Original Research

Improved prediction of drug-target interactions based on ensemble learning with fuzzy local ternary pattern

Zheng-Yang Zhao[1], Wen-Zhun Huang[1,*], Xin-Ke Zhan[1], Yu-An Huang[1], Shan-Wen Zhang[1], Chang-Qing Yu[1]

[1]*School of Information Engineering, Xijing University, 710123 Xi'an, Shaanxi, China*

TABLE OF CONTENTS

1. Abstract

Introduction: The prediction of interacting drug-target pairs plays an essential role in the field of drug repurposing, and drug discovery. Although biotechnology and chemical technology have made extraordinary progress, the process of dose-response experiments and clinical trials is still extremely complex, laborious, and costly. As a result, a robust computer-aided model is of an urgent need to predict drug-target interactions (DTIs). **Methods**: In this paper, we report a novel computational approach combining fuzzy local ternary pattern (FLTP), Position-Specific Scoring Matrix (PSSM), and rotation forest (RF) to identify DTIs. More specially, the target primary sequence is first numerically characterized into PSSM which records the biological evolution information. Afterward, the FLTP method is applied in extracting the highly representative descriptors of PSSM, and the combinations of FLTP descriptors and drug molecular fingerprints are regarded as the complete features of drug-target pairs. **Results**: Finally, the entire features are fed into rotation forests for inferring potential DTIs. The experiments of 5-fold cross-validation (CV) achieve mean accuracies of 89.08%, 86.14%, 82.41%, and 78.40% on Enzyme, Ion Channel, GPCRs, and Nuclear Receptor datasets. **Discussion**: For further validating the model performance, we performed experiments with the state-of-art support vector machine (SVM) and light gradient boosting machine (LGBM). The experimental results indicate the superiorities of the proposed model in effectively and reliably detect potential DTIs. There is an anticipation that the proposed model can establish a feasible and convenient tool to identify high-throughput identification of DTIs.

2. Introduction

The identification of DTIs has turned into a focal point of pharmaceutical science to support screening the drug candidates and solving the problems of etiologies. The strikingly improved biochemical technologies have dramatically promoted the process of therapeutic drug discovery. In the last few years, Food and Drug Administration (FDA)

Submitted: 25 May 2021 Revised: 21 June 2021 Accepted: 2 July 2021 Published: 30 July 2021

has just approved a limited quantity of medicines due to the efficiency issues and harmful side effects [1]. Detecting interacting drug-target pairs is still of great significance to select the promising molecule drugs. The researchers have put much effort into exploring the DTIs based on traditional experiments. Nevertheless, the biochemical methods remain to be expensive and cumbersome. Furthermore, these methods need to face the contingency of serial results. Hence, the novel computer-aided drug development (CADD) models are essential to be constructed for stably and reliably inferring DTIs [2].

With the breakthrough of protein sequencing and drug molecular structure determination technologies, various sorts of databases including PubChem [3], ChEMBL [4], Therapeutic Target Database (TTD) [5], Kyoto Encyclopedia of Genes and Genomes (KEGG) [6], and DrugBank [7] are continuously enriching the public data of target proteins and drug sub-structures. Previously, computational-based prediction models mainly focused on molecular docking, ligand, and data mining [8]. However, there are some limitations in these traditional methods. For instance, the molecular docking method mainly predicts the binding sites by energy and geometry matching, it predicts the affinity of binding sites by computational simulation [9]. This method plays a critical role in determining the mode of drug actions. However, molecular docking requires all proteins in the model to have a complete 3D prediction structure that seriously limits the versatility of the model. The ligand-based method combines the chemical structure and pharmacological activity of a specific object through quantitative-structure activity relationships (QSAR), each model can only predict the relationship of one target [10]. The poor physical interoperability of the single model makes the method is hardly to be widely utilized in large-scale cross prediction. The data mining method collects DTIs by text mining and data matching [11]. The method is limited by mining algorithm and database authority, so it cannot achieve further promotion and application in DTIs prediction. In conclusion, the development of effective and robust models has become the essential requirement of DTIs prediction.

Bolgár *et al*. [12] proposed Variational Bayesian Multiple Kernel Logistic Matrix Factorization which embedded multiple kernel learning, weighted observation, and graph Laplacian regularization to model DTIs. Shi *et al*. [13] develop two-layer multiple classifier system (TLMCS) which focuses on fully utilizing heterogeneous features for better predicting DTIs. Xia *et al*. [14] proposed a novel model namely Self-Paced Learning with Collaborative Matrix Factorization based on weighted low-rank approximation (SPLCMF) to predict DTIs. Specifically, this framework employed regularized least squares to fuse the related networks and reduce the complexity of samples by soft weighting. Yan *et al*. [15] developed (substructure-drug-target Kronecker product kernel regularized least squares) SDTRLS model which integrates RLS-Kron model, chemical substructure similarity fusion, and Gaussian Interaction Profile (GIP) kernels to detect interacting drug-target pairs. Cui *et al*. [16] proposed L2,1-GRMF which is a developed GRMF method to identify the DTIs by combining L2,1-norm. Hao *et al*. [17] construct dual network integrated logistic matrix factorization (DNILMF) for drug structure matrix and target sequence kernel matrix to predict DTIs.

We established a novel *in silico* mothed to infer DTIs within this paper, this method mainly integrates PSSM, FLTP, and RF classifier. Specifically, the target primary sequences are first converted into numerical PSSM metrics which record the frequencies of amino acids that appear in different positions. Then, we employed FLTP approach to excavate the potential characteristics of PSSMs. Subsequently, we merge them and drug fingerprints as entire feature vectors of drug-target pairs. Finally, the full feature describers are fed into rotation forest to detect DTIs. We verified our model on the benchmark data sets, viz. Enzymes, Ion Channels, GPCRs, and Nuclear Receptors by utilizing 5-fold Cross-validation. Furthermore, we compared the established model with another advanced feature describer and various classifiers including LGBM and RF. The different experimental results illustrate that the proposed model has an outstanding effect on predicting DTIs, this model can reliably screen candidates for clinical trials. The flowchart of the established model is depicted in Fig. 1.

Table 1. Statistical description of benchmark dataset.

Statistics	Enzyme	Ion channel	GPCRs	Nuclear receptor
Drugs	445	210	223	54
Target proteins	664	204	95	26
Interactions	2926	1467	635	90

3. Materials and methods

3.1 Datasets

In this paper, the databases, viz. DrugBank [7], SuperTarget [18], BRENDA [19], and KEGG BRITE [6] provide four benchmark datasets including Enzyme, Ion Channel, GPCRs, and Nuclear Receptor for us to execute the established model. Enzyme data set stores 445 drugs, 664 proteins, and 2926 DTIs. Ion channel data set stores 210 drugs, 204 proteins, and 1467 DTIs. GPCRs data set stores 223 drugs, 95 proteins, and 635 DTIs. Nuclear Receptor data set stores 54 drugs, 26 proteins, and 90 DTIs. Table 1 clearly listed the experimental statistics of these benchmark datasets.

Drug and protein interactions were represented as a bipartite graph; drugs and proteins formed the nodes of the graph, and the verified interactions between them were denoted by edges within the graph. In the experiments, all drug-target pairs which are connected by edges are categorized to positive dataset, the other pairs are treated as neg-

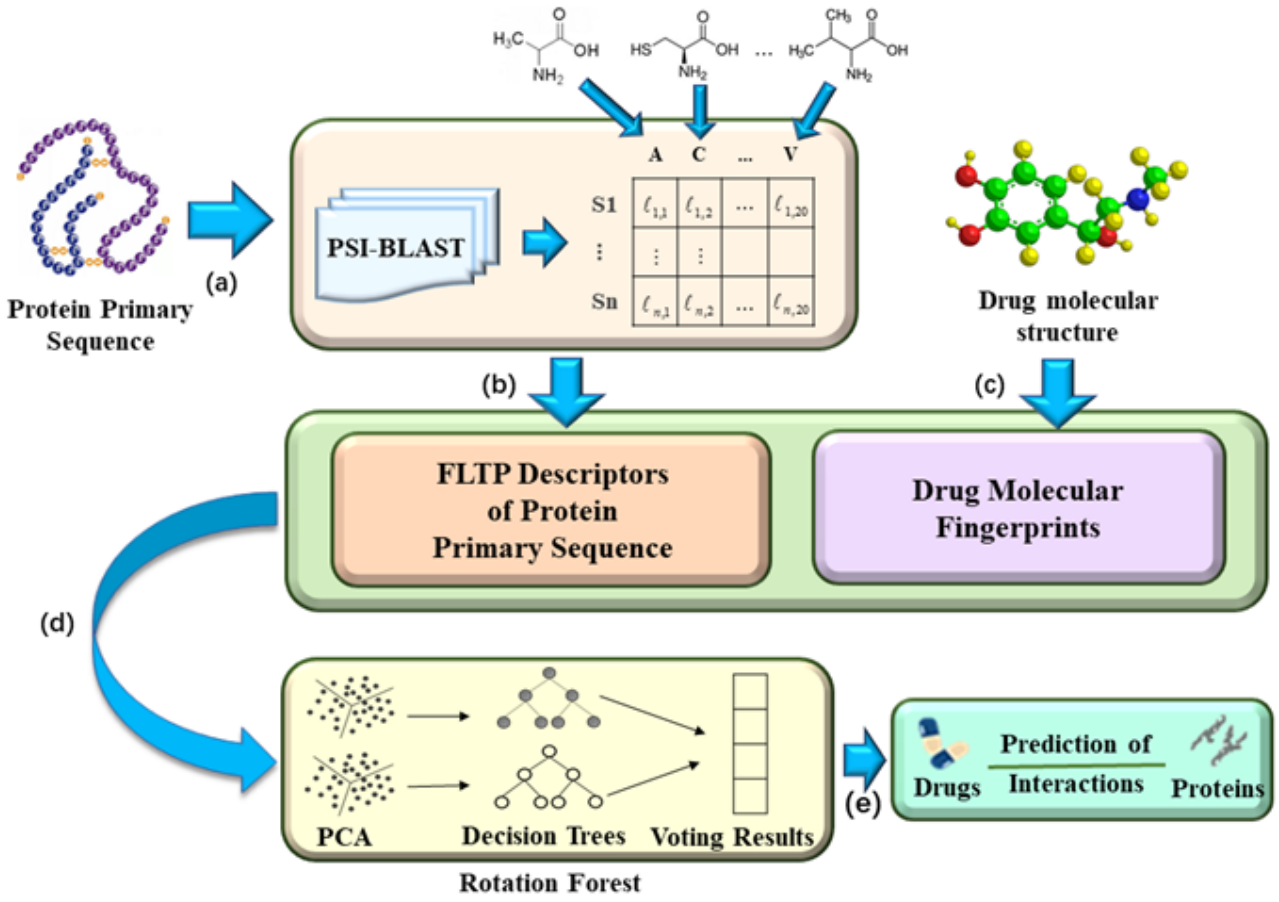

Fig. 1. Workflow of our model. (a) numerical convert the proteins to PSSMs. (b) characterize PSSMs by FLTP. (c) extract the molecular fingerprints of drugs. (d) feed the entire features into rotation forest. (e) predict DTIs.

ative samples. Considering the number of the nodes, the known interactions only take a little account of all relationships of drug-target pairs. Take GPCRs dataset for an example, there are 42,840 (210 × 204) types of relationships exist in the network. However, the interactions which have been certified by biotechnology are 1467 which accounts for 3.42%. Consequently, the down sampling algorithm is employed to extract the same number of negative samples as the positive data set to ensure sample balance. The 1467 positive samples are verified by clinical experiments, and the remaining samples were verified as negative samples. However, it is hardly to ignore the false verification caused by the error in the clinical experiment. Meanwhile, considering the number of negative samples only accounted for 3.55% of the remaining samples, the possibility that positive samples which exist in the remaining samples are assigned to negative data set can be ignored for the huge quantity gap.

3.2 Drug substructure characterization

Recently, the molecular fingerprints which contain chemical substructure information can effectively reflect drug structure [20]. It transforms the molecular structures into a series of binary fingerprint sequences by detecting specific fragments in the molecular structure [21]. Although the molecular is divided into several independent parts, it still ensures the integrality of the entire drug structural information [22]. Studies substantiate that the molecular fingerprints inhibit the information loss and accumulated error of screening procedures. Meanwhile, it also reduces the complexity of the calculation in the description process. Specifically, when the fraction matches a molecular substructure, the corresponding position of carrier will be assigned as 1. Mature fingerprint databases provide reliable tools for the generation of molecular fingerprints. We selected the fingerprint map which contains 881 substructures from Pubchem system (https://pubchem.ncbi.nlm.nih.gov/) [23]. Therefore, the describers of drug molecules are completely converted into a series of 881-dimentional Boolean vectors. Fig. 2 gives the transformation of Zanamivir into a fingerprint.

3.3 Position-specific scoring matrix

In recent years, various Physico-chemical methods are applied to numerically characterize protein which is composed of 20 types of letters [24]. Position-Specific

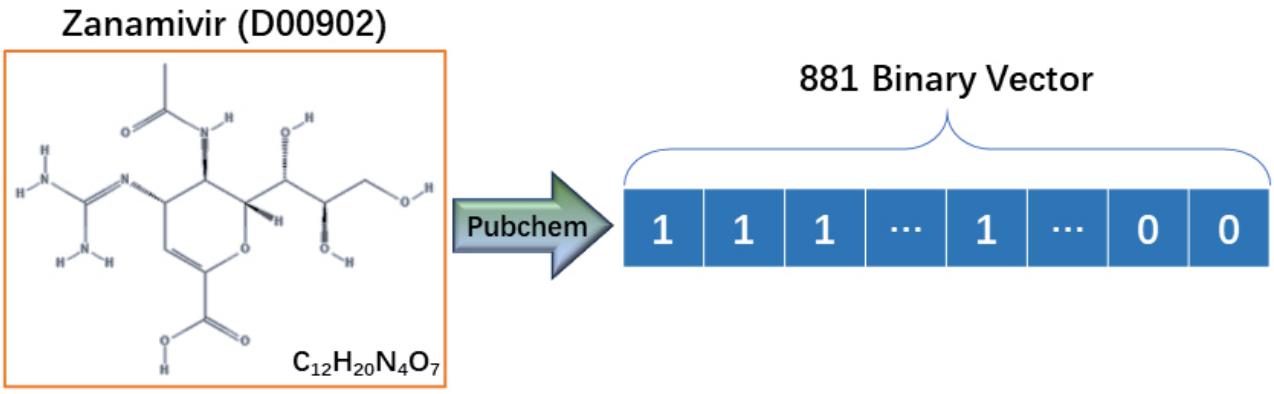

Fig. 2. The transformation of Zanamivir into a fingerprint.

Scoring Matrix (PSSM) is extensively utilized in protein binding site prediction, protein secondary structure prediction, and protein subcellular localization [25]. In this section, PSSM is employed to excavate the evolutionary information by calculating the probability of an amino acid emerges in a specific location of protein primary sequence. PSSM matrix is showed as follows.

$$PSSM = \begin{bmatrix} \ell_{1,1} & \ell_{1,2} & \cdots & \ell_{1,20} \\ \ell_{2,1} & \ell_{2,2} & \cdots & \ell_{2,20} \\ \vdots & \vdots & \ddots & \vdots \\ \ell_{n,1} & \ell_{n,2} & \cdots & \ell_{n,20} \end{bmatrix} \quad (1)$$

where PSSM is a matrix, where $L \times 20$ denotes the length of the target, and 20 represents the number of amino acids. $\ell_{i,j}$ represents the evolutionary score that ith residue mutate into jth amino acid during the evolutionary process. After optimizing Position-Specific Iterated Basic Local Alignment Search Tool (PSI-BLAST), parameter *e* is set to 0.001 and the iteration frequency is set to 3. Fig. 3 gives the example of Lipoprotein Lipase converting into PSSM.

3.4 Fuzzy local ternary pattern

Fuzzy Local Ternary Pattern (FLTP) can be utilized to precisely describe the texture feature, and it has a wide application in preventing face spoofing and image tampering areas [26]. For the anti-rotation ability of FLTP, it is also robust to the noise in the image. This method dynamically calculates the threshold based on Weber's law to extract multiple features. Meanwhile, it can be extended to circles and neighborhoods with different radius. In this paper, FLTP is employed to describe the characteristics of PSSMs. The algorithm converts the difference between neighborhood pixels and center pixels into the upper and lower binary codes. The upper binary code can be expressed as $FLTP\_S_{P,R}^{upper}$, the lower one can be expressed as $FLTP\_S_{P,R}^{lower}$. The complete describer is defined as follows:

$$FLTP\_S_{P,R} = \left[FLTP\_S_{P,R}^{upper}, FLTP\_S_{P,R}^{lower} \right. \quad (2)$$

where $FLTP\_S_{P,R}^{upper}$ can be calculated as follows.

$$FLTP\_S_{P,R}^{\text{upper}}\ (x_c, y_c) = \sum_{i=0}^{P-1} s\left(i_i - (i_c + \tau)\right) 2^i \quad (3)$$

$$s(x) = \begin{cases} 1 & , x \geq 0 \\ 0 & , \textit{otherwhise} \end{cases} \quad (4)$$

where $FLTP\_S_{P,R}^{lower}$ can be calculated as follows.

$$FLTP\_S_{P,R}^{lower}\ (x_c, y_c) = \sum_{i=0}^{P-1} s\left(i_i - (i_c - \tau)\right) 2^i \quad (5)$$

$$s(x) = \begin{cases} 1 & , x < 0 \\ 0 & , \textit{otherwhise} \end{cases} \quad (6)$$

where (x_c, y_c) represents the circular central pixel, and i_i represents the gray value of neighborhood pixel. When calculating the upper binary code, if the gray value of neighborhood pixel is greater than $i_c + \tau$, the neighborhood pixel is marked as 1. When calculating the lower binary code, if the gray value of the neighborhood pixel is less than $i_c - \tau$, the neighborhood pixel is marked as 0. The gray value of the circular central pixel i_i which was generated by the nonlinear interpolation algorithm and dynamic threshold τ are calculated as follows.

$$i_i = I\left(x_c + R\sin\frac{2\pi i}{P}, y_c - R\cos\frac{2\pi i}{P}\right) \quad (7)$$

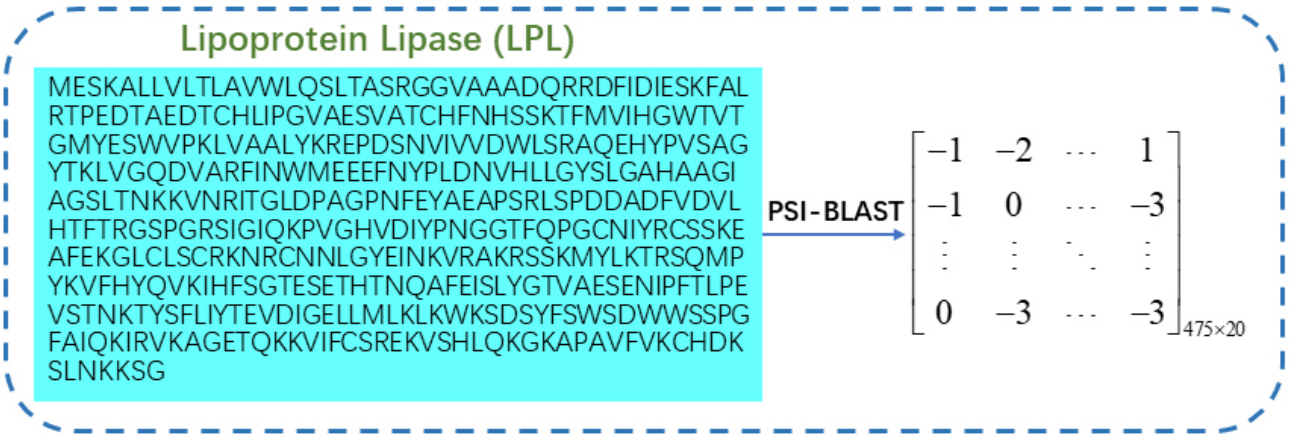

Fig. 3. The example of Lipoprotein Lipase converting into PSSM.

$$\frac{|i_i - i_c|}{i_c} = \tau \tag{8}$$

Finally, the FLTP feature vector can be obtained as follow.

$$H(k) = \sum_{i=0}^{I}\sum_{j=0}^{J} f\left(FLTP\_S_{P,R}(i,j),k\right), \quad k \in [0,255] \tag{9}$$

In this experiment, the radius of the circular domain R = 1, the number of pixels in circular domain P = 8. The upper and lower binary codes are transformed into 256 dimensional vectors respectively. Hence, the entire descriptor of PSSM is a matrix of 1×512.

3.5 Rotation forest

Rodriguez *et al.* [27] proposed rotation forest (RF) based on integrated forest [27, 28]. This ensemble classifier succeeds in the classification of small-sized data set. Significantly, RF also has good effects on promoting sample difference [29]. Within the experiments, we utilized rotation forest to detect DTIs. Firstly, RF stochastically separates the sample set into L disjoint subsets. Subsequently, Principal Component Analysis (PCA) approaches to convert subsets to generate rotation forest. Finally, send them to different base classifiers for scorning each subtree. The matrix C of $n \times N$ is regarded to be the train set containing N features of n samples, and $T = (t_1, t_2, \cdots, t_n)^T$ gathers the labels of different samples. The method has K base classifiers R_i. The sequential training steps of the base classifier are as follows.

(I) Follow obtaining the optimized parameter L, dataset P is separated to L disjoint subsets stochastically, each subset has N/L features.

(II) Let $P_{i,j}$ represents jth the subset of P, and $C_{i,j}$ denotes the feature set of $P_{i,j}$. Then calculate the new training features set $C'_{i,j}$ by bootstrap sampling on 75% of $C_{i,j}$.

(III) Execute PCA on $C'_{i,j}$ to get the principal component coefficients $a_{i,j}^{(1)}, a_{i,j}^{(2)}, \cdots a_{i,j}^{(m_j)}$.

(IV) These coefficients make up the sparse rotation matrix Q_i as:

$$Q_i = \begin{bmatrix} a_{i,1}^{(1)}, a_{i,1}^{(2)}, \cdots a_{i,1}^{(m_1)} & 0 & \cdots & 0 \\ 0 & a_{i,2}^{(1)}, a_{i,2}^{(2)}, \cdots a_{i,2}^{(m_2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{i,K}^{(1)}, a_{i,K}^{(2)}, \cdots a_{i,K}^{(m_K)} \end{bmatrix} \tag{10}$$

In the process of classification, the possibility that sample x belongs to category t_i is $d_{i,j}\left(xQ_i^a\right)$ yielded by base classifier R_i. Afterword, calculate the confidence degrees that x belongs to different classes as follows:

$$\theta_j(x) = \frac{1}{K}\sum_{i=1}^{K} d_{i,j}\left(xQ_i^a\right) \tag{11}$$

Finally, the sample x will be classified in accordance with the degree.

4. Results and discussion

4.1 Evaluation criteria

For improving the reliability of the experimental performance, the evaluative indices, viz. accuracy (Acc.), precision (Prec.), sensitivity (Sen.), specificity (Spec.), and Matthews correlation coefficient (MCC) are utilized to analyze the results of 5-fold CV.

$$Acc. = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

$$\Pr ec. = \frac{TP}{TP + FP} \tag{13}$$

$$Sen. = \frac{TP}{TP + FN} \tag{14}$$

$$Spec. = \frac{TN}{TP + FP} \tag{15}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FN) \times (TN + FP)}} \tag{16}$$

where true positive (TP) records the aggregate of interacting drug-target pairs which were assigned to positive set; true negative (TN) denotes the sum of non-interacting drug-target pairs which were assigned to negative set; false positive (FP) is the quantity of non-interacting drug-target pairs which were assigned in positive set; false negative (FN) denotes the count of interacting drug-target pairs which were assigned to negative set. In addition, the receiver operating characteristic (ROC) curves were pictured to visualize the prediction results [30], the area under the curves (AUC) was also attached to ROC for justifying the established model [31]. We also utilized PR curves and AUPR values to indicate the sample balance and model performance.

4.2 Parameter discussion

In RF classifier, the main parameters K and L denote the numbers of feature sub-sets and decision trees which affect the classification accuracy. To get the optimal parameters, this paper employs grid-search algorithm to study the influence of parameters on prediction results [32]. When L-value increased from 0 to 38, the experimental results show that the accuracy was increasing, then it decreased sharply. Meanwhile, the accuracy was growing with the increase of K-value. In consideration of the model efficiency, the optimal parameters K and L are set to 18 and 38, respectively. Fig. 4 depicts the prediction accuracy surface with factors of K-value and L-value.

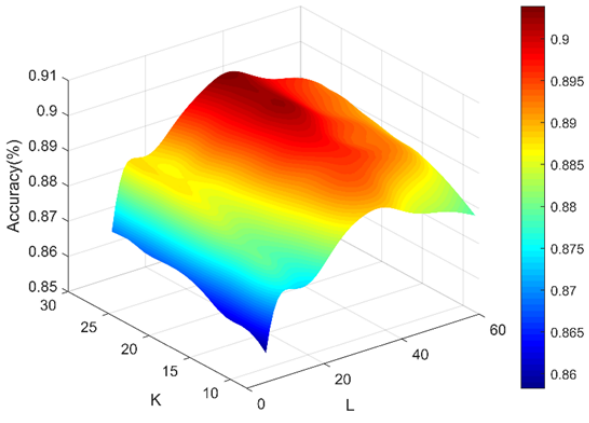

Fig. 4. Accuracy surface of the optimization on K-value, and L-value.

4.3 Five-fold CV results on four datasets

To certify the feasibility of the established model and avoid over-fitting, we executed 5-fold CV on four benchmark data sets with the same parameters. Specifically, each data set is separated into 5 equal-sized and disjointed fractions. The independent fractions take turns to be treated as test sets, while the other fractions serve as train sets. Tables 2,3,4,5 display the experimental results of our method on four standard data sets.

The statistics of results has been shown in Table 6. The average criteria of accuracy, sensitivity, precision, specificity and Matthews correlation coefficient are 89.08%, 90.32%, 87.52%, 90.62%, and 78.17% on *Enzyme* data set. Their standard deviations are 0.68%, 0.59%, 1.21%, 0.43%, and 1.32%. We obtained the average criteria of 86.14%, 86.46%, 85.69%, 86.60%, and 72.28% on *Ion Channel* data set. Their standard deviations are 1.67%, 2.61%, 1.18%, 2.42%, and 3.37%. On *GPCRs* data set, our model generated the average criteria of 82.41%, 82.10%, 81.97%, 82.96%, and 64.85% with standard deviation of 2.20%, 3.48%, 3.12%, 1.60%, and 4.43%. In terms of *Nuclear Receptor* dataset, the average criteria are 78.40%, 76.33%, 77.78%, 76.43%, and 56.02%, respectively, with standard deviation of 5.07%, 7.02%, 14.65%, 5.99%, and 12.21%. As can be noted, the small size of *Nuclear Receptor* data set leads to a higher standard deviation. Figs. 5,6,7,8 record the performance of our model on four benchmark datasets, while the average AUC values of 0.9535, 0.9292, 0.8901, and 0.8534 are also attached to them. Figs. 9,10,11,12 plot the PR curve of our model on four golden standard datasets, while the average AUPR values of 0.9608, 0.9345, 0.8941, and 0.8636 are also attached to them.

4.4 Comparison of FLTP and ZMs models

For strictly validating the feature describing ability of fuzzy local ternary pattern (FLTP) method. We constructed the comparative experiment by replacing FLTP descriptors with Zernike Moments (ZMs) descriptors which have strong Rotational Invariance [33, 34]. ZMs method is widely utilized in the field of edge detection by extracting global feature information at different scales [35]. Table 7 shows the comparison of ZMs and FLTP with the same classifier. These experimental statistic shows that FLTP method has a significant performance improvement compared with Zernike Moments on benchmarks. The criteria values entirely get promoted on *Enzyme*, *Ion Channel*, and *GPCRs* dataset. Fig. 13 displays the mean ROC curves of FLTP model and ZMs model by an interpolation method. It is noteworthy that the AUC values of FLTP-embedded model are comprehensive greater than ZMs model, and the mean value gaps attain 2.55%, 0.89%, 1.17%, and 4.09%, respectively. The results indicate that our model provides an effective way to characterize PSSM for detecting potential DTIs.

Table 2. Experimental results yield by 5-fold CV on *Enzyme* dataset.

Test set	Acc. (%)	Pre. (%)	Sen. (%)	Spec. (%)	MCC (%)
1	88.46	90.51	86.05	90.89	77.02
2	89.15	89.61	87.18	90.91	78.23
3	88.55	89.88	87.14	89.98	77.14
4	89.06	91.12	87.88	90.38	78.15
5	90.17	90.46	89.35	90.95	80.33
Average	89.08 ± 0.68	90.32 ± 0.59	87.52 ± 1.21	90.62 ± 0.43	78.17 ± 1.32

Table 3. Experimental results yield by 5-fold CV on *Ion Channel* dataset.

Test set	Acc. (%)	Pre. (%)	Sen. (%)	Spec. (%)	MCC (%)
1	88.31	90.21	86.29	90.38	76.69
2	84.58	85.33	84.49	84.67	69.14
3	87.12	87.46	86.87	87.37	74.24
4	84.41	83.16	84.34	84.47	68.77
5	86.27	86.15	86.44	86.10	72.54
Average	86.14 ± 1.67	86.46 ± 2.61	85.69 ± 1.18	86.60 ± 2.42	72.28 ± 3.37

Table 4. Experimental results yield by 5-fold CV on *GPCRs* dataset.

Test set	Acc. (%)	Pre. (%)	Sen. (%)	Spec. (%)	MCC (%)
1	79.92	77.77	80.99	78.95	59.87
2	84.65	84.92	84.25	85.04	69.29
3	84.65	86.29	82.95	86.40	69.36
4	80.63	81.20	81.82	79.34	61.18
5	82.21	80.30	84.80	79.69	64.54
Average	82.41 ± 2.20	82.10 ± 3.48	81.97 ± 3.12	82.96 ± 1.60	64.85 ± 4.43

Table 5. Experimental results yield by 5-fold CV on *Nuclear Receptors* dataset.

Test set	Acc. (%)	Pre. (%)	Sen. (%)	Spec. (%)	MCC (%)
1	73.88	81.25	56.52	76.92	42.32
2	86.11	78.95	93.75	80.00	73.41
3	80.56	83.33	78.95	82.35	61.21
4	74.29	66.67	71.43	76.19	47.14
5	77.14	71.43	88.24	66.67	56.01
Average	78.40 ± 5.07	76.33 ± 7.02	77.78 ± 14.65	76.43 ± 5.99	56.02 ± 12.21

Table 6. The statistics of results yield by 5-fold CV on four benchmark datasets.

Statistics	Evaluation criteria	Acc.	Pre.	Sen.	Spec.	MCC
Enzyme	Average	89.08	90.32	87.52	90.62	78.17
	Standard deviation	0.68	0.59	1.21	0.43	1.32
Ion Channel	Average	86.14	86.46	85.69	86.60	72.28
	Standard deviation	1.67	2.61	1.18	2.42	3.37
GPCRs	Average	82.41	82.10	81.97	82.96	64.85
	Standard deviation	2.20	3.48	3.12	1.60	4.43
Nuclear Receptors	Average	78.40	76.33	77.78	76.43	56.02
	Standard deviation	5.07	7.02	14.65	5.99	12.21

4.5 Comparison with other classifiers

Thus far, some machine learning-based classifiers are utilized to identify DTIs. To fairly verify the performance of the proposed model, we embed the state of art support vector machine (SVM) and light gradient boosting machine (LGBM) algorithm into our model with fuzzy local ternary pattern. Within RF classifier, we set parameters K = 18, L = 38 which was discussed above. The SVM utilized inner product kernel function instead of nonlinear mapping to high dimensional space, it also adopts small-sample learning method to greatly simplify the process of classification and regression. There are 400 experiments with different combinations of parameters c and g were car-

Table 7. Performance comparison of fuzzy local ternary pattern with Zernike Moments.

Dataset	Model	Acc. (%)	Prec. (%)	Sen. (%)	Spec. (%)	MCC (%)	AUPR (%)
Enzyme	FLTP + RF	89.08 ± 0.68	90.32 ± 0.59	87.52 ± 1.21	90.62 ± 0.43	78.17 ± 1.32	96.08 ± 0.32
	ZMs + RF	86.13 ± 0.53	87.15 ± 0.79	85.25 ± 1.76	87.40 ± 1.13	72.69 ± 1.13	93.97 ± 0.38
Ion Channel	FLTP + RF	86.14 ± 1.67	86.46 ± 2.61	85.69 ± 1.18	86.60 ± 2.42	72.28 ± 3.37	93.45 ± 0.92
	ZMs + RF	84.00 ± 1.14	84.14 ± 2.83	84.00 ± 3.02	84.11 ± 3.11	68.13 ± 2.23	91.88 ± 1.26
GPCRs	FLTP + RF	82.41 ± 2.20	82.10 ± 3.48	81.97 ± 3.12	82.96 ± 1.60	64.85 ± 4.43	89.41 ± 2.24
	ZMs + RF	81.50 ± 3.15	81.27 ± 5.73	81.46 ± 5.84	81.62 ± 3.70	63.06 ± 6.33	88.21 ± 1.65
Nuclear Receptor	FLTP + RF	78.40 ± 5.07	76.33 ± 7.02	77.78 ± 14.65	76.43 ± 5.99	56.02 ± 12.21	86.36 ± 5.86
	ZMs + RF	75.15 ± 5.34	76.19 ± 10.26	75.90 ± 9.22	75.78 ± 11.93	51.83 ± 11.41	81.09 ± 6.24

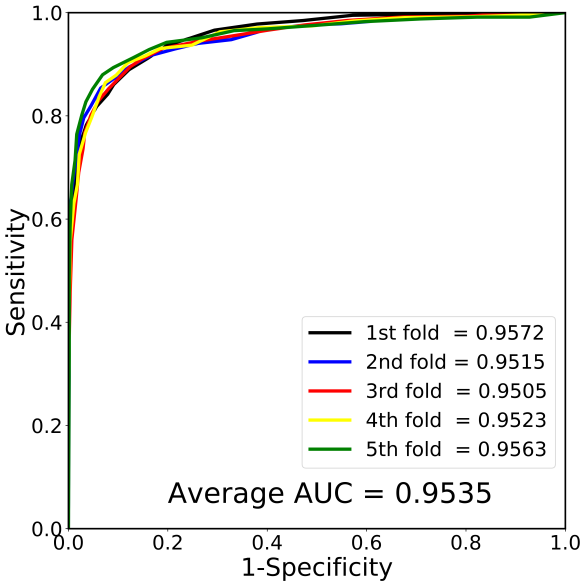

Fig. 5. The ROC curves generated by 5-fold CV on *Enzyme* dataset.

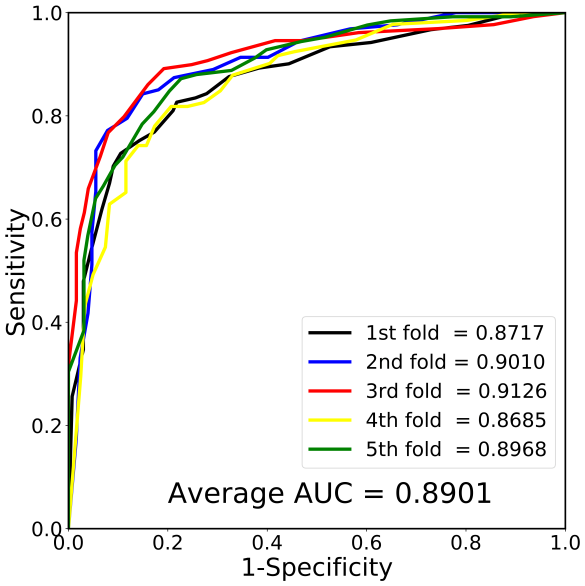

Fig. 7. The ROC curves generated by 5-fold CV on *GPCRs* dataset.

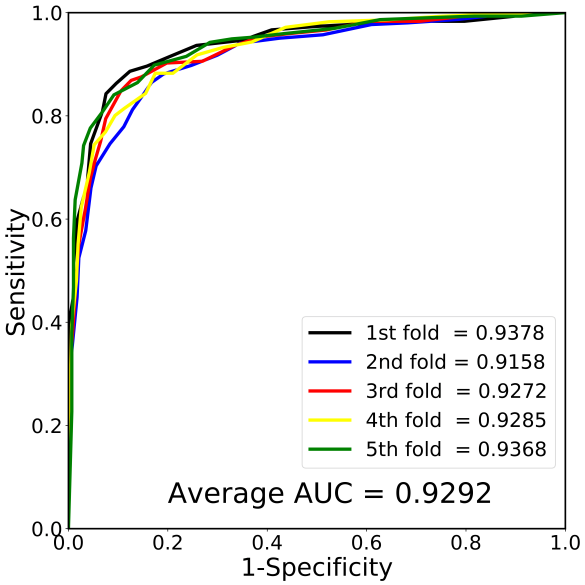

Fig. 6. The ROC curves generated by 5-fold CV on *Ion Channel* dataset.

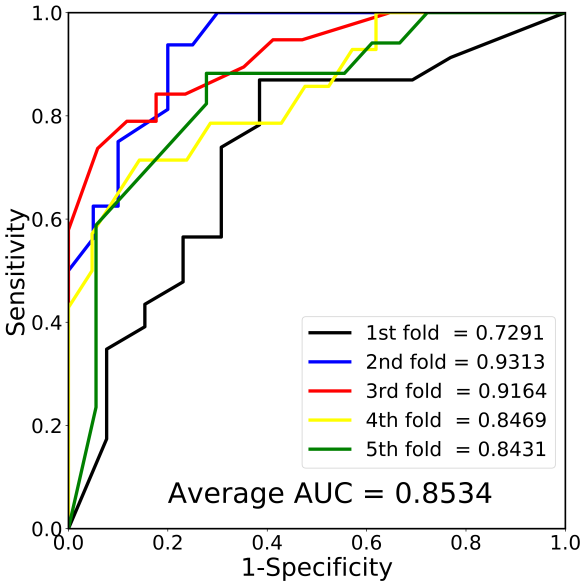

Fig. 8. The ROC curves generated by 5-fold CV on *Nuclear Receptors* dataset.

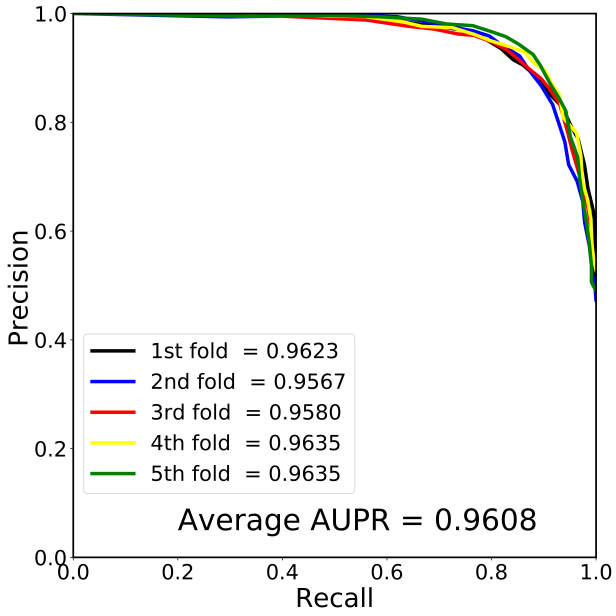

Fig. 9. The PR curves generated by 5-fold CV on *Enzyme* dataset.

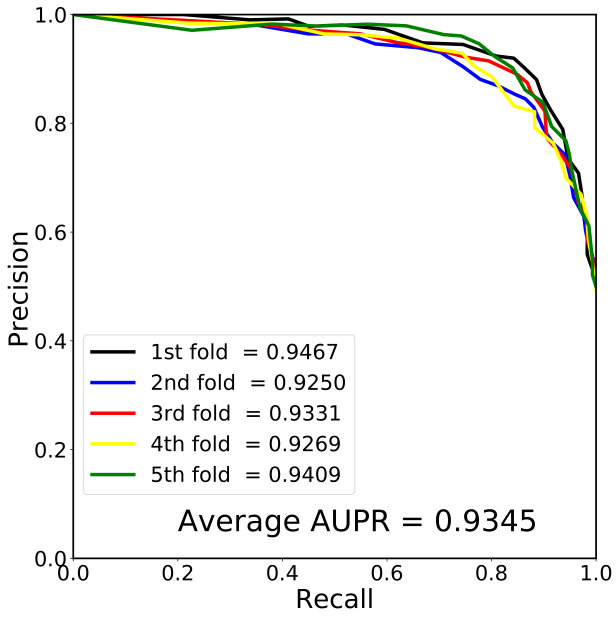

Fig. 10. The PR curves generated by 5-fold CV on *Ion Channel* dataset.

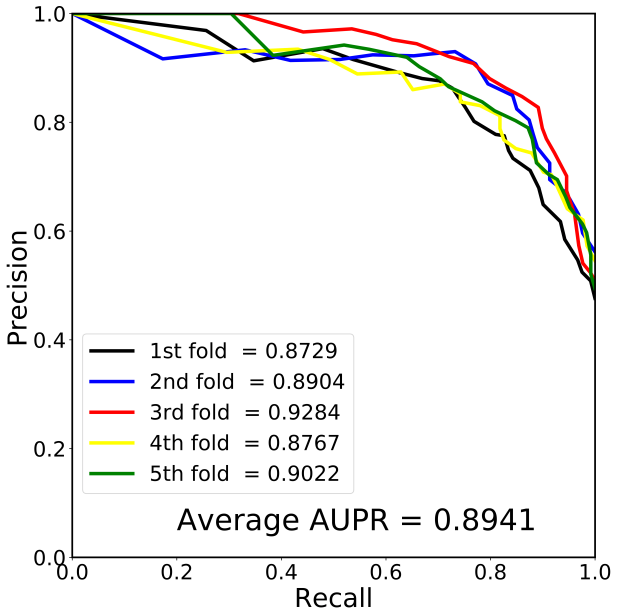

Fig. 11. The PR curves results generated by 5-fold CV on GPCRs dataset.

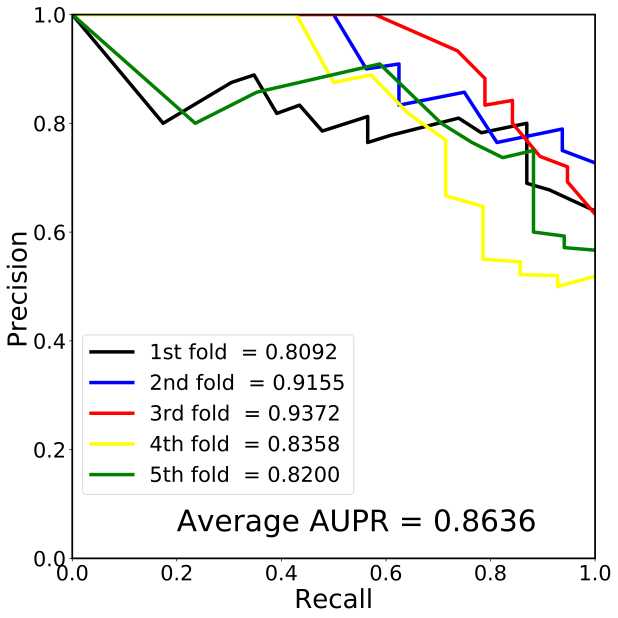

Fig. 12. The PR curves generated by 5-fold CV on *Nuclear Receptors* dataset.

ried out to get the highest accuracy, and we set *c*-value, *g*-value to 0.7 and 40, respectively. The kernel of SVM was select as radial basis function (RBF) based on LIBSVM tool. The LGBM method is the improved gradient boosting decision trees (GBDT) algorithm to reduce the time cost and power consumption in industrial applications. After parameter optimizations, the leaves-number, the learning rate, and the training rounds were set to 55, 0.05, and 37, respectively.

Fig. 14 records the comparison between RF, LGBM, and SVM on *Enzyme*, *Ion Channel*, *GPCRs*, and *Nuclear Receptor* data sets. The results indicate that model which embeds RF classifier has higher prediction accuracy. Compared with SVM classifier, the average accuracy promotions of RF are 10.49%, 10.57%, 8.40%, and 15.20%, the accuracy gaps between RF and LGBM are 3.93%, 3.24%, 3.21%, 6.77% on four benchmark dataset. Figs. 15,16 plot the ROC curves of the golden standard datasets based on the rates of 1-specificity against sensitivity. The model which has higher AUC values predict

more accurate. As shown in Figs. 15,16, the AUC value gaps of four data sets attain to 0.1051, 0.1162, 0.0944, and 0.2232 between RF and SVM, the value gaps between RF and LGBM attain to 0.1013, 0.1013, 0.0910, and 0.1329, respectively. Therefore, it is considered that the proposed model is more efficient at predicting DTIs.

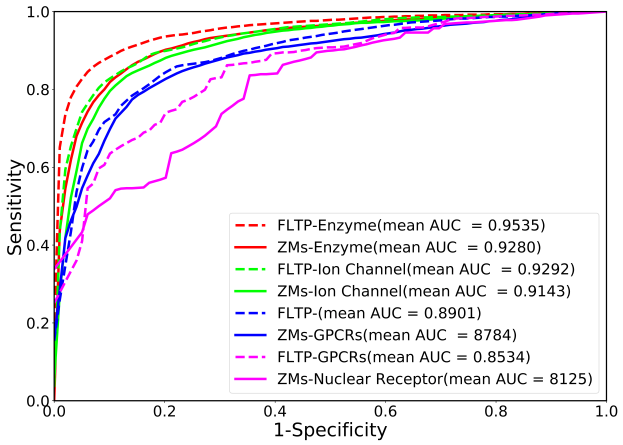

Fig. 13. Comparison of average AUC values on FLTP and ZMs.

Table 8. Comparison between our model with state-of-art methods in terms of benchmark data sets.

Dataset	Method	AUC	AUPR
Enzyme	SIMCOMP	0.876	0.358
	DCT	0.909	0.873
	Bigram-PSSM	0.948	0.546
	LOOP	0.951	0.933
	Our method	0.954	0.961
Ion Channel	SIMCOMP	0.767	0.274
	DCT	0.893	0.812
	Bigram-PSSM	0.889	0.39
	LOOP	0.935	0.921
	Our method	0.929	0.935
GPCRs	SIMCOMP	0.867	0.452
	DCT	0.867	0.793
	Bigram-PSSM	0.872	0.282
	LOOP	0.886	0.865
	Our method	0.890	0.894
Nuclear Receptor	SIMCOMP	0.856	0.435
	DCT	0.799	0.628
	Bigram-PSSM	0.869	0.411
	LOOP	0.819	0.822
	Our method	0.853	0.864

4.6 Comparison with previous methods

So far, numerous advanced models have been established to predict DTIs and assist drug design. In this section, we compared our model with partial state-of-art models for fully evaluating the model performance by adopting 5-fold CV. After experimenting the previous methods such as SIMCOMP [36], DCT [37], Bigram-PSSM [38], LOOP [39] on benchmark datasets. Table 8 gives the comparison of AUC value and AUPR values. It is clearly that the performance of the established model has risen significantly. Although the AUC value of our model is 0.006 lower than LOOP on *Ion Channel* dataset, the AUC values of *Enzyme*, *GPCRs*, and *Nuclear Receptors* have grown 0.003, 0.004, and 0.034, respectively, and the AUPR values of four benchmark datasets have grown 0.028, 0.014, 0.029, and 0.042, respectively. As a result, the experiments substantiate that the model which combining FLTP descriptors and rotation forest can remarkably enhance the performance of predicting DTIs.

5. Conclusions

In summary, this paper integrates Position-Specific Scoring Matrix, fuzzy local ternary pattern, and rotation forest as a novel prediction algorithm for identifying the relationships between drugs and targets. Specifically, the fusions which combine FLTP describers of PSSMs and drug molecular fingerprints are fed into RF for inferring DTIs. The mean accuracies of our model were 89.08%, 86.14%, 82.41%, and 78.40% on standard data sets. We also made systematic comparisons to ensure the superiority of our model. First, the Zernike Moments (ZMs) method was utilized to alter the FLTP method to validate the feature description ability. Second, the state-of-art SVM, LGBM with FLTP features are experimented to access the performance of RF. The results indicate that this computational can be regarded as a significantly reliable tool for screening feasible candidates for medical trials.

6. Limitation and future work

Besides achieving more accurate prediction results than previous models, we also noticed the limitations of our model. This section will analyze these limitations from two aspects. On one side, the fuzzy local ternary pattern only describes the local texture characteristics. This feature descriptor is hardly to capture the global information of the sample, which leads to the singleness of the feature of PSSM. To extract more excellent feature vectors, future work will focus on fusion features. We will study a variety of local and global feature extraction methods and combine them to build a prediction model. On the other side, the loss and noise of data samples have a great effect on the accuracy of the model. We will explore two-dimensional data sample filtering algorithms to reduce data noise and improve data robustness. Meanwhile, we will further optimize the parameters to keep the integrity of the samples for accurate prediction. In general, the subsequent work will concentrate on extracting more accurate supervised classifiers and more fusion features which integrate the texture features and contour features of PSSMs. The growth of high

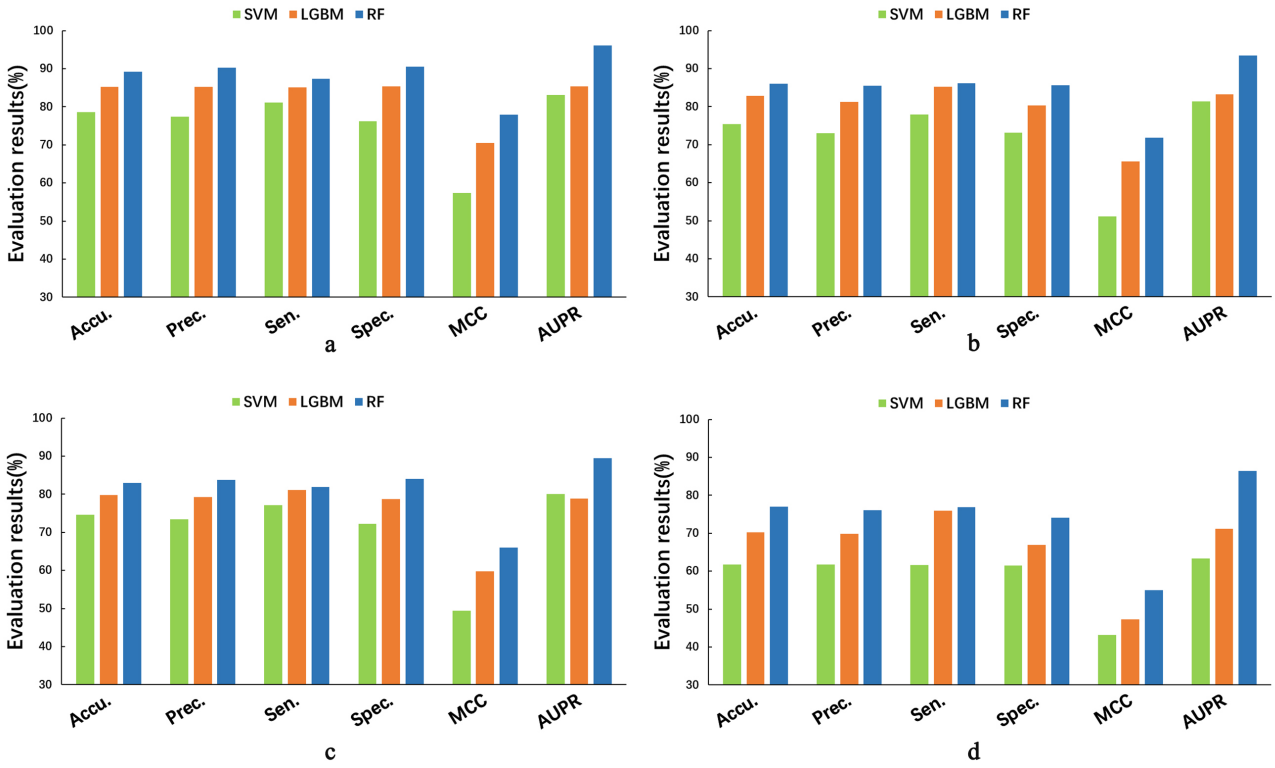

Fig. 14. Comparison of advanced classifiers on gold standard data sets. (a) 5-fold CV results on *Enzyme* data set. (b) 5-fold CV results on *Ion Channel* data set. (c) 5-fold CV results on *GPCRs* data set. (d) 5-fold CV results on *Nuclear Receptors* data set.

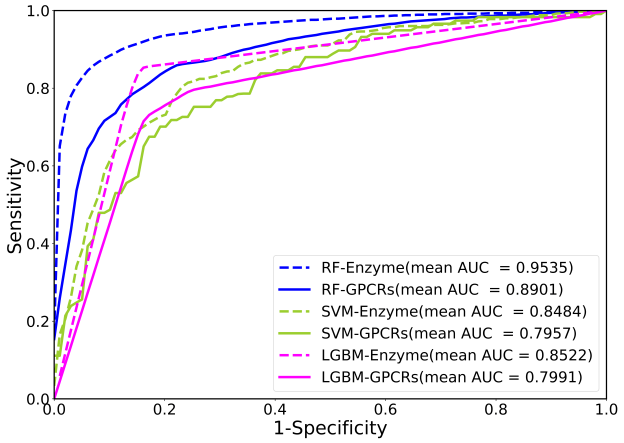

Fig. 15. ROC curves obtained by different classifiers on *Enzyme* and *GPCRs* datasets.

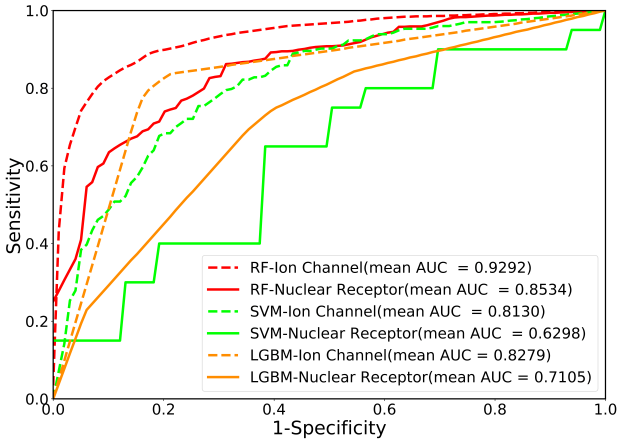

Fig. 16. ROC curves obtained by different classifiers on *Ion Channel* and *Nuclear Receptor* datasets.

throughput data set will create favorable circumstances and challenges for constructing auxiliary tools to enhance the accuracy of identification.

7. Author contributions

ZYZ handled the Conceptualization. ZYZ and XKZ performed the methodology, software, and validation. YAH curated the data. WZH, SWZ, and CQY administrated the project. WZH handled the funding acquisition.

8. Ethics approval and consent to participate

Not applicable.

9. Acknowledgment

We thank Zhu-Hong You for technical assistance. Thanks to all the peer reviewers for their opinions and suggestions.

10. Funding

This research was supported by the National Natural Science Foundation of China under Grant No. 62072378.

11. Conflict of interest

The authors declare no conflict of interest.

12. Data and software

https://github.com/zhaozhiya-20/Predict-the-interaction-of-DTIs-combining-FLTP-and-RF.

13. References

[1] Redkar S, Mondal S, Joseph A, Hareesha KS. A Machine Learning Approach for Drug-target Interaction Prediction using Wrapper Feature Selection and Class Balancing. Molecular Informatics. 2020; 39: 1900062.

[2] Zeng H, Wu X. Alzheimer's disease drug development based on Computer-Aided Drug Design. European Journal of Medicinal Chemistry. 2016; 121: 851–863.

[3] Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, *et al*. PubChem Substance and Compound databases. Nucleic Acids Research. 2016; 44: D1202–D1213.

[4] Willighagen EL, Waagmeester A, Spjuth O, Ansell P, Williams AJ, Tkachenko V, *et al*. The ChEMBL database as linked open data. Journal of Cheminformatics. 2013; 5: 23.

[5] Zhu F, Han B, Kumar P, Liu X, Ma X, Wei X, *et al*. Update of TTD: Therapeutic Target Database. Nucleic Acids Research. 2010; 38: D787–D791.

[6] Kanehisa MGS. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research. 2000; 28: 27–30.

[7] Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, *et al*. DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Research. 2008; 36: D901–D906.

[8] Yamanishi Y, Kotera M, Kanehisa M, Goto S. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. Bioinformatics. 2010; 26: i246–i254.

[9] Morris GM, Lim-Wilby M. Molecular docking. Methods in Molecular Biology. 2008; 443: 365–382.

[10] Alsenan SA, Al-Turaiki IM, Hafez AM. Feature Extraction Methods in Quantitative Structure–Activity Relationship Modeling: A Comparative Study. IEEE Access. 2020; 8: 78737–78752.

[11] Percha B GY, ALTMAN RB. Discovery and explanation of drug-drug interactions via text mining. Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing. 2012; 410–421.

[12] Bolgar B, Antal P. VB-MK-LMF: fusion of drugs, targets and interactions using variational Bayesian multiple kernel logistic matrix factorization. BMC Bioinformatics. 2017; 18: 440.

[13] Shi JY, Li JX, Mao KT, Cao JB, Lei P, Lu HM, *et al*. Predicting combinative drug pairs via multiple classifier system with positive samples only. Computer Methods and Programs in Biomedicine. 2019; 168: 1–10.

[14] Xia LY, Yang ZY, Zhang H, Liang Y. Improved Prediction of Drug-Target Interactions Using Self-Paced Learning with Collaborative Matrix Factorization. Journal of Chemical Information and Modeling. 2019; 59: 3340–3351.

[15] Yan C, Wang J, Lan W, Wu F-X, Pan Y. SDTRLS: Predicting Drug-Target Interactions for Complex Diseases Based on Chemical Substructures. Complexity. 2017; 2017: 1–10.

[16] Cui Z, Gao YL, Liu JX, Dai LY, Yuan SS. L2,1-GRMF: an improved graph regularized matrix factorization method to predict drug-target interactions. BMC Bioinformatics. 2019; 20: 287.

[17] Hao M, Bryant SH, Wang Y. Predicting drug-target interactions by dual-network integrated logistic matrix factorization. Scientific Reports. 2017; 7: 40376.

[18] Hecker N, Ahmed J, von Eichborn J, Dunkel M, Macha K, Eckert A, *et al*. SuperTarget goes quantitative: update on drug-target interactions. Nucleic Acids Research. 2012; 40: D1113–D1117.

[19] Schomburg I, Chang A, Ebeling C, Gremse M, Heldt C, Huhn G, *et al*. BRENDA, the enzyme database: updates and major new developments. Nucleic Acids Research. 2004; 32: D431–D433.

[20] Wu Z, Cheng F, Li J, Li W, Liu G, Tang Y. SDTNBI: an integrated network and chemoinformatics tool for systematic prediction of drug-target interactions and drug repositioning. Briefings in Bioinformatics. 2017; 18: 333–347.

[21] Liang X, Zhu W, Lv Z, Zou Q. Molecular Computing and Bioinformatics. Molecules. 2019; 24: 2358.

[22] Shen J CF, Xu Y, Li W, Tang Y. Estimation of ADME properties with substructure pattern recognition. Journal of Chemical Information and Modeling. 2010; 50: 1034–1041.

[23] Xu C, Cheng F, Chen L, Du Z, Li W, Liu G, *et al*. In silico prediction of chemical Ames mutagenicity. Journal of Chemical Information and Modeling. 2012; 52: 2840–2847.

[24] Altschul S.F. KEV. Iterated profile searches with PSI-BLAST-a tool for discovery in protein databases. Trends in Biochemical Sciences. 1998; 23: 444–447.

[25] Gribskov M, McLachlan AD, Eisenberg D. Profile analysis: detection of distantly related proteins. Proceedings of the National Academy of Sciences of the United States of America. 1987; 84: 4355–4358.

[26] Kavitha P, Vijaya K. Fuzzy local ternary pattern and skin texture properties based countermeasure against face spoofing in biometric systems. Computational Intelligence. 2020; 37: 559–577.

[27] Rodriguez JJ KL, Alonso CJ. Rotation Forest: A New Classifier Ensemble Method. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2006; 28: 1619–1630.

[28] Wei L, Su R, Wang B, Li X, Zou Q, Gao X. Integration of deep feature representations and handcrafted features to improve the prediction of N6-methyladenosine sites. Neurocomputing. 2019; 324: 3–9.

[29] Zhang F, Wang Y, Ni J, Zhou Y, Hu W. SAR Target Small Sample Recognition Based on CNN Cascaded Features and AdaBoost Rotation Forest. IEEE Geoscience and Remote Sensing Letters. 2019; 17: 1008–1012.

[30] Wei L, Zhou C, Chen H, Song J, Su R. ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. Bioinformatics. 2018; 34: 4007–4016.

[31] Wei L, Hu J, Li F, Song J, Su R, Zou Q. Comparative analysis and prediction of quorum-sensing peptides using feature representation learning and machine learning algorithms. Briefings in Bioinformatics. 2020; 21: 106–119.

[32] Brito JA, McNeill FE, Webber CE, Chettle DR. Grid search: an innovative method for the estimation of the rates of lead exchange between body compartments. Journal of Environmental Monitoring. 2005; 7: 241–247.

[33] Khotanzad A HY. Invariant Image Recognition by Zernike Moments. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1990; 12: 489–497.

[34] YS KWaK. Region-based shape descriptor using Zernike moments. Signal Processing: Image Communication. 2000; 16: 95–102.

[35] Kumar Y, Aggarwal A, Tiwari S, Singh K. An efficient and robust approach for biomedical image retrieval using Zernike

moments. Biomedical Signal Processing and Control. 2018; 39: 459–473.
[36] Ozturk H, Ozkirimli E, Ozgur A. A comparative study of SMILES-based compound similarity functions for drug-target interaction prediction. BMC Bioinformatics. 2016; 17: 128.
[37] Wang L, You ZH, Li LP, Yan X, Zhang W. Incorporating chemical sub-structures and protein evolutionary information for inferring drug-target interactions. Scientific Reports. 2020; 10: 6641.
[38] Mousavian Z, Khakabimamaghani S, Kavousi K, Masoudi-Nejad A. Drug-target interaction prediction from PSSM based evolutionary information. Journal of Pharmacological and Toxicological Methods. 2016; 78: 42–51.
[39] Zhan X, You Z-H, Cai J, Li L, Yu C, Pan J, *et al*. Prediction of Drug-Target Interactions by Ensemble Learning Method From Protein Sequence and Drug Fingerprint. IEEE Access. 2020; 8: 185465–185476.

Abbreviations: DTIs, drug-target interactions; FLTP, fuzzy local ternary pattern; PSSM, Position-Specific Scoring Matrix; RF, rotation forest; CV, cross-validation; SVM, support vector machine; LGBM, light gradient boosting machine; FDA, food and drug administration; CADD, computer-aided drug development; TTD, therapeutic target database; KEGG, Kyoto encyclopedia of genes and genomes; QSAR, quantitative-structure activity relationships; TLMCS, two-layer multiple classifier system; SDTRLS, substructure-drug-target Kronecker product kernel regularized least squares; DNILMF, dual network integrated logistic matrix factorization; PSI-BLAST, position-specific iterated basic local alignment search tool; PCA, principal component analysis; TP, true positive; TN, true negative; FP, false positive; FN, false negative; ROC, receiver operating characteristic; AUC, area under the curves; ZMs, Zernike moments; RBF, radial basis function; GBDT, gradient boosting decision trees.

Keywords: Drug-target interactions; Fuzzy local ternary pattern; Drug molecular fingerprints; Rotation forest

Send correspondence to: Wen-Zhun Huang, School of Information Engineering, Xijing University, 710123 Xi'an, Shaanxi, China, E-mail: huangwenzhun@xijing.edu.cn

Temporary page!

LaTeX was unable to guess the total number of pages correctly. As there was some unprocessed data that should have been added to the final page this extra page has been added to receive it.

If you rerun the document (without altering it) this surplus page will go away, because LaTeX now knows how many pages to expect for this document.