Hindawi
Scientific Programming
Volume 2021, Article ID 6686409, 10 pages
https://doi.org/10.1155/2021/6686409

Research Article

A Sparse Feature Extraction Method with Elastic Net for Drug-Target Interaction Identification

Zheng-Yang Zhao, Wen-Zhun Huang, Jie Pan, Yu-An Huang, Shan-Wen Zhang, and Chang-Qing Yu

School of Information Engineering, Xijing University, Xi'an 710123, China

Correspondence should be addressed to Wen-Zhun Huang; huangwenzhun@xijing.edu.cn

Received 9 December 2020; Revised 30 January 2021; Accepted 9 February 2021; Published 24 February 2021

Academic Editor: Wenzheng Bao

The identification of drug-target interactions (DTIs) plays a crucial role in drug discovery. However, the traditional high-throughput techniques based on clinical trials are costly, cumbersome, and time-consuming for identifying DTIs. Hence, new intelligent computational methods are urgently needed to surmount these defects in predicting DTIs. In this paper, we propose a novel computational method that combines position-specific scoring matrix (PSSM), elastic net based sparse features extraction, and rotation forest (RF) classifier. Specifically, we converted each protein primary sequence into PSSM, which contains biological evolutionary information. Then we extract the hidden sparse feature descriptors in PSSM by elastic net based sparse feature extraction method (ESFE). After that, we fuse them with the features of drug, which are represented by molecular fingerprints. Finally, rotation forest classifier works on detecting the potential drug-target interactions. When performing the proposed method by the experiments of fivefold cross validation (CV) on enzyme, ion channel, G protein-coupled receptors (GPCRs), and nuclear receptor datasets, this method achieves average accuracies of 90.32%, 88.91%, 80.65%, and 79.73%, respectively. We also compared the proposed model with the state-of-the-art support vector machine (SVM) classifier and other effective methods on the same datasets. The comparison results distinctly indicate that the proposed model possesses the efficient and robust ability to predict DTIs. We expect that the new model will be able to take effects on predicting massive DTIs.

1. Introduction

Identification of DTIs plays an increasingly critical part in drug development. Drug-target interactions guarantee the health promotion by preventing and treating diseases. Although the biological research has made a great progress in identifying DTIs, the pharmaceutical research method is still time-consuming and expensive [1]. Meanwhile, it is supposed that the market demand for new drugs remains strong. Consequently, computer-aided drug development (CADD) [2] methods are developed to reduce the cost and complexity of DTIs prediction on a large scale.

In past years, the databases including DrugBank [3], PubChem [4], Therapeutical Target Database (TTD) [5], and ZINC [6] have provided the data of small molecule drugs and biotechnology drugs. Furthermore, they also provide biological and chemical information such as molecular structures, drug-target interactions, and characteristics of relevant drug [7]. These data can be downloaded in various formats, and many reliable models have been designed to predict DTIs based on these databases.

By the time, a series of limitations still exist in the traditional computational models to detect interacting drug-target pairs. The model that creates a quantitative relationship based on the structures and pharmacological activities of compounds is hard to achieve the accuracy requirements of DTIs prediction for the lacks of physical interpretation [8–10]. The molecular docking model has a poor performance on large-scaled DTIs prediction for it cannot fit the proteins without three-dimensional (3D) structure information [11–14]. The model with ligand [15] is based on pharmacophore mapping. It is difficult to efficiently apply this model on account of the small number of the known ligands. The text mining methods [16, 17]

including information matching and other retrieval technologies are limited by the mining algorithms to detect new interactions in the literature database. However, proteins metastructure provides a new biologic way to describe the proteins with primary sequences. In fact, it has a better application prospect in the identification of DTIs.

In the last few years, researchers have achieved many advances in DTIs identification. Paska et al. [18] contributed a prediction model based on Bayesian personalized matrix decomposition. Ding et al. [19] proposed a double Laplacian regularized least-squares (DLapRLS) method based on Hilbert-Schmidt independence criterion and multikernel learning (HSIC-MKL) model. Specifically, it builds kernels for multiple information sources and then uses alternating least squares to train it. Shi et al. [20] developed a novel approach with triple matrix factorization (TMF) to find out the characteristics that include dominant feature pairs, frequent substructures, and conservative amino acid triplets. Zheng et al. [21] proposed the multiple similarities collaborative matrix factorization (MSCMF) model, which projects drugs and targets into a common low-rank feature space. Keum and Nam [22] used self-training SVM to classify the feature vectors extracted by bipartite local model (BLM). Lately, Wang et al. [23] developed a machine learning approach to excavate DTIs. In this method, the protein sequences were transformed into PSSM by counting the occurrence frequency of amino acids in the same position. Then, discrete cosine transform (DCT) is utilized to describe the PSSM, while encoding drug molecules as feature vectors.

In the experiments, we proposed a novel computational model that combines PSSM, elastic net based sparse feature extraction (ESFE) method, and RF classifier to identify drug-target identifications. Firstly, we converted the protein primary sequences into PSSM to retain the biological evolution information. After that, we combined the sparse features of PSSM with drug molecular fingerprint information. Finally, RF classifier is used to predict the DTIs. In order to evaluate the model, this paper uses fivefold CV method on the datasets of enzyme, ion channel, GPCRs, and nuclear receptor. As complement to the evaluation, we also compared the proposed model with support vector machine (SVM) and several previous models. The comprehensive results show that the proposed model effectively generates accurate predictions of DTIs. Figure 1 shows the flow chart for detecting interacting drug-target pairs by the proposed model.

2. Materials and Methods

2.1. Datasets. In this work, we choose the datasets containing enzyme, ion channel, GPCRs, and nuclear receptor, which were collected from DrugBank [2], SuperTarget [24], BRENDA [25], and KEGG BRITE [26] by Yamanshi. In the datasets, the number of drugs is 445, 210, 223, and 54, respectively, and the number of target proteins is 664, 204, 95, and 26, respectively. It has been demonstrated that the number of interacting drug-target pairs which has been verified for the datasets is 2926, 1467, 635, and 90, respectively. Table 1 concretely gives the statistics of drugs, target proteins, and interacting drug-target pairs on the datasets.

In general, we utilize a bipartite graph [27] to depict the network of DTIs. The nodes of the bipartite graph match the drug molecules, and the connections between the nodes match the relationships of drugs-target. For example, 2926 real edges exist in the sparse network on enzyme dataset. These edges represent the drug-target interactions, which have been verified. The total number of drug-target pairs is 295,480 (445×664); 2926 drug-target pairs with interactions are regarded as positive samples, and the residual 292,554 (295480−2926) pairs represent the potential negative samples. For ensuring the balance of samples, we randomly selected the same number of negative samples as positive samples. In theory, a few interacting drug-target pairs exist in negative samples. Considering that the selected samples only account for about 1% of the negative samples, we ignore the possibility that the interacting drug-target pairs are selected as negative samples. The numbers of negative samples selected on four datasets are 2926, 1467, 635, and 90, respectively.

2.2. Drug Substructure Characterization. The previous studies indicate that we can use topological, constitutional, and quantum chemical properties to describe drug compounds. In this paper, we established the fingerprints of drug molecular substructure to effectively encode the substructures of drug compounds. The fingerprint takes the form of a feature vector containing 881 Boolean values, and each Boolean value in the vector corresponds to a certain molecular substructure. When the drug molecule has a certain molecular substructure, the corresponding Boolean value in the vector will be set to 1; otherwise, it will be set to 0. This paper utilized PubChem system to achieve the chemical structure of the drug molecular fingerprints [28].

2.3. Position-Specific Scoring Matrix (PSSM). On the basis of the physical and chemical properties of amino acids [29, 30], we can transform the primary protein sequence into a multidimensional matrix. Position-specific scoring matrix (PSSM) was introduced by Gribskov et al. [31] in 1987. It provides an effective way to excavate the information of biological evolution from different kinds of amino acids. In this section, PSSM was utilized to extract the features of target proteins. We convert primary sequences of proteins into PSSM with position-specific iterated basic local alignment search tool (PSI-BLAST) [32]. The matrix is expressed as follows:

$$\mathrm{PSSM} = \begin{bmatrix} \ell_{1,1} & \ell_{1,2} & \cdots & \ell_{1,20} \\ \ell_{2,1} & \ell_{2,2} & \cdots & \ell_{2,20} \\ \vdots & \vdots & \ddots & \vdots \\ \ell_{n,1} & \ell_{n,2} & \cdots & \ell_{n,20} \end{bmatrix}, \tag{1}$$

where PSSM is a matrix of $n \times 20$, $n$is the length of protein sequence, and the number 20 represents the quantity of

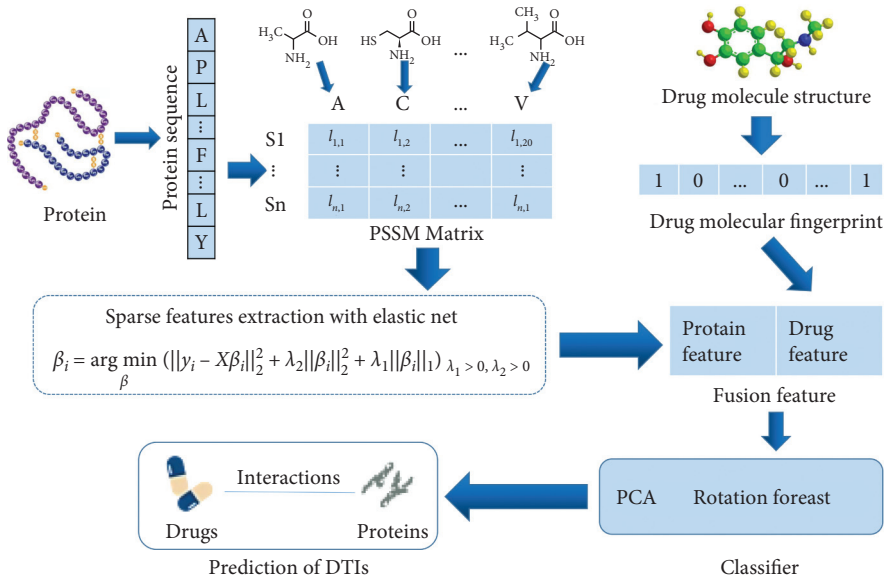

Figure 1: Flow chart for detecting DTIs by the novel model.

Table 1: The statistics of drugs, target proteins, and interacting drug-target pairs on the benchmark datasets.

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

amino acids. When using PSI-BLAST, the parameter e and iteration are, respectively, set as 0.001 and 3 to get wide and high homologous sequences.

2.4. Sparse Feature Extraction with Elastic Net. In order to improve the accuracy of the classifier, it is necessary to extract the most obvious features from the original data. Zou et al. [33] proposed a sparse principal component analysis method based on elastic net. Firstly, principal component analysis was performed on PSSM; then the principal component coefficients were sparsely processed by elastic net. Suppose that PSSM is $X = [x_1, x_2, \ldots, x_p]_n$, n is the length of the protein sequence, and p is the number of amino acids. After centralizing the matrix, this method performs singular value decomposition (SVD) in X as follows:

$$X = U\ DV^T, \tag{2}$$

where $D = \begin{pmatrix} D_1 & 0 \\ 0 & 0 \end{pmatrix}$, $D_1 = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r)$, $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r \geq 0$, and $r = \mathrm{rank}(X)$. Then we got the following equations:

$$XV = \left[x_1, x_2, \ldots, x_p\right] \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1p} \\ v_{21} & v_{22} & \cdots & v_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ v_{p1} & v_{p2} & \cdots & v_{pp} \end{bmatrix},$$

$$= \left[\sum_{j=1}^{p} v_{j1}X_j, \sum_{j=1}^{p} v_{j2}X_j, \cdots, \sum_{j=1}^{p} v_{jp}X_j\right] = U\ D,$$

$$\sum_{j=1}^{p} v_{ji}X_j = \frac{\sigma_i^2}{n-1}, \ (i = 1, 2, \cdots p). \tag{3}$$

We got σ_i from the above equations and then made $y_i = \sigma_i u_i$, $(i = 1, 2, \cdots, p)$ as the principal component to establish the ridge regression model. The model adds $L2$ norm to the least-square model, which can shrink the regression coefficient. Supposing that $Y = [y_1, y_2, \cdots, y_p]$ and $\hat{Y} = X\beta$, the ridge regression model is as follows:

$$\hat{\beta}_{\mathrm{ridge}} = \arg\min_{\beta}\left(\|Y - X\beta\|_2^2 + \lambda\|\beta\|_2^2\right). \tag{4}$$

Estimation coefficients of ridge regression are still not sparse in the ridge regression model. Adding $L1$ norm constraint to ridge regression model, we get elastic net model to reduce some coefficients to 0. Therefore, the elastic net model without scaling factor is used as the sparse approximation of principal component analysis as

$$\hat{\beta} = \arg\min_{\beta}\left(\|Y - X\beta\|_2^2 + \lambda_2\|\beta\|_2^2 + \lambda_1\|\beta\|_1\right), \quad (5)$$

where $\lambda_1, \lambda_2 (\lambda_1 > 0, \lambda_2 > 0)$ are regular coefficients. The normalized elastic net estimation is used as the approximate value of the principal component coefficient:

$$\frac{\hat{\beta}}{\|\hat{\beta}\|} = v_i. \quad (6)$$

Finally, we replace PSSM with the principal component coefficients for further process of feature fusion and classification. In this paper, the PSSM of each target protein sequence would be described by 20 feature vectors, when using the elastic net based sparse feature descriptor.

2.5. Rotation Forest (RF). Rotation forest is a kind of supervised learning algorithm, which is an improvement of the early integrated forest [34]. It has high accuracy of classification for small- and medium-scale datasets. When executing RF classifier, the dataset is randomly divided into different sample subsets of K by selecting disjoint features. Then bootstrap algorithm and sparse component analysis (PCA) are utilized to generate sparse rotation matrices based on subsets. Lastly, train each base classifier by utilizing the matrices, and vote to give the result of RF classifier by counting the prediction of different base decision tree classifiers.

Let $X = (x_1, x_2, x_3, \cdots, x_n)^T$ be the training feature set taking the form of an $n \times m$matrix to carry out the m features of n samples, and let $Y = (y_1, y_2, y_3, \cdots, y_n)^T$ be the label matrix of $1 \times n$. The base classifiers of RF are represented as $D_1, D_2, D_3, \cdots, D_L$, respectively. The training steps of base classifiers are as follows:

(I) The sample set M is randomly divided into K sample subsets; each subset contains $C = m/K$ features.

(II) Suppose that $M_{i,j}$ denotes the sample subset, and $X_{i,j}$ denotes the feature set of $M_{i,j}$. Perform bootstrap method to rebuild a new training feature set $X_{i,j}^{'}$ on 75% of the original feature set $X_{i,j}$.

(III) Perform principal component analysis (PCA) on the set $X_{i,j}^{'}$. When the index of feature is j, we got the principal component coefficients $a_{i,j}^{(1)}, a_{i,j}^{(2)}, a_{i,j}^{(3)}, \cdots, a_{i,j}^{(C_j)}$.

(IV) The principal component coefficients are put into the sparse rotation matrix R_i as follows:

$$R_i = \begin{bmatrix} a_{i,1}^{(1)}, a_{i,1}^{(2)}, \cdots a_{i,1}^{(C_1)} & 0 & \cdots & 0 \\ 0 & a_{i,2}^{(1)}, a_{i,2}^{(2)}, \cdots a_{i,2}^{(C_2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{i,K}^{(1)}, a_{i,K}^{(2)}, \cdots a_{i,K}^{(C_k)} \end{bmatrix}. \quad (7)$$

Base classifier D_i gives that the possible result of test sample x is $d_{i,j}(xR_i^a)$. Then calculate the confidence degree to which x belongs to different categories as $\mu_j(x)$; the formula is as follows:

$$\mu_j(x) = \frac{1}{L}\sum_{i=1}^{L} d_{i,j}(xR_i^a). \quad (8)$$

Finally, the sample x is distributed into a most likely class.

3. Results and Discussion

3.1. Evaluation Criteria. In general, we used four evaluation criteria, accuracy (Acc.), sensitivity (Sen.), precision (Pre.), and Matthews correlation coefficient (MCC), to measure the effect of the prediction model on four datasets.

$$\begin{aligned} \text{Acc.} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}, \\ \text{Sen.} &= \frac{\text{TP}}{\text{TP} + \text{TN}}, \\ \text{Pre.} &= \frac{\text{TP}}{\text{FP} + \text{TP}}, \\ \text{MCC} &= \frac{\text{TN} \times \text{TP} - \text{FN} \times \text{FP}}{\sqrt{(\text{TN} + \text{FN}) \times (\text{TP} \times \text{FP}) \times (\text{TN} + \text{FP}) \times (\text{TP} \times \text{FN})}}, \end{aligned} \quad (9)$$

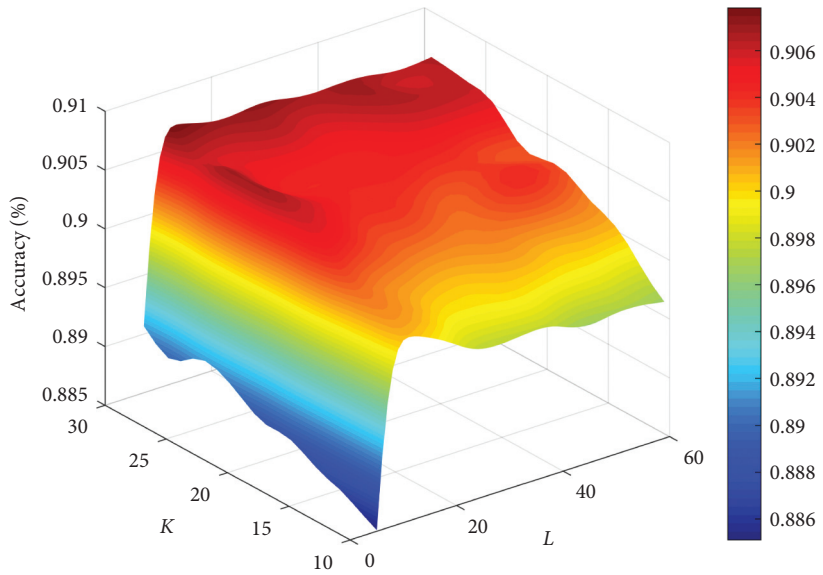

Figure 2: Influence of parameters K and L on the proposed model.

Table 2: 5-fold CV results on enzyme dataset obtained by the proposed model.

Test set	Acc. (%)	Pre. (%)	Sen. (%)	MCC (%)	AUC (%)
1	91.45	92.78	90.30	82.94	96.42
2	88.38	91.36	84.81	76.95	94.36
3	92.05	91.93	91.77	84.09	96.89
4	90.51	91.71	88.68	81.05	96.05
5	89.23	89.30	89.61	78.45	95.93
Average	90.32 ± 1.52	91.41 ± 1.29	89.03 ± 2.62	80.70 ± 2.99	95.93 ± 1.10

where TP (true positive) is the aggregate of true samples predicted correctly; TN (true negative) is the aggregate of true samples predicted incorrectly; FP (false negative) represents the number of false samples predicted incorrectly; FN (false negative) represents the number of false samples predicted correctly. We employ receiver operating characteristic (ROC) [35] curves to image results and count the area under the curve (AUC) [36] to quantify the performance of model.

3.2. Parameter Discussion. The numbers of decision trees K and random subsets L in RF are relevant to the performance of the proposed model. We adopt grid search method [37] to optimize the suitable parameters. Figure 2 shows that the accuracy rapidly increases in the beginning and tends to be flat as the K-value increases. On the other hand, we can see that the accuracy increases when the L-value is between 0 and 20 and then gradually descends in a fluctuation. Considering the cost of time and calculation, the K-value and L-value of RF are, respectively, set as 22 and 20. Figure 2 shows the influence generated by different parameters of RF on the proposed model.

3.3. Fivefold Cross-Validation Results on Four Datasets. Fivefold CV was applied in evaluating the prediction ability of model on four gold standard datasets. This method is a distinct sampling verification where each sample has only one chance to be selected into the testing dataset. Specifically, the whole dataset is equally divided into five sections, one of which is regarded as the testing data and the others are treated as the training data. During the validation, the model keeps $K = 22$ and $L = 20$, which are generated as Figure 2 by considering the influence on the proposed model. Tables 2–5 give the validation results of four datasets.

After the application of fivefold cross validation on enzyme dataset, the average accuracy, sensitivity, precision, MCC, and AUC come to be 90.32%, 91.41%, 89.03%, 80.70%, and 95.93% with standard deviations of 1.52%, 1.29%, 2.62%, 2.99%, and 1.10%, respectively. On the ion channel dataset, the average values of accuracy, sensitivity, precision, MCC, and AUC reached 88.91%, 88.37%, 89.61%, 77.82%, and 94.86%, respectively, and the average standard deviations are 0.49%, 1.47%, 1.13%, 0.97%, and 0.17%. With regard to GPCRs dataset, the average values of accuracy, sensitivity, precision, MCC, and AUC reached 80.65%, 80.89%, 79.93%, 61.07%, and 88.16%, with standard deviations of 1.65%, 2.79%, 3.25%, 3.45%, and 0.94%, respectively. On nuclear receptor dataset, the average values of accuracy, sensitivity, precision, MCC, and AUC reached 79.73%, 78.56%, 80.99%, 60.60%, and 85.72% with standard deviations of 7.90%, 9.74%, 12.21%, 16.70%, and 4.65%, respectively. Figures 3–6 show the ROC curves of the four datasets generated by 5-fold CV.

Table 3: 5-fold CV results on ion channel dataset obtained by the proposed model.

Test set	Acc. (%)	Pre. (%)	Sen. (%)	MCC (%)	AUC (%)
1	88.64	87.86	88.17	77.23	94.94
2	89.32	90.17	88.67	78.66	94.88
3	88.81	88.85	90.54	77.48	94.77
4	88.31	86.18	90.66	76.72	95.07
5	89.49	88.78	90.01	78.99	94.62
Average	88.91 ± 0.49	88.37 ± 1.47	89.61 ± 1.13	77.82 ± 0.97	94.86 ± 0.17

Table 4: 5-fold CV results on GPCRs dataset obtained by the proposed model.

Test set	Acc. (%)	Pre. (%)	Sen. (%)	MCC (%)	AUC (%)
1	80.71	80.74	82.58	61.34	88.70
2	83.46	84.80	82.17	66.97	89.45
3	79.53	76.92	81.97	59.21	87.04
4	79.44	81.30	77.52	58.98	87.99
5	80.11	80.70	75.41	58.87	87.59
Average	80.65 ± 1.65	80.89 ± 2.79	79.93 ± 3.25	61.07 ± 3.45	88.16 ± 0.94

Table 5: 5-fold CV results on nuclear receptor dataset obtained by the proposed model.

Test set	Acc. (%)	Pre. (%)	Sen. (%)	MCC (%)	AUC (%)
1	86.11	80.95	94.44	73.25	88.27
2	88.89	93.32	82.35	78.06	90.87
3	72.22	68.75	65.95	43.75	80.78
4	71.42	78.95	71.43	42.15	80.78
5	80.01	70.83	90.81	65.79	87.91
Average	79.73 ± 7.90	78.56 ± 9.74	80.99 ± 12.21	60.60 ± 16.70	85.72 ± 4.65

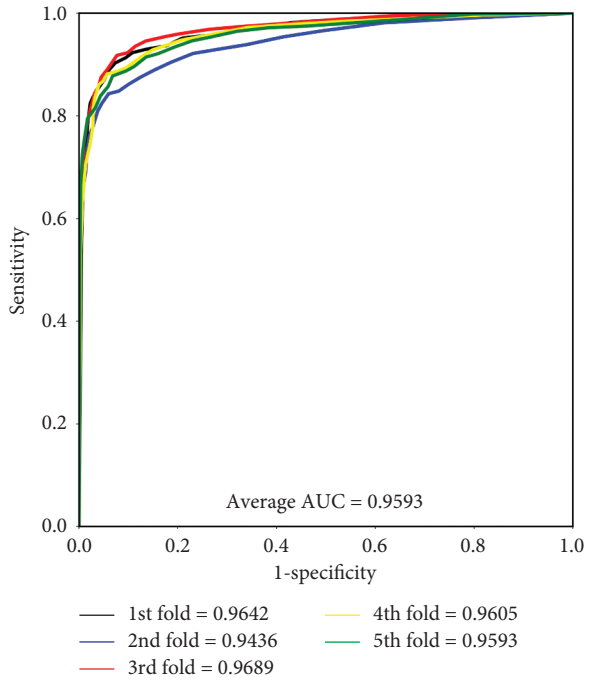

Figure 3: ROC curves are received by 5-fold CV on enzyme dataset.

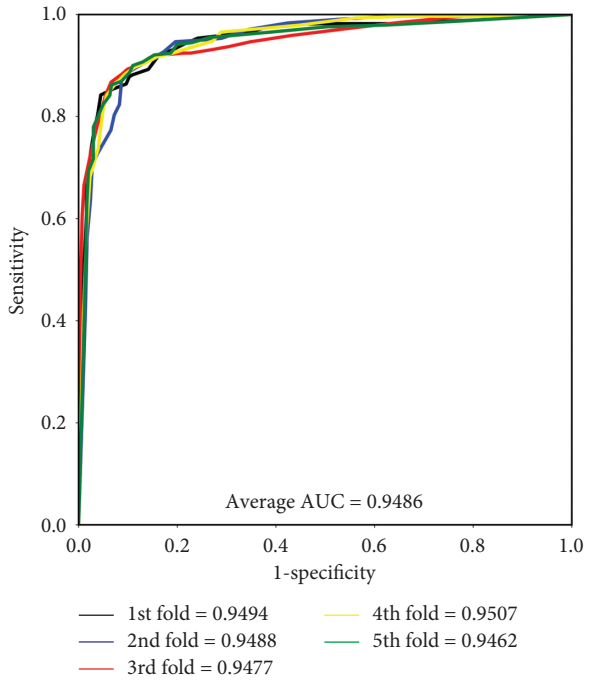

Figure 4: ROC curves are received by 5-fold CV on ion channel dataset.

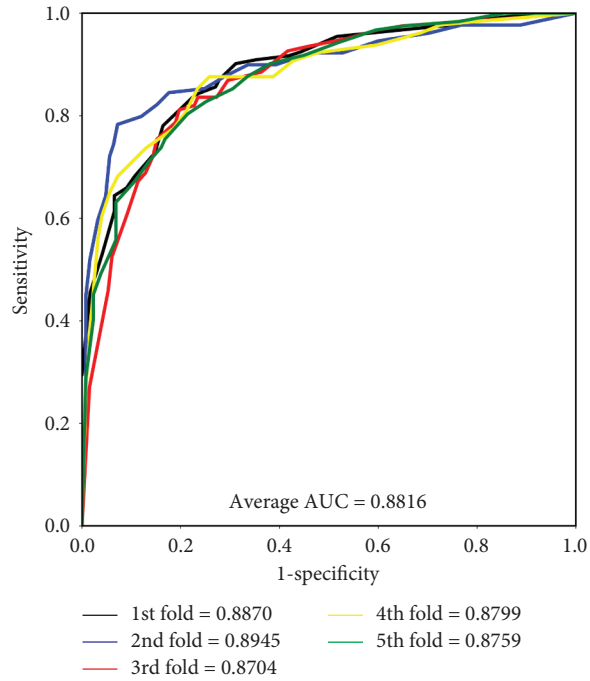

Figure 5: ROC curves are received by 5-fold CV on GPCRs dataset.

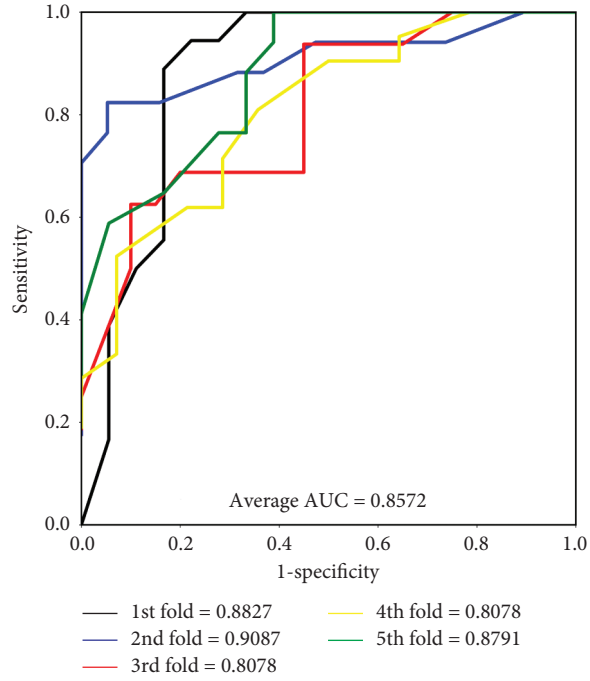

Figure 6: ROC curves are received by 5-fold CV on nuclear receptor dataset.

3.4. Comparison with Support Vector Machine Classifier. Many traditional classifiers can effectively predict DTIs. In this section, we combined SVM [38] with elastic net based sparse feature extraction (ESFE) as a compared model for further assessing the performances of the prediction model. When building the support vector machine classifier, the Gaussian kernel is adopted in SVM to predict DTIs. Meanwhile, the parameter c also affects the performance of the model. In addition, we utilized grid search method to tune and monitor the established model and set the best parameter c as 0.1. Table 6 illustrates the results predicted by the comparison model on enzyme dataset. The average accuracy, sensitivity, precision, MCC, and AUC are 86.44%, 84.64%, 89.02%, 72.99%, and 91.62%, respectively, with

Table 6: 5-fold cross-validation results on enzyme dataset by RF and SVM classifiers.

Test set	Acc. (%)	Pre. (%)	Sen. (%)	MCC (%)	AUC (%)
PSSM + ESFE + RF					
1	91.45	92.78	90.30	82.94	96.42
2	88.38	91.36	84.81	76.95	94.36
3	92.05	91.93	91.77	84.09	96.89
4	90.51	91.71	88.68	81.05	96.05
5	89.23	89.30	89.61	78.45	95.93
Average	90.32 ± 1.52	91.41 ± 1.29	89.03 ± 2.62	80.70 ± 2.99	95.93 ± 1.10
PSSM + ESFE + SVM					
1	86.92	84.26	90.02	74.03	92.70
2	87.52	86.56	89.72	75.04	91.82
3	85.98	82.90	89.86	72.25	91.41
4	87.01	85.89	89.33	74.04	92.11
5	84.78	83.58	86.18	69.61	90.06
Average	86.44 ± 1.08	84.64 ± 1.54	89.02 ± 1.61	72.99 ± 2.14	91.62 ± 0.99

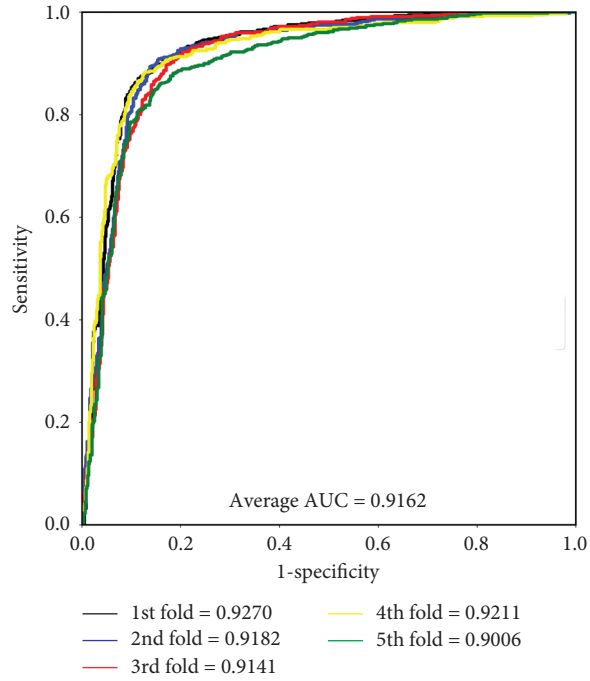

Figure 7: ROC curves yielded by the SVM classifier on enzyme dataset.

Table 7: The average AUC generated by our method, MSCMF, NetLapRLS, KBMF2K, and Yamanishi on benchmark datasets.

Dataset	Our method	MSCMF	NetLapRLS	KBMF2K	Yamanishi
Enzyme	**0.9593**	0.9142	0.9013	0.832	0.821
Ion channel	**0.9486**	0.9054	0.9165	0.799	0.692
GPCRs	**0.8816**	0.8363	0.7701	0.849	0.811
Nuclear receptor	**0.8572**	0.6867	0.6772	0.824	0.814

standard deviations of 1.08%, 1.54%, 1.61%, 2.14%, and 0.99%, respectively. The accuracy of RF classifier has improvements of 1.08%, 1.54%, 1.61%, 2.14%, and 0.99%, respectively, compared with the SVM classifier. The improvements indicate that the RF is more stable and effective than support vector machine.

Figure 7 gives the ROC curves of the compared model on enzyme dataset. The comparison of multiple criteria shows that the RF classifier has a better performance compared with SVM classifier.

3.5. Comparison with Other Methods. Thus far, many models are presented for identifying DTIs. In this part, we tested four models, MSCMF [14], NetLapRLS [15], KBMF2K [39], and Yamanishi [40], to better assess the

prediction ability of our model. These methods also adopt fivefold CV. Table 7 provides the average AUC of these models performed on enzyme, ion channel, GPCRs, and nuclear receptor datasets. In this table, the AUC values of these methods are less than our method to varying degrees. The average AUC has growth of 0.0922, 0.1204, 0.0650, and 0.1067, respectively. The results illustrate that the proposed model has a significant improvement in detecting DTIs.

4. Conclusions

In this paper, we proposed a novel computational model combining position-specific scoring matrix (PSSM), elastic net based sparse feature extraction, and rotation forest classifier to identify drug-target interactions. The fivefold CV method comprehensively assessed the prediction ability of the proposed model on the datasets. Our model achieves average accuracies of 90.32%, 88.91%, 80.65%, and 79.73% on such datasets as enzyme, ion channel, GPCRs, and nuclear receptor. In addition, we conduct support vector machine (SVM) and other previous models on the same datasets. The results illustrate that our model can stably and precisely predict DTIs. Although we perform many experiments, this paper still has some limitations. On the one hand, the sparse feature descriptors represent the local information of PSSM, and the overall information is hardly to get. On the other hand, the grid search method was utilized on rotation forest to keep the most import information, and it shows that the differences of features obtained from the feature extraction need to be improved. The future research will focus on finding more suitable feature extraction methods and better classifiers to optimize the model. Furthermore, we will study the influence of noise on the results to improve the accuracy and feasibility of the proposed model.

Data Availability

The data are original, and the data source is restricted.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Authors' Contributions

Zheng-Yang Zhao conceived the algorithm and software; Jie Pan prepared the datasets; Shan-Wen Zhang carried out the experiments; Yu-A Huang analyzed the experiments; Chang-Qing Yu visualized the results and wrote the draft of the paper; Wen-Zhun Huang administrated the project and supervised the funding.

Acknowledgments

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

References

[1] G. Emilien, M. Ponchon, C. Caldas et al., "Impact of genomics on drug discovery and clinical medicine," *Qjm*, vol. 93, no. 7, pp. 391–423, 2000.

[2] H. Zeng and X. Wu, "Alzheimer's disease drug development based on computer aided drug design," *European Journal of Medicinal Chemistry*, vol. 121, pp. 851–863, 2017.

[3] D. S. Wishart, "Drugbank and its relevance to pharmacogenomics," *Pharmacogenomics*, vol. 9, no. 8, pp. 1155–1162, 2008.

[4] E. E. Bolton, Y. Wang, P. A. Thiessen, and S. H. Bryant, "PubChem: integrated platform of small molecules and biological activities," *Annual Reports in Computational Chemistry*, vol. 4, pp. 217–241, 2008.

[5] Y. Hong, Q. Chu, Y. Li, L. Tao et al., "Therapeutic target database update 2016: enriched resource for bench to clinical drug target and targeted pathway information," *Nucleic Acids Research*, vol. 4, no. 44, pp. 1069–1074, 2016.

[6] J. J. Irwin and B. K. Shoichet, "Zinc a free database of commercially available compounds for virtual screening - journal of chemical information and modeling," *American Chemical Society*, vol. 36, no. 16, pp. 177–182, 2005.

[7] A. J. Williams, S. Ekins, and V. Tkachenko, "Towards a gold standard: regarding quality in public domain chemistry databases and approaches to improving the situation," *Drug Discovery Today*, vol. 17, no. 13-14, pp. 685–701, 2012.

[8] C. Hansch, B. R. Telzer, and L. Zhang, "Comparative QSAR in toxicology: examples from teratology and cancer chemotherapy of aniline mustards," *Critical Reviews in Toxicology*, vol. 25, no. 1, pp. 67–89, 1995.

[9] A. Valencia, J. Prous, O. Mora, N. Sadrieh, and L. G. Valerio, "A novel QSAR model of Salmonella mutagenicity and its application in the safety assessment of drug impurities," *Toxicology and Applied Pharmacology*, vol. 273, no. 3, pp. 427–434, 2013.

[10] S. Kar, A. Harding, K. Roy et al., "QSAR with quantum topological molecular similarity indices: toxicity of aromatic aldehydes to Tetrahymena pyriformis," *SAR and QSAR in Environmental Research*, vol. 21, no. 1-2, pp. 149–168, 2010.

[11] W. Bao, C. Yuan, Y. Zhang et al., "Mutli-features prediction of protein translational modification sites," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 15, no. 5, pp. 1453–1460, 2018.

[12] H. Luo, T. Du, P. Zhou et al., "Molecular docking to identify associations between drugs and class I human leukocyte antigens for predicting idiosyncratic drug reactions," *Combinatorial Chemistry and High Throughput Screening*, vol. 18, no. 3, pp. 296–304, 2015.

[13] W. Bao, Z. Jiang, and D. S. Huang, "Novel human microbe-disease association prediction using network consistency projection," *BMC Bioinformatics*, vol. 18, p. 543, 2017.

[14] W. Bao, B. Yang, D. S. Huang et al., "IMKPse: identification of protein malonylation sites by the key features into general PseAAC," *IEEE Access*, p. 1, 2019.

[15] 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, no. 2, pp. 197–206, 2007.

[16] S. Zhu, Y. Okuno, G. Tsujimoto et al., "A probabilistic model for mining implicit 'chemical compound-gene' relations from literature," *Bioinformatics*, vol. 21, pp. 245–251, 2005.

[17] Y. Pan, Y. Zhang, and J. Liu, "Text mining-based drug discovery in cutaneous squamous cell carcinoma," *Oncology Reports*, vol. 40, no. 6, pp. 3830–3842, 2018.

[18] L. Peska, K. Buza, J. Koller et al., "Drug-target interaction prediction: a Bayesian ranking approach," *Computer Methods and Programs in Biomedicine*, vol. 152, pp. 15–21, 2017.

[19] Y. Ding, "Identification of drug–target interactions via dual laplacian regularized least squares with multiple kernel fusion," *Knowledge-Based Systems*, vol. 204, Article ID 106254, 2020.

[20] J. Y. Shi, A. Q. Zhang, S. W. Zhang, K. T. Mao, and S. M. Yiu, "A unified solution for different scenarios of predicting drug-target interactions via triple matrix factorization," *BMC Systems Biology*, vol. 12, no. S9, p. 136, 2019.

[21] X. Zheng, H. Ding, H. Mamitsuka et al., "Collaborative matrix factorization with multiple similarities for predicting drug-target interactions," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1025–1033, Chicago, IL, USA, August 2013.

[22] J. Keum and H. Nam, "S. E. L. F.-B. L. M.: Prediction of drug-target interactions via self-training SVM," *PLoS One*, vol. 12, no. 2, Article ID e0171839, 2017.

[23] L. Wang, Z.-H. You, X. Chen, X. Yan, G. Liu, and W. Zhang, "Rfdt: a rotation forest-based predictor for predicting drug-target interactions using drug structure and protein sequence information," *Current Protein and Peptide Science*, vol. 19, no. 5, pp. 445–454, 2018.

[24] H. Nikolai, A. Jessica, V. E. Joachim et al., "SuperTarget goes quantitative: update on drug–target interactions," *Nucleic Acids Research*, vol. 40, no. D1, pp. D1113–D1117, 2012.

[25] I. Schomburg, "BRENDA, enzyme data and metabolic information," *Nucleic Acids Research*, vol. 30, no. 1, pp. 47–49, 2002.

[26] M. Kanehisa, "From genomics to chemical genomics: new developments in KEGG," *Nucleic Acids Research*, vol. 34, no. 90001, pp. D354–D357, 2006.

[27] M. He and M. Nourani, "Drug-target interaction prediction using semi-bipartite graph model and deep learning," *BMC Bioinformatics*, vol. 21, no. S4, 2020.

[28] J. Shen, F. Cheng, Y. Xu, W. Li, and Y. Tang, "Estimation of ADME properties with substructure pattern recognition," *Journal of Chemical Information and Modeling*, vol. 50, no. 6, pp. 1034–1041, 2010.

[29] W. Bao, Y. Chen, and D. Wang, "Prediction of protein structure classes with flexible neural tree," *Bio-Medical Materials and Engineering*, vol. 24, no. 6, pp. 3797–3806, 2014.

[30] W. Bao, D. Wang, and Y. Chen, "Classification of protein structure classes on flexible neutral tree," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 14, no. 5, 2017.

[31] M. Gribskov, A. D. McLachlan, and D. Eisenberg, "Profile analysis: detection of distantly related proteins," *Proceedings of the National Academy of Sciences*, vol. 84, no. 13, pp. 4355–4358, 1987.

[32] S. F. Altschul and E. V. Koonin, "Iterated profile searches with PSI-BLAST-a tool for discovery in protein databases," *Trends in Biochemical Sciences*, vol. 23, no. 11, pp. 444–447, 1998.

[33] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *Journal of Computational and Graphical Statistics*, vol. 15, no. 2, pp. 265–286, 2006.

[34] J. J. Rodriguez, L. I. Kuncheva, and C. J. Alonso, "Rotation forest: a new classifier ensemble method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 10, pp. 1619–1630, 2006.

[35] M. H. Zweig and G. Campbell, "Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine," *Clinical Chemistry*, vol. 39, no. 4, pp. 561–577, 1993.

[36] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[37] J. A. A. Brito, F. E. Mcneill, C. E. Webber, and D. R. Chettle, "Grid search: an innovative method for the estimation of the rates of lead exchange between body compartments," *Journal of Environmental Monitoring*, vol. 7, no. 3, pp. 241–247, 2005.

[38] C. Saunders, M. O. Stitson, J. Weston et al., "Support vector machine," *Computer Science*, vol. 1, no. 4, pp. 1–28, 2002.

[39] M. Gönen, "Predicting drug-target interactions from chemical and genomic kernels using Bayesian matrix factorization," *Bioinformatics*, vol. 28, no. 18, pp. 2304–2310, 2012.

[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, no. 13, pp. i232–i240, 2008.