DTI-CDF: a CDF model towards the prediction of DTIs based on hybrid features

Yan-Yi Chu [1], Yu-Fang Zhang[1], Wei Wang[1,2], Xian-Geng Wang[1], Xiao-Qi Shan[1], Yi Xiong[1,*], and Dong-Qing Wei[1,*]

[1]*State Key Laboratory of Microbial Metabolism, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China*

[2]*School of Mathematical Sciences, Shanghai Jiao Tong University, Shanghai, China*

**Address correspondence to the authors at Room 4-321, Life Science Building, Shanghai Jiaotong University, 800 Dongchuan Road, 200240, Minhang District, Shanghai, China; Phone\Fax: +86 21-34204573; Email: xiongyi@sjtu.edu.cn, dqwei@sjtu.edu.cn*

Abstract

Drug-target interactions play a crucial role in target-based drug discovery and exploitation. Computational prediction of DTIs has become a popular alternative strategy to the experimental methods for identification of DTIs of which are both time and resource consuming. However, the performances of the current DTIs prediction approaches suffer from a problem of low precision and high false positive rate. In this study, we aimed to develop a novel DTIs prediction method, named DTI-CDF, for improving the prediction precision based on a cascade deep forest model which integrates hybrid features, including multiple similarity-based features extracted from the heterogeneous graph, fingerprints of drugs, and evolution information of target protein sequences. In the experiments, we built five replicates of 10 fold cross-validations under three different experimental settings of data sets, namely, corresponding DTIs values of certain drugs (S_D), targets (S_T), or drug-target pairs (S_P) in the training set are missed, but existed in the test set. The experimental results show that our proposed approach DTI-CDF achieved significantly higher performance than the state-of-the-art methods.

Keywords: Drug-target interaction prediction, cascade deep forest, hybrid features

1 Introduction

Drug discovery is the process of identifying new candidate compounds with potential therapeutic effects, during which the prediction of drug-target interactions (DTIs) is an essential step [1, 2]. Drugs are significant in the human body by interacting with various targets. Proteins represent an important type of target, and their functions can be enhanced or inhibited by drugs to achieve phenotypic effects for therapeutic purposes [3, 4]. However, the number of drug candidates approved by the FDA is relatively small [4, 5], mainly due to the possible adverse effects of multi-targeting of drugs. Currently, large quantities of researches have focused on DTIs prediction because it is an essential tool in the context of drug repurposing. Since the high cost of experimental determination of DTIs, it is thus necessary to develop efficient computational methods by making use of the heterogeneous biological data on known DTIs to understand the mechanisms of action of drugs in the human body.

The recent computational approaches are mainly concerned with the ligand-based approaches [6], docking simulation approaches [7], and chemogenomic approaches [8]. The ligand-based interactions are motivated by the idea that similar molecules generally bind to similar proteins. However, these methods perform poorly especially when the number of known ligands is insufficient. The docking simulation methods need three-dimensional structure information of proteins for simulation, thus being intractable when numerous proteins are not access to their structure information. The chemogenomic methods have attracted much interest because it maps both the chemical feature space and the genomic feature space into a unified Euclidean space, namely pharmacological space and obtains good performance in DTIs prediction. These methods can mainly be divided into three classes: graph-based method, network-based method, and machine learning-based methods [9]. The former two methods sometimes have encountered the dilemma for the scarcity of known DTIs and unidentified negative DTIs data, thus the results are not highly confidential. However, machine learning-based methods can better overcome this dilemma for its advanced capability to sufficiently exploiting the sample information, resulting in more reliable prediction results than the former two methods. On the other hand, there are large quantities of high efficient machine learning methods have been employed in various research fields and obtain good performances, which motivates us to make full use of these methods such that enhances the prediction accuracy. Moreover, the machine learning models are being studied intensively which thus provides broaden possibilities for the improvement of DTIs prediction.

The recent machine learning-based methods are composed of semi-supervised models and supervised models. As for semi-supervised machine learning method, Xia et al. [10] first designed a manifold Laplacian regularized the least square (LapRLS) by using the concept of the bipartite local model (BLM) [11] with the assist of labeled and unlabeled data. Subsequently, some improved LapRLS-based models were used for DTIs prediction, such as NetLapRLS [10] and ILRLS [12]. In addition, there are some interesting work, such as the Network-Consistency-based Prediction method [13] carried out by maximizing the known interaction's rank coherence, the PUDT method [14] predicted DTIs by using positive unlabeled data, a set of Graph Auto-Encoder-based models applying multi-view similarities [15], and the NormMulInf method [16] developed a principal component analysis model based collaborative filtering method by using low-rank similarity matrix. These methods are easy to be implemented. However, these methods are unable to apply to drugs without any targets information and more time-consuming as the result of the unlabeled negative interactions and thus cannot be implemented on a large-scale database.

In the context of supervised learning of which known interactions are labeled positive samples and other interactions are labeled negative ones, there are two categories methods: similarity-based methods and feature vector-based methods. As for the similarity-based methods[17], the key assumption is the "guilt-by-association", i.e., similar drugs tend to share similar targets and vice versa. Based on the nearest neighbor method, a variety of similarity measures such as Jaccard similarity, Cosine similarity, and Pearson correlation similarity are exploited to calculate the similarity score such that enhances the model prediction ability. Shi et at. [18]designed a Similarity-Rank-based predictor to present the likelihood of each drug-target pair tending to interact or not.

The merit of this method does not need the complex parameter optimization. Chen et al. [19] employed an inference model based on the random walk on the heterogeneous network. Cheng et al. [20] developed three inferring networks consisting of drug-based similarity inference network, target-based similarity inference network and network-based inference to discover DTIs. The method cannot predict new drug or targets. Based on the BLM, the DTIs prediction problem can be transformed into a binary classification problem. Thus, various classifiers have been exploited such as regularized least square classifier [21], support vector machine (SVM) classifier [22]. Mei et al. [23] further exploited BLM with Neighbor-based Interaction Profile Inferring, which adds a preprocessing component to infer training data from neighbors' interaction profiles. In addition, in the matrix factorization methods that are typically utilized in recommendation systems to find a potential user-item relationship, DTIs can be transformed into a matrix completion problem aiming to find missing interactions. Gönen et al. [24] presented a pioneer work of Kernelized Bayesian Matrix Factorization with Twin Kernels. There are other methods such as Probabilistic Matrix Factorization [25], multiple Similarities Collaborative Matrix Factorization [26], Graph regularized matrix factorization [27], weighted GRMF [27] and Variational Bayesian Multiple Kernel Logistic Matrix Factorization [28]. These methods improve the interpretability to some extent by using the knowledge of matrix theory, however, are not practical in large-scale samples for the high cost of the matrix completion problem. Based on the feature vector, a variety of methods have been proposed. Wang et al. [29] employed a restricted Boltzmann machine to find the data distribution such that identifies DTIs relationship as well as drug modes of action. Based on meta-path-based topological features, Fu et al. [30] employed random forest classifier to do prediction, and Zong et al. [31] calculated it with SkipGram model. These methods are difficult to find new targets or drugs in known networks. Olayan et al. [32] exploited a novel method called DDR to solve the above problem and improve the prediction accuracy which executes graph mining technique firstly to acquire the comprehensive feature vectors and then applies the random forest model by using different graph-based features extracted from the drug-target heterogeneous graph. Farshid Rayhan et al. [33] explored sampling method to avoid data imbalanced problem and used Adaboost model to do prediction, and it is the first time combine evolutionary and structural information of proteins as a part of the feature vector. Ezzat et al. [34] carried out the ensembling learning method which using Decision Tree and Kernel Ridge Regression as base classifiers. Besides, Wen et al. [35] proposed a deep learning framework by deep belief network for the first time applying in this field, which needs further exploration. Based deep learning method, Hu et al. [36] used Auto-Encoders to learn representations as SVM's feature vector, and Tian et al. [37] used a deep neural network to learn and predict.

Motivated by the previous studies [38, 39], we propose a cascade deep forest (CDF) model that further improves the performance of predicting DTIs. This method combines the above two machine learning-based methods. First, we utilize FP2 fingerprint (FP2), to extract the structural information of drugs, Pseudo-position specific scoring matrix (PsePSSM) to extract evolution information of protein sequence, and adds Path-category-based multi-similarities feature (PathCS) based on the heterogeneous graph of DTIs. Then, we apply the CDF model under three experimental settings through five repeated 10-fold cross-validations in four representative data sets, and the performance evaluation is performed using both AUPR and AUC metrics. Besides, the statistical hypothesis test is used to evaluate the results' significance. Finally, we verify that the proposed DTI-CDF method is significantly better than the current state-of-art methods available.

2 Materials and Methods

2.1 Data sets

This study uses four data sets compiled by Yamanishi et al. [40] to evaluate the performance of the proposed DTI-CDF method in DTIs prediction. The four data sets are distinguished and named by the target protein of the drug: enzymes (E), ion channels (IC), G-protein-couples receptors (GPCR) and nuclear receptors (NR). These data sets contain known human DTIs retrieved from the KEGG BRITE [41], BRENDA [42], SuperTarget[43] and DrugBank [44] databases. Therefore, it is generally considered as the gold-standard data sets.

In order to simulate more practically, in these four data sets, we consider the entire space of the DTIs, where the number of known DTIs (or positive samples) is much lower than the number of unknown DTIs (or negative samples). Thus, these four data sets are very unbalanced, as shown in Table 1.

Table 1. Summary of quantitative information for the four data sets

Data sets	Known interactions	Unknown interactions	Drugs	Targets
NR	90 (6.41%)	1314 (93.59%)	54	26
GPCR	635 (3%)	20550 (97%)	223	95
IC	1476 (3.45%)	41364 (96.55%)	210	204
E	2926 (1%)	292554 (99%)	445	664

2.2 Feature construction

2.2.1 PathCS

PathCS [32] is based on the heterogeneous DTIs weighted graph, containing drugs, targets, and their similarities or interactions. In this graph, the edge between two target nodes or two drug nodes represents their similarities and the weight w_x is their similarity value. The edge between a target and a drug denotes a known drug-target interaction and the weight is equal to 1.

There are six kernels used in this study to generate similarity profiles for drugs and targets which have been proved a more robust and less redundant similarity set [32], their information as follows:

(1) Protein kernels. We use the proteins' amino acid sequence information to generate the spectrum kernel [45] and set the subsequence length k as 4.

(2) Drug kernels. There are three side-effects kernels as drug information sources. The first resource obtained from the SIDER database [46], which contains information on marketed drugs and their adverse reactions. For each side-effect classification, a binary (absence or presence) profile was used to represent drugs. The other two pharmacological profiles are derived from the FDA's adverse event reporting system [47] based on the frequency and binary information, respectively, of side-effects classifications. These three pharmacological profiles are used to generate similarity profiles through the weighted cosine correlation coefficient. And if a drug is not in the data resources, its assigned similarity is 0.

(3) Gaussian Interaction Profile kernel (GIP kernel). The GIP kernel [21] is a binary matrix based on the DTIs network for drugs or targets, in which the absence or presence of interaction in the

network for each drug or each target is described as 0 or 1, respectively. However, this kernel cannot be computed for a new drug (or a new target), which do not have any drugs (or targets) to interact with the training data set. To solve this problem, we adopt the method of neighbor-based interaction-profile inferring [23] to calculate this kernel.

After obtaining the above similarity measures, the first step is to combine the drugs' (or targets') multiple similarity measures into one fused matrix [48] to build a heterogeneous DTIs graph, then extract PathCS for each drug-target pair. The path category is defined by a path structure that starts at a drug node and ends up at a target node such as to set the path length to 2 or 3. Path categories are as follows: drug-drug-target, drug-target-target, drug-drug-drug-target, drug-drug-target-target, drug-target-drug-target and drug-target-target-target. We define two normalized matrices N_1^h and N_2^h according to the above 6 path categories C^h, $h = 1,2,\cdots,6$. For a specific drug d_i and a specific target t_j, we denote one path from d_i to t_j as p_q and the set of paths is R_{ijh}. In addition, the path between d_i and t_j is built by the intermediate nodes which are restricted to be the 5 nearest neighbors of d_i and t_j, respectively. Thus, the N_1^h and N_2^h with elements $n_1^h(\mathrm{i},\mathrm{j})$ and $n_2^h(\mathrm{i},\mathrm{j})$, respectively, are computed as follows:

$$n_1^h(i,j) = \frac{\sum_{\forall q: p_q \in R_{ijh}} \prod_{\forall w_x \in p_q, p_q \in R_{ijh}} w_x}{\sum_j \sum_{\forall q: p_q \in R_{ijh}} \prod_{\forall w_x \in p_q, p_q \in R_{ijh}} w_x} \tag{1}$$

$$n_2^h(i,j) = \frac{\max_{\forall q: p_q \in R_{ijh}} \prod_{\forall w_x \in p_q, p_q \in R_{ijh}} w_x}{\sum_j \max_{\forall q: p_q \in R_{ijh}} \prod_{\forall w_x \in p_q, p_q \in R_{ijh}} w_x} \tag{2}$$

2.2.2 PsePSSM

There are varieties of methods that can be used to extract characteristic information on proteins based on the amino acid sequence. In this study, we select PsePSSM [49] as our protein description method. This method combines the position-specific scoring matrix (PSSM) [50] and the pseudo amino acid composition (PseACC) [51], which represent evolution information and sequence information, respectively. For an amino acid sequence A with L residues, the dimension of a normalized PSSM is $\mathrm{L} \times 20$ as follows:

$$A_{PSSM} = \begin{bmatrix} E_{1,1} & E_{1,2} & \cdots & E_{1,j} & \cdots & E_{1,20} \\ E_{2,1} & E_{2,2} & \cdots & E_{2,j} & \cdots & E_{2,20} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ E_{i,1} & E_{i,2} & \cdots & E_{i,j} & \cdots & E_{i,20} \\ \vdots & \vdots & \cdots & \vdots & \ddots & \vdots \\ E_{L,1} & E_{L,2} & \cdots & E_{L,j} & \cdots & E_{L,20} \end{bmatrix} \tag{3}$$

where i is the position of the residue in the amino acid sequence, j is the type of one of the 20 native amino acids, $E_{i,j}$ is the score at which the i-th residue in the amino acid sequence is mutated to the j-th amino acid.

For proteins with different amino-acid sequence length, the number of rows of this matrix is different. To solve this problem, one can turn this matrix into a vector of length 20 as

$$\bar{A}_{PSSM} = [E_1, E_2, \cdots, E_{20}], \quad E_j = \frac{1}{L}\sum_{i=1}^{L} E_{i,j} \tag{4}$$

where E_j is the average occurrence score of all residues in the amino-acid sequence that is mutated to the j-th amino acid during evolution. The PseACC is used to describe an amino-acid sequence A using Eq. (3) as

$$G_j^{\lambda} = \frac{1}{L-\lambda}\sum_{i=1}^{L-\lambda}\left|E_{i,j} - E_{i+\lambda,j}\right|^2,\ j=1,2,\cdots,20;\ 0\le\lambda\le L \tag{5}$$

where G_j^{λ} is the λ-tier correlation factor of the jth amino acid and λ denotes the difference order along each column of the matrix A_{PSSM}. For clearly, G_j^1 represents the relevant factor calculated by the 1-order difference of row elements along the jth column of the matrix, i.e., the closest PSSM score, on the protein sequence of amino acid type j, and G_j^2 represents the 2-order difference along the j-th column of the matrix, namely the second closest PSSM score and so on. In this study, $\lambda \in \{1,2, \dots 10\}$ because of its best performance [49], thus generating a vector containing 200 components. Hence, a protein using the PsePSSM method obtains a feature vector with 220 elements, namely $[\bar{A}_{PSSM}, G_1^1, G_2^1, \dots G_1^2, G_2^2, \dots G_{19}^{10}, G_{20}^{10}]$.

2.2.3 FP2

FP2 [52] is a path-based fingerprint to express drugs. This fingerprint identifies all ring and linear substructures of lengths from 1 to 7 in the molecule, among them, the C and N substructures should be excluded. Then, these substructures are mapped to a 1024-bit string. Thus, for each molecule, FP2 is a hexadecimal 256-dimensional vector.

2.3 Classification algorithm

Firstly, we generate the feature vector for each DTI. The GIP similarity is constructed according to training data, then PathCS is obtained. Based on the drug-target pair, the PathCS, FP2, and PsePSSM are merged to form input feature vector, called hybrid features.

Secondly, CDF classifier [53] is used to predict DTIs. The input of the CDF model is hybrid features. Then, the new category probability vector link input feature vector is used as the next layer input, and the final category probability vector is output through multiple learners. When building a CDF model (Fig.1), it is important to determine the machine learner used for each layer. In the model, we set the number of learners of each layer from 2 to 6, and Random Forest (RF) [54] and XGBoost (XGB) [55] are used as learners to follow the "good but distinguishable" principle. In addition, the depth of layers is identified automatic by the trend of evaluation metrics.

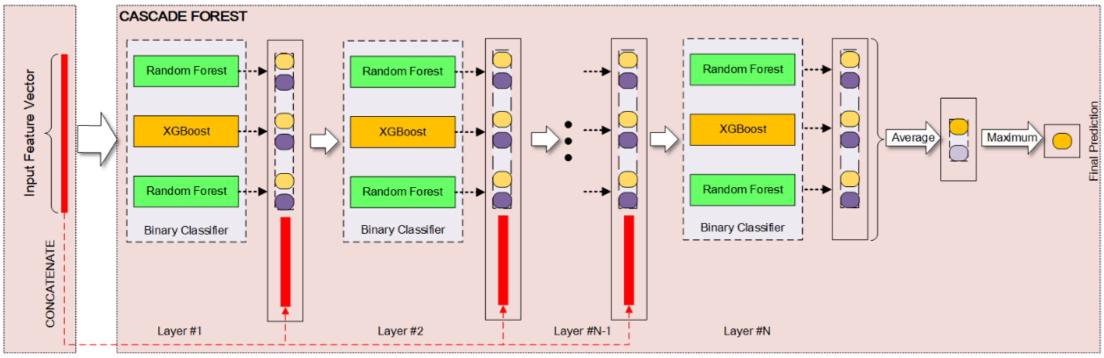

Fig. 1. This machine learning model is composed of an input feature vector, a CDF classifier, and a final prediction. In particular, CDF is the core unit of the model, which has five variants in this study.

In each variant, each layer consists of a different number of RF and XGB binary classifiers and different layers own the same structure. The figure shows one special model in which each layer has two RF learners and one XGB learner, denoted as RF2-XGB1. Other variants are RF2, RF3, RF3-XGB2, RF4-XGB2, respectively.

2.4 Experimental settings and cross-validation

In this study, we evaluate three experimental settings as Table 2 shows, which includes most of the conditions for DTIs prediction. In Table 2, objects are new indicates that no corresponding DTIs in the training data, and known vice versa. In order to facilitate the comparison with other methods, we followed previous studies [32, 56-58] as the benchmark and conducted the 10-fold cross-validations (CVs) test for each experimental setting of each data set, and the above process was repeated 5 times using different random seeds.

Table 2. Summary the corresponding DTIs information in test data of three experimental settings

Experimental settings	Drugs	Targets	Interactions
S_P	Known	Known	New
S_D	New	Known	New
S_T	Known	New	New

2.5 Performance evaluation

For each fold of each predictive model, the following metrics are calculated:

$$\text{Precision} = \frac{TP}{TP+FP} \tag{6}$$

$$\text{True Positive Rate} = \text{Recall} = \frac{TP}{TP+FN} \tag{7}$$

$$\text{False Positive Rate} = \frac{FP}{FP+TN} \tag{8}$$

where *TP* is true positive, *FP* is false positive, *FN* is false negative and *TN* is true negative. We plot the precision-recall curve (PR curve) based on different precision and recall, and the receiver operating characteristic curve (ROC curve) based on different recall and false positive rate, respectively, under the condition of different classified cutoff values. We define AUPR and AUC as the area under the PR curve and the ROC curve, respectively. For each experiment setting of each data set, the AUPR and AUC are calculated as a measure of model performance as follows:

$$AUPR = \frac{\sum_{i=1}^{5}\sum_{j=1}^{10} AUPR_{i,j}}{50} \tag{9}$$

$$AUC = \frac{\sum_{i=1}^{5}\sum_{j=1}^{10} AUC_{i,j}}{50} \tag{10}$$

where i represents the i-th repeated trials, j represents the j-th fold of CVs. Since the positive samples and negative samples in each data set are highly unbalanced, the AUPR provides a better performance estimate relative to AUC because it more severely penalizes the false positives[59].

3 Results and Discussion

3.1 Predictive ability of different types of features

The aim of this section is to evaluate the effect of adding FP2 and PsePSSM into a similarity-based feature vector. For computational method in DTI prediction, extracting the features of drugs and targets effectively is very important. In previous studies, there are two main methods for generating features of drugs and targets: (i) based on the chemical structure for drugs and the amino acid sequence for proteins to extract features. For chemical structures, various molecular fingerprints of compounds can be used, such as FP2 [49], Extended-Connectivity Fingerprints [60], etc. The amino acid sequences can be represented by amino acid composition [61], PseAAC [62], PsePSSM [49], etc. (ii) based on association rules, drugs with similar chemical structures tend to bind similar proteins. It is based on heterogeneous networks of DTIs, using single or fusion similarity measures as features [23, 31, 32, 58].

The similarity-based feature plays a crucial role in predicting DTIs. Large quantities of studies use evolution information for targets' sequence and structure information for drugs, PsePSSM and FP2 are reported that they can effectively extract information on drugs and targets, respectively, as good results are obtained by using the above two features [32, 49]. In this study, we combined the above two feature extraction method to generate the input feature vector. Based on PathCS, we constructed hybrid features by adding FP2 and PsePSSM to describe drugs and targets, respectively. The results show that this hybrid features achieves higher AUPR and AUC than using PathCS along (Fig. 2). Furthermore, it demonstrates that in the prediction of DTIs, combining the structure information of drugs and evolution information of targets' sequence with similarity information could increase accuracy than similarity information used alone.

In addition, we use a statistical hypothesis test [63] to further explore the extent to which hybrid features outperform single PathCS used in the DTI-CDF method. Differences in the results of different prediction methods are caused by a variety of factors, such as data composition, training model and experimental setting, etc. In order to exclude other factors and only consider the difference caused by the prediction method, the one-sided paired t-test, that is a pairwise comparison method based on paired data, is employed. Firstly, the difference $d_i \in \mathbf{D}, (\mathrm{i} = 1,2,\cdots,24)$ of AUPR and AUC based on 12 experimental conditions (i.e. four data sets under three experimental settings) between the above two methods are calculated. It is assumed that the difference d_i are all from the normal distribution $N(\mu_d, \sigma^2)$, where both μ_d and σ^2 are unknown. Then, a statistical hypothesis test is performed based on the data obtained above. If the hybrid features used in DTI-CDF method is not different from the single PathCS, the difference d_i between each pair of data belongs to a random error, and the random error can be considered to obey a normal distribution with a mean of zero. Assuming that there is no difference between the above two methods, the test hypothesis is as follows:

$$H_0 : \mu_d = 0, H_1 : \mu_d > 0 + \Delta \tag{11}$$

By the *t*-test of a single population means using the normal distribution, the rejection domain is:

$$t = \frac{\bar{D}}{s/\sqrt{n}} \geq t_\alpha\left(n-1\right) \tag{12}$$

where $\bar{D}$ is the mean of the sample, s is the standard deviation of the sample, n is the sample size, α is the significance level and Δ is equivalent to the effect size of mean difference, defined as $\Delta = \frac{\mu_d - 0}{\sigma}$. In order to ensure that only when the hybrid features are far superior to the single PathCS used in DTI-CDF method, it can be tested with a high probability $1-\beta$, we set $\alpha = 0.05$, $\Delta = 0.9$, $\beta = 0.01$. Under these conditions, a sample size n not less than 21 is required, and the actual sample size n is 24 satisfying the requirement. The rejection domain and the actual effect size of mean difference are

$$t = \frac{\bar{D}}{s/\sqrt{n}} \geq t_{0.05}(23) = 1.71 \tag{13}$$

$$\delta = \frac{\bar{D}}{s} \tag{14}$$

Substituting d_i into the above formula yields the observed value t_0 of t, then the p-value of the right-tailed t-test can be calculated by

$$p = P\{t \geq t_0\} \tag{15}$$

The calculation results show in Table 3, illustrated that at the significance level $\alpha = 0.05$, t falls within the rejection domain, so the null hypothesis H_0 is rejected and the alternative hypothesis H_1 is accepted. By calculation, it is known that the minimum significance level p-value of rejecting the null hypothesis H_0 is far less than α, and the effect size of the mean difference δ more than 0.8. We can reasonably believe that when training the model, using the hybrid features as the input feature vector is significantly better than just applying PathCS.

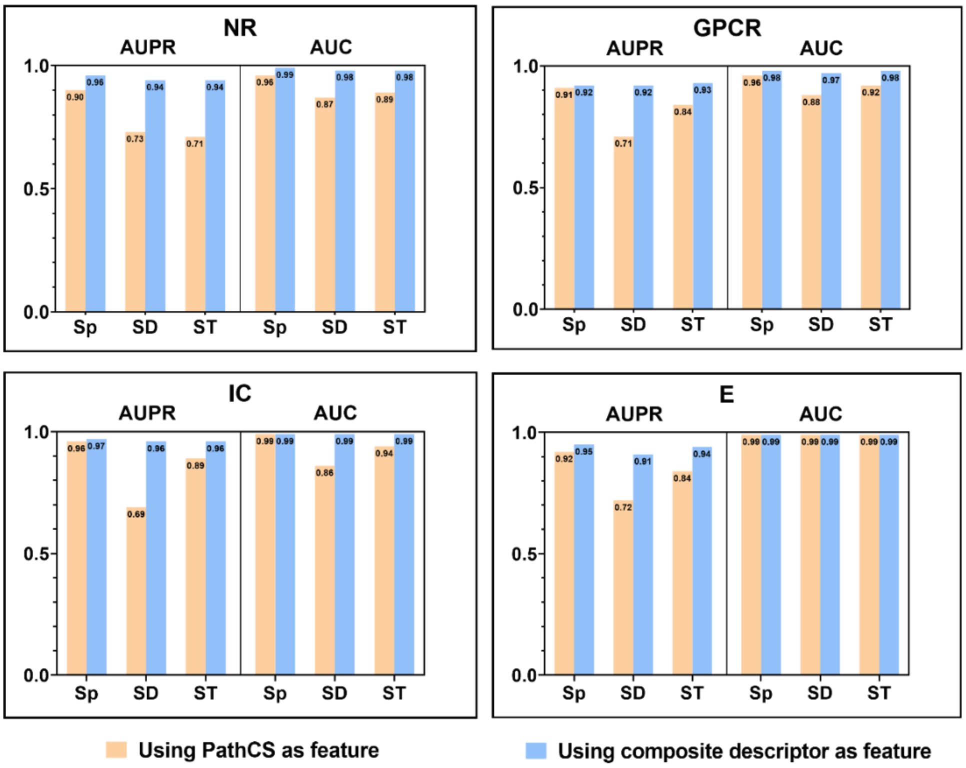

Fig. 2. Comparison of DTI-CDF methods that using PathCS only and hybrid features as a feature vector

In order to explore the reason why the combination of the three types of features improves the performance of the model, we analyzed the correlation between these three types of features. We performed a cross-correlation function analysis on the three types of features used in the model training on 12 experimental conditions. The cross-correlation function ρ_{xy} of two vectors x and y is defined as:

$$\rho_{xy} = \frac{R_{xy}(k)}{\sqrt{R_{xx}(0) \cdot R_{yy}(0)}} \tag{16}$$

$$R_{xx}(0) = \frac{1}{\|x\|} \sum_{i=1}^{x} x_i^2 \tag{17}$$

$$R_{yy}(0) = \frac{1}{\|y\|} \sum_{j=1}^{y} y_j^2 \tag{18}$$

$$R_{xy}(k) = \frac{1}{\|x\|} \sum_{i=1}^{\|x\|} x(i) y(i+k) \tag{19}$$

for $k = 0,1,2 \cdots, \|y\| - \|x\| + 1$, where $\|x\|$ is the length of x, $\|y\|$ is the length of y, and assume $\|y\| \geq \|x\|$. For each experimental condition, the data is first separated into three parts according to the three types of features and flattened separately to obtain three corresponding vectors with different lengths, of $N \times M$, where N is the number of the samples and M is the dimension of a feature vector. Then, the cross-correlation function among the three types of features is calculated according to Eq. (17-20). According to the characteristics of the flattened data, we make $k = 0, N, 2N, \cdots, \left\lfloor \frac{\|y\| - \|x\| + 1}{N} \right\rfloor N$ to reduce the computational cost. We choose the maximum value of the absolute value from ρ_{xy} as the correlation coefficient to measure the linear correlation between two types of features to ensure the reliability of the results, this value is in the range of [0,1], equal to 0 and 1 represent linear independent and complete linear correlation, respectively. Since the correlation coefficients of three experimental settings on a particular data set are very close, we selected the maximum value shown in Fig. 3. It can be seen that under the four data sets, the correlation coefficients among the three types of features are all below 10^{-4}, indicating that the data of different features are very close to irrelevant, thus, these three types of features can be considered to be orthogonal to each other, which can produce the greatest information when describing DTIs. Therefore, after adding FP2 and PsePSSM into a feature vector, the representation of DTIs is enriched without redundancy, thereby avoiding over-fitting and improving the performance of the model.

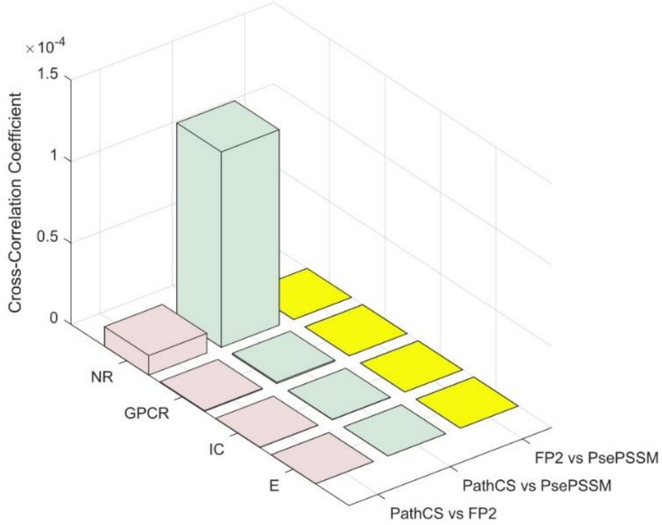

Fig. 3. Comparison of cross-correlation coefficients among three types of features

3.2 Effect of CDF model on the DTI-CDF performance

In the CDF model, one or more RF or XGB learners are added based on a single RF learner. Compared to a single RF, we demonstrate that the performance of CDF is better at predicting DTIs. In this regard, we used PathCS as input feature vector to train CDF model, then compared with the DDR method through AUPR and AUC metrics, as the DDR method uses PathCS to train the single RF model. Note that the experimental conditions of the above two methods are identical, such as the division of data sets, random state, etc. We observed that in the CDF model, the results are higher than those in the RF model except that the AUPR of the IC and E data sets under the experimental setting S_D are approximately equal to that of the RF model (as shown in Fig. 4). In the comparison of AUC, it was found that half of the results showed that the CDF model was higher than the RF model. In the other half, except for the IC data set under the experimental setting S_D, the difference between other results is no more than 3%. It is worth noting that for highly unbalanced data sets, AUPR is more accurate than AUC. Therefore, we have reason to think that the CDF model is better than the RF model. In addition, we prove this statistically with the hypothesis test.

Firstly, the difference between the above two methods of AUPR and AUC is calculated. Assuming that there is no significant difference between these two methods, the statistical hypothesis is still expressed by Eq. (12). According to t-test, the corresponding rejection domain is Eq. (14). The test results are shown in Table 3.

The calculation results show that at the significance level $\alpha = 0.05$, t falls within the rejection domain, so the null hypothesis H_0 is rejected and the alternative hypothesis H_1 is accepted. By calculation, it is known that the minimum significance level p-value of rejecting the null hypothesis H_0 is less than α, and the effect size of mean difference δ is 0.41. We can reasonably believe that using CDF model is medium significantly better than using the RF model for DTIs prediction.

It is not difficult to find that all the results under the experimental setting S_p are very good (i.e. both AUC and AUPR are more than 90%) when using the CDF model, while the results of the S_T and S_D experimental settings are not, a similar phenomenon also appears in the results of the DDR method. The reason is that the model has been over-fitting under the experimental settings of S_D and S_T, because the AUPR and AUC on the training set are both more than 0.9 when using CDF model. There are many reasons for over-fitting, the most important reason is that the model is too complex or the feature is not sufficient to describe the sample. From the above results, we can see that although the CDF model introduced by us is more complex than the RF model, it still achieves better results. Therefore, the crux of the over-fitting problem lies in the feature, which will be echoed in Section 3.1 and Section 3.3.

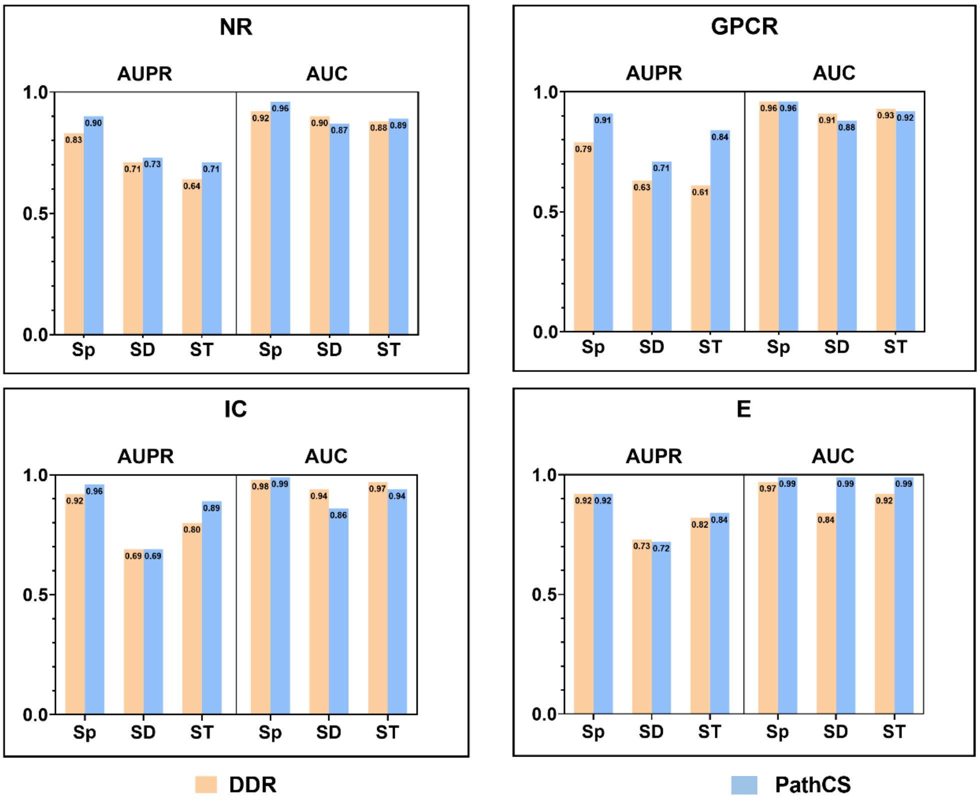

Fig 4. Comparison of DTI-CDF method that using PathCS as a feature vector and DDR method

3.3 Comparisons with the state-of-the-art algorithms (CDF vs. DDR)

Based on these four data sets, the DDR method [32] is proved as the most advanced method for predicting DTIs under the same experimental conditions (i.e. 5 repeated trails of 10-fold CVs under three experimental settings of each data set), so we only compare the proposed DTI-CDF method with the DDR method. Experiments show that DTI-CDF achieves better performance than DDR under the same conditions (Fig. 5), and all AUPRs exceeds 0.90, all AUCs are more than 0.96.

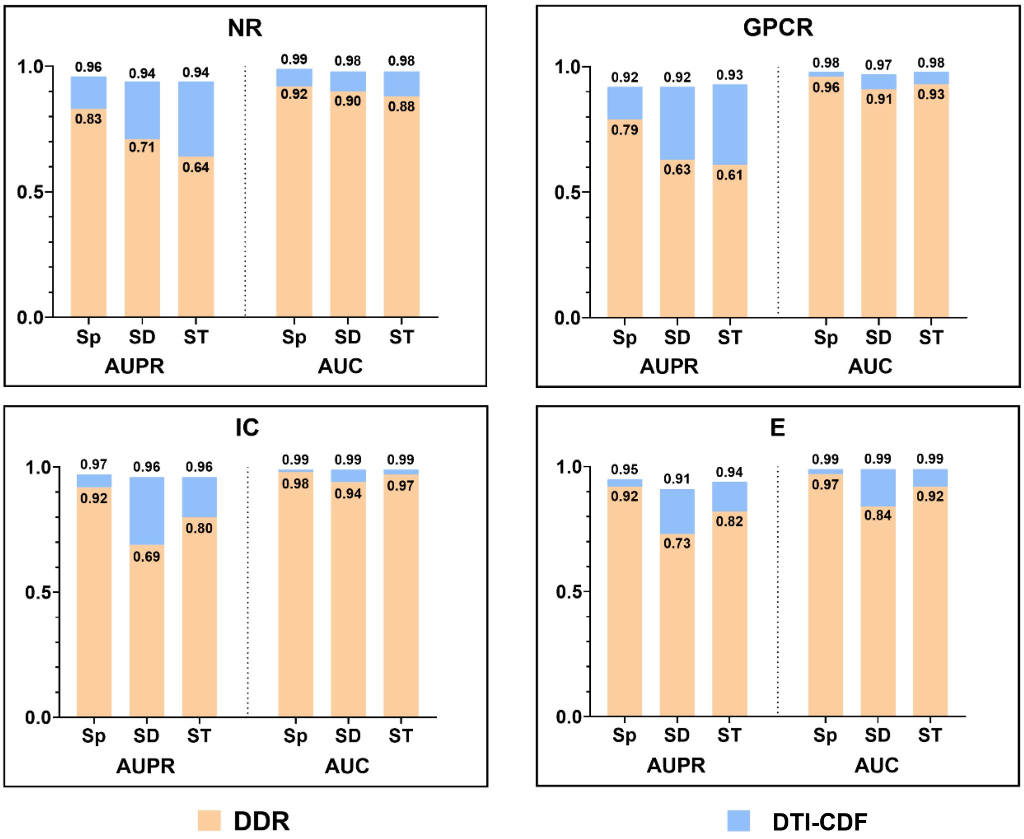

Fig. 5. Comparison of the AUPR and AUC of DTI-CDF with the DDR methods.

In order to compare the degree of difference between the DTI-CDF method and the DDR methods, we also carried out the one-sided paired t-test. Similarly, the AUPR and AUC of two methods under the same experimental conditions are subtracted in pairs. Assuming that there is no significant difference between the DTI-CDF method and the DDR method, Eq. (12) and Eq. (14) are still used to the test hypothesis and the rejection domain. The test results are listed in Table 3. The calculation results show that at the significance level $\alpha = 0.05$, t falls within the rejection domain, so the null hypothesis H_0 is rejected and the alternative hypothesis H_1 is accepted. By calculation, it is known that the minimum significance level p-value of rejecting the null hypothesis H_0 is far less than α, and the effect size of mean difference δ is larger than 0.8. We can reasonably believe that the DTI-CDF method is highly significantly better than the DDR method.

Table 3. The hypothesis test results

	t_0	p-value	δ
Compare hybrid features with PathCS.	6.11	1.56×10^{-6}	1.25
Compare CDF model with RF model.	2.03	0.03	0.41
Compare DTI-CDF method with the DDR method.	6.11	1.57×10^{-6}	1.25

4 Conclusions

We propose a DTI-CDF method to predict DTIs, which utilizes similarity information for drugs and targets, structural information for drugs, and evolution information for targets' sequence to obtain feature vector as the input of CDF algorithm for DTIs prediction. We use AUPR and AUC to evaluate the performance of the DTI-CDF method under three different experimental settings based on gold-standard data sets, all of them more than 0.9 and superior than the current state-of-art DDR method. It is further proved that the performance of the DTI-CDF method is significantly better than other existing methods when a known DTI is missing from the training data, especially in searching targets for new drugs (S_D setting) and finding drugs for new targets (S_T setting). This demonstrates that the DTI-CDF method has a higher predictive ability for the real scene of DTIs statistically. In addition, we have demonstrated that combining the similarity, structural and sequence information as feature vectors can better describe DTIs. We firstly use CDF algorithm in this field and prove its superiority. In the future, we plan to use the DTI-CDF method to deal with the regression problem such as calculating the affinity between drugs and targets.

References

[1] S. Anusuya, M. Kesherwani, K. V. Priya, A. Vimala, G. Shanmugam, D. Velmurugan, and M. M. Gromiha, "Drug-Target Interactions: Prediction Methods and Applications," *Current protein & peptide science,* vol. 19, p. 537, 2018-01-01 2018.

[2] X. Chen, C. C. Yan, X. Zhang, X. Zhang, F. Dai, J. Yin, and Y. Zhang, "Drug－target interaction prediction: databases, web servers and computational models," *Briefings in Bioinformatics,* vol. 17, pp. 696-712, 2016.

[3] R. Santos, O. Ursu, A. Gaulton, A. P. Bento, R. S. Donadi, C. G. Bologa, A. Karlsson, B. Al-Lazikani, A. Hersey, T. I. Oprea, and J. P. Overington, "A comprehensive map of molecular drug targets," *Nature Reviews Drug Discovery,* vol. 16, pp. 19-34, 2017.

[4] J. P. Overington, B. Al-Lazikani and A. L. Hopkins, "How many drug targets are there?" *Nature Reviews Drug Discovery,* vol. 5, pp. 993-996, 2006.

[5] D. C. Swinney and J. Anthony, "How were new medicines discovered?" *Nature Reviews Drug Discovery,* vol. 10, pp. 507-519, 2011.

[6] M. J. Keiser, B. L. Roth, B. N. Armbruster, P. Ernsberger, J. J. Irwin, and B. K. Shoichet, "Relating protein pharmacology by ligand chemistry," *Nature Biotechnology,* vol. 25, pp. 197-206, 2007.

[7] Huang, Sheng You , S. Z. Grinter , and X. Zou . "Scoring functions and their evaluation methods for protein–ligand docking: recent advances and future directions." PHYSICAL CHEMISTRY CHEMICAL PHYSICS 12.40(2010):12899-0.

[8] Y. Yamanishi, "Chemogenomic approaches to infer drug-target interaction networks," *Methods in molecular biology (Clifton, N.J.),* vol. 939, p. 97, 2013-01-01 2013.

[9] R. Chen, X. Liu, S. Jin, J. Lin, and J. Liu, "Machine Learning for Drug-Target Interaction Prediction," *Molecules,* vol. 23, p. 2208, 2018-08-31 2018.

[10] Z. Xia, L. Y. Wu, X. Zhou, and S. T. Wong, "Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces," *BMC Syst Biol,* vol. 4 Suppl 2, p. S6, 2010-09-13 2010.

[11] K. Bleakley and Y. Yamanishi, "Supervised prediction of drug－target interactions using bipartite local models," *Bioinformatics,* vol. 25, pp. 2397-2403, 2009-09-15 2009.

[12] G. Quan, D. Yongsheng, Z. Tongliang, and H. Tao, "Prediction drug-target interaction networks based on semi-supervised learning method,", 2016, pp. 7185-7188.

[13] H. Chen and Z. Zhang, "A semi-supervised method for drug-target interaction prediction with consistency in networks," *PloS one,* vol. 8, p. e62975, 2013-01-01 2013.

[14] W. Lan, J. Wang, M. Li, J. Liu, Y. Li, F. Wu, and Y. Pan, "Predicting drug－target interaction using

positive-unlabeled learning," *Neurocomputing*, vol. 206, pp. 50-57, 2016.

[15] T. Ma, C. Xiao, J. Zhou, and F. Wang, "Drug Similarity Integration Through Attentive Multi-view Graph Auto-Encoders," 2018-01-01 2018.

[16] L. Peng, B. Liao, W. Zhu, Z. Li, and K. Li, "Predicting Drug – Target Interactions With Multi-Information Fusion," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, pp. 561-572, 2017.

[17] H. Ding, I. Takigawa, H. Mamitsuka, and S. Zhu, "Similarity-based machine learning methods for predicting drug – target interactions: a brief review," *Briefings in Bioinformatics*, vol. 15, pp. 734-747, 2014-09-01 2014.

[18] J. Shi and S. Yiu, "SRP: A concise non-parametric similarity-rank-based model for predicting drug-target interactions,", 2015, pp. 1636-1641.

[19] X. Chen, M. Liu and G. Yan, "Drug – target interaction prediction by random walk on the heterogeneous network," *Molecular BioSystems*, vol. 8, p. 1970, 2012.

[20] F. Cheng, C. Liu, J. Jiang, W. Lu, W. Li, G. Liu, W. Zhou, J. Huang, and Y. Tang, "Prediction of drug-target interactions and drug repositioning via network-based inference," *PLoS Comput Biol*, vol. 8, p. e1002503, 2012-01-20 2012.

[21] T. van Laarhoven, S. B. Nabuurs and E. Marchiori, "Gaussian interaction profile kernels for predicting drug – target interaction," *Bioinformatics*, vol. 27, pp. 3036-3043, 2011-11-01 2011.

[22] S. Kim, D. Jin and H. Lee, "Predicting drug-target interactions using drug-drug interactions," *PLoS One*, vol. 8, p. e80129, 2013-01-20 2013.

[23] J. Mei, C. Kwoh, P. Yang, X. Li, and J. Zheng, "Drug – target interaction prediction by learning from local information and neighbors," *Bioinformatics*, vol. 29, pp. 238-245, 2013-01-15 2013.

[24] M. Gonen, "Predicting drug-target interactions from chemical and genomic kernels using Bayesian matrix factorization," *Bioinformatics*, vol. 28, pp. 2304-2310, 2012-09-15 2012.

[25] M. C. Cobanoglu, C. Liu, F. Hu, Z. N. Oltvai, and I. Bahar, "Predicting Drug – Target Interactions Using Probabilistic Matrix Factorization," *Journal of Chemical Information and Modeling*, vol. 53, pp. 3399-3409, 2013-12-23 2013.

[26] X. Zheng, H. Ding, H. Mamitsuka, and S. Zhu, "Collaborative matrix factorization with multiple similarities for predicting drug-target interactions,", 2013, pp. 1025-1033.

[27] A. Ezzat, P. Zhao, M. Wu, X. Li, and C. Kwoh, "Drug-Target Interaction Prediction with Graph Regularized Matrix Factorization," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 14, pp. 646-656, 2017.

[28] B. Bolgár and P. Antal, "VB-MK-LMF: fusion of drugs, targets and interactions using variational Bayesian multiple kernel logistic matrix factorization," *BMC Bioinformatics*, vol. 18, 2017.

[29] Y. Wang and J. Zeng, "Predicting drug-target interactions using restricted Boltzmann machines," *Bioinformatics*, vol. 29, pp. i126-i134, 2013-07-01 2013.

[30] G. Fu, Y. Ding, A. Seal, B. Chen, Y. Sun, and E. Bolton, "Predicting drug target interactions using meta-path-based semantic network analysis," *BMC Bioinformatics*, vol. 17, 2016.

[31] N. Zong, H. Kim, V. Ngo, and O. Harismendy, "Deep mining heterogeneous networks of biomedical linked data to predict novel drug – target associations," *Bioinformatics*, vol. 33, pp. 2337-2344, 2017-08-01 2017.

[32] R. S. Olayan, H. Ashoor and V. B. Bajic, "DDR: efficient computational method to predict drug – target interactions using graph mining and machine learning approaches," *Bioinformatics*, vol. 34, pp. 1164-1173, 2018-04-01 2018.

[33] F. Rayhan, S. Ahmed, S. Shatabda, D. M. Farid, Z. Mousavian, A. Dehzangi, and M. S. Rahman,

"iDTI-ESBoost: Identification of Drug Target Interaction Using Evolutionary and Structural Features with Boosting," *Scientific Reports,* vol. 7, 2017.

[34] A. Ezzat, M. Wu, X. Li, and C. Kwoh, "Drug-target interaction prediction using ensemble learning and dimensionality reduction," *Methods,* vol. 129, pp. 81-88, 2017.

[35] M. Wen, Z. Zhang, S. Niu, H. Sha, R. Yang, Y. Yun, and H. Lu, "Deep-Learning-Based Drug – Target Interaction Prediction," *Journal of Proteome Research,* vol. 16, pp. 1401-1409, 2017-04-07 2017.

[36] P. Hu, K. C. C. Chan and Z. You, "Large-scale prediction of drug-target interactions from deep representations,", 2016, pp. 1236-1243.

[37] K. Tian, M. Shao, Y. Wang, J. Guan, and S. Zhou, "Boosting compound-protein interaction prediction by deep learning," *Methods,* vol. 110, pp. 64-72, 2016.

[38] Y. Guo, S. Liu, Z. Li, and X. Shang, "BCDForest: a boosting cascade deep forest model towards the classification of cancer subtypes based on gene expression data," *BMC Bioinformatics,* vol. 19, 2018.

[39] R. Su, X. Liu, L. Wei, and Q. Zou, "Deep-Resp-Forest: A deep forest model to predict anti-cancer drug response," *Methods,* 2019.

[40] Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda, and M. Kanehisa, "Prediction of drug-target interaction networks from the integration of chemical and genomic spaces," *Bioinformatics,* vol. 24, pp. i232-i240, 2008-07-01 2008.

[41] M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, and Y. Yamanishi, "KEGG for linking genomes to life and the environment," *Nucleic Acids Research,* vol. 36, pp. D480-D484, 2007-12-23 2007.

[42] I. Schomburg, "BRENDA, the enzyme database: updates and major new developments," *Nucleic Acids Research,* vol. 32, pp. 431D-433, 2004-01-01 2004.

[43] S. Gunther, 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, and R. Preissner, "SuperTarget and Matador: resources for exploring drug-target relationships," *Nucleic Acids Research,* vol. 36, pp. D919-D922, 2007-12-23 2007.

[44] D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, and M. Hassanali, "DrugBank: a knowledgebase for drugs, drug actions and drug targets," *Nucleic Acids Research,* vol. 36, pp. D901-D906, 2008-01-01 2008.

[45] C. Leslie, E. Eskin and W. S. Noble, "The spectrum kernel: a string kernel for SVM protein classification," *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing,* p. 564, 2002-01-01 2002.

[46] M. Kuhn, I. Letunic, L. J. Jensen, and P. Bork, "The SIDER database of drugs and side effects," *Nucleic Acids Research,* vol. 44, pp. D1075-D1079, 2016-01-04 2016.

[47] M. Takarabe, M. Kotera, Y. Nishimura, S. Goto, and Y. Yamanishi, "Drug target prediction using adverse event report systems: a pharmacogenomic approach," *Bioinformatics*, vol. 28, pp. i611-i618, 2012-09-15 2012.

[48] B. Wang, A. M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains, and A. Goldenberg, "Similarity network fusion for aggregating data types on a genomic scale," *Nature Methods,* vol. 11, pp. 333-337, 2014.

[49] H. Shi, S. Liu, J. Chen, X. Li, Q. Ma, and B. Yu, "Predicting drug-target interactions using Lasso with random forest based on evolutionary information and chemical structure," *Genomics,* 2018.

[50] D. T. Jones, "Protein secondary structure prediction based on position-specific scoring matrices," *J Mol Biol,* vol. 292, pp. 195-202, 1999-09-17 1999.

[51] K. C. Chou, "Prediction of protein cellular attributes using pseudo-amino acid composition," *Proteins,* vol. 43, pp. 246-55, 2001-05-15 2001.

[52] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, "Open Babel: An open chemical toolbox," *J Cheminform,* vol. 3, p. 33, 2011-10-07 2011.

[53] F. J. Zhou Z H, "Deep Forest: Towards an Alternative to Deep Neural Networks," 2017.

[54] H. T. K, "Random Decision Forests," in *International Conference on Document Analysis & Recognition*, 1995.

[55] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System,", 2016, pp. 785-794.

[56] M. Hao, S. H. Bryant and Y. Wang, "Predicting drug-target interactions by dual-network integrated logistic matrix factorization," *Scientific Reports,* vol. 7, 2017.

[57] Y. Liu, M. Wu, C. Miao, P. Zhao, and X. Li, "Neighborhood Regularized Logistic Matrix Factorization for Drug-Target Interaction Prediction," *PLOS Computational Biology,* vol. 12, p. e1004760, 2016-02-12 2016.

[58] A. C. A. Nascimento, R. B. C. Prudêncio and I. G. Costa, "A multiple kernel learning algorithm for drug-target interaction prediction," *BMC Bioinformatics,* vol. 17, 2016.

[59] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves,", 2006, pp. 233-240.

[60] D. Rogers and M. Hahn, "Extended-Connectivity Fingerprints," *Journal of Chemical Information and Modeling,* vol. 50, pp. 742-754, 2010-05-24 2010.

[61] H. Nakashima and K. Nishikawa, "Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies," *Journal of molecular biology,* vol. 238, p. 54, 1994-01-01 1994.

[62] K. C. Chou, "Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes," *Bioinformatics,* vol. 21, pp. 10-19, 2005-01-01 2005.

[63] K. P. Burnham and D. R. Anderson, *Model Selection and Multimodel Inference A Practical Information-Theoretic Approach*, 2nd ed ed.: Springer-Verlag, 2002.