Nat. Rev. Drug Discov. 5 (10) (2006) 821–834.
[79] M. Bińkowska, J. Woroń, Progestogens in menopausal hormone therapy, Menopausal Review 14 (2) (2015) 134–143.
[80] R. Ferdousi, R. Safdari, Y. Omidi, Computational prediction of drug-drug interactions based on drugs functional similarities, J. Biomed. Inform. 70 (2017) 54–64.
[81] S. Ayvaz, J. Horn, O. Hassanzadeh, Q. Zhu, J. Stan, N.P. Tatonetti, S. Vilar, M. Brochhausen, M. Samwald, M. Rastegar-Mojarad, M. Dumontier, R.D. Boyce, Toward a complete dataset of drug-drug interaction information from publicly available sources, J. Biomed. Inform. 55 (2015) 206–217.