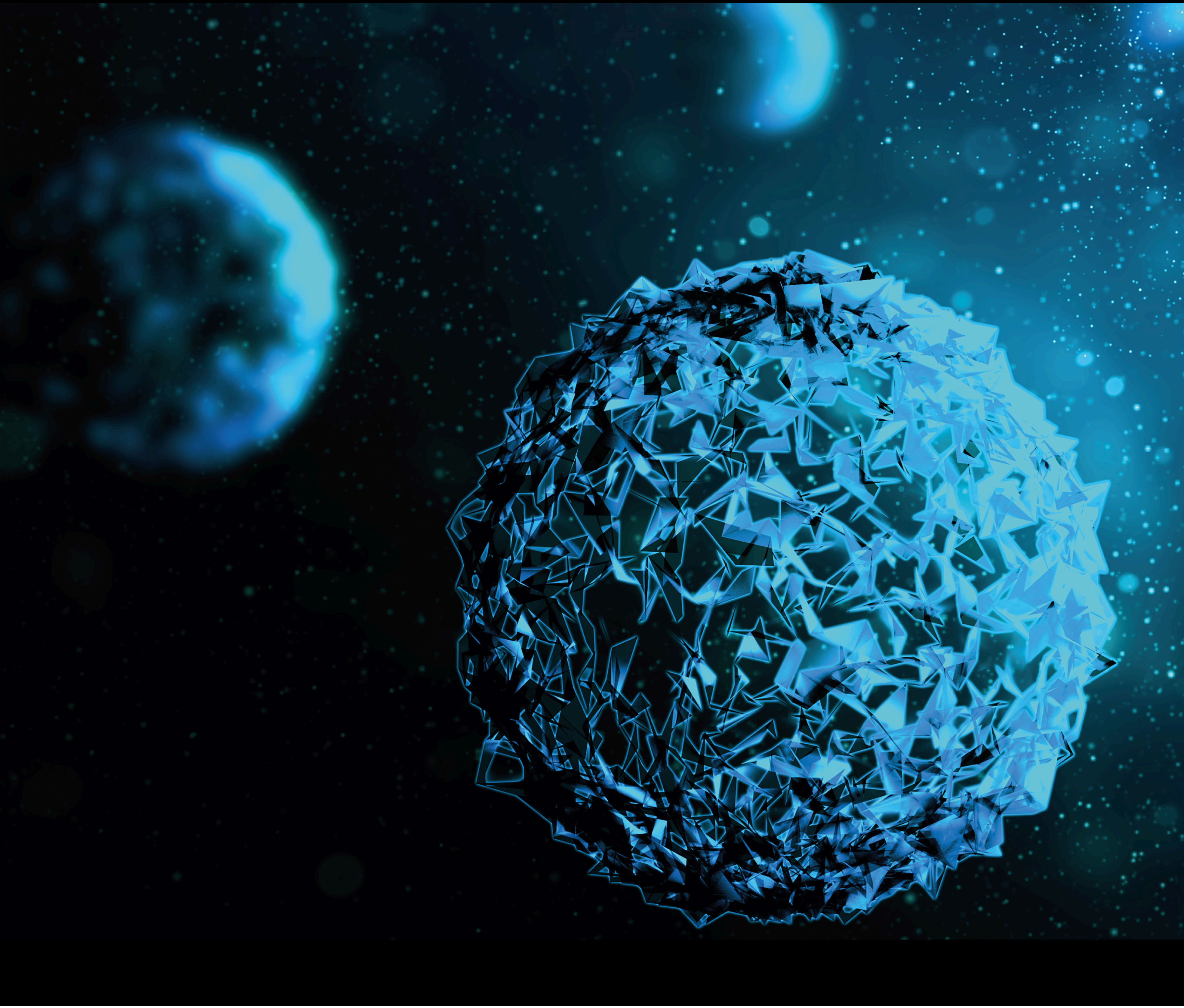
BioMed Research International

Scalable Machine Learning Algorithms in Computational Biology and Biomedicine 2021

Lead Guest Editor: Quan Zou
Guest Editors: Dariusz Mrozek, Qin Ma, and Yungang Xu



Scalable Machine Learning Algorithms in Computational Biology and Biomedicine 2021

BioMed Research International

Scalable Machine Learning Algorithms in Computational Biology and Biomedicine 2021

Lead Guest Editor: Quan Zou
Guest Editors: Dariusz Mrozek, Qin Ma, and Yungang Xu

Section Editors

Academic Editors

Computational Biology

Bioinformatics

Contents

Hindawi
BioMed Research International
Volume 2022, Article ID 9044793, 7 pages
https://doi.org/10.1155/2022/9044793

Research Article

Application of DNA-Binding Protein Prediction Based on Graph Convolutional Network and Contact Map

Weizhong Lu,[1,2] Nan Zhou,[1] Yijie Ding,[1] Hongjie Wu,[1] Yu Zhang,[3] Qiming Fu,[1] and Haiou Li[2]

[1]*School of Electronic and Information Engineering, Suzhou University of Science and Technology, Suzhou, China*
[2]*Provincial Key Laboratory for Computer Information Processing Technology, Soochow University, Suzhou, China*
[3]*Suzhou Industrial Park Institute of Services Outsourcing, Suzhou, China*

Correspondence should be addressed to Yijie Ding; wuxi_dyj@163.com

Received 13 April 2021; Accepted 24 December 2021; Published 17 January 2022

Academic Editor: Khac-Minh Thai

DNA contains the genetic information for the synthesis of proteins and RNA, and it is an indispensable substance in living organisms. DNA-binding proteins are an enzyme, which can bind with DNA to produce complex proteins, and play an important role in the functions of a variety of biological molecules. With the continuous development of deep learning, the introduction of deep learning into DNA-binding proteins for prediction is conducive to improving the speed and accuracy of DNA-binding protein recognition. In this study, the features and structures of proteins were used to obtain their representations through graph convolutional networks. A protein prediction model based on graph convolutional network and contact map was proposed. The method had some advantages by testing various indexes of PDB14189 and PDB2272 on the benchmark dataset.

1. Introduction

The sequence of a protein determines its structure and different structures determine different functions. There is about 18% of the weight of protein in the human body. As the carrier of life, it plays a very important role in human production and life. As a major component of life, proteins are involved in almost all activities of cells, including DNA replication and transcription, chromatin formation, cell growth, and a series of activities, all of which cannot be separated by specific proteins [1]. These proteins that bind to and interact with DNA are called DNA-binding proteins. It has a strong affinity with single-stranded DNA, but a small affinity with double-stranded DNA. Therefore, DNA-binding proteins are also called helical instability proteins, single-stranded DNA-binding proteins [2].

With the development of gene sequencing, various sequencing studies have left many DNA and proteins, including DNA-binding proteins. Using machine learning and deep learning methods to predict DNA-binding proteins has reached a good level, but there is still room for improvement.

At present, many methods based on machine learning have emerged to distinguish DNA-binding proteins, which are divided into structure and sequence methods. Yubo et al. [3] proposed a DBD-Hunter method that combines structural comparison with an assessment of statistical potential to measure the interaction between DNA bases and protein residues. Zhou et al. [4] used random forest for classification by adopting amino acid preservation pattern, potential electrostatic, and other features. However, these methods are too dependent on the protein structure, so the practical operation is difficult. Therefore, sequence-based studies were carried out. Liu et al. [5] proposed a new method for predicting DNA-binding proteins, IDNA-Pro, by integrating features into pseudoamino acids from protein sequences and classifying them through random forest. Zhao et al. [6] classified DNA-binding proteins based on the physicochemical properties of amino acids by using random forest to recognize the sequence features generated by PseAcc. Although the method based on machine learning can identify DNA-binding proteins well, it needs a lot of human intervention in the process of feature selection and could not properly grasp the relationship between data and features. To overcome this difficulty, deep learning techniques were introduced into protein prediction. Loo et al. [7] proposed a new prediction method MsDBP, which input the fused multiscale features into a deep neural network for learning and classification. The classification was tested with 67% accuracy on a separate dataset PDB2272. Compared with machine learning method, it can save the necessary manual intervention, but the prediction result needs to be improved.

Although there are many methods used to predict DNA-binding proteins at present, the results still have room for improvement. The main problem is how to obtain the high-precision protein structure from the protein sequence, because the accuracy of protein structure and feature has a great impact on the prediction results. In addition, the graph convolution network (GCN) has been widely used in the research of bioinformatics. Graph composed of nodes and edges serves as the input of the network without any requirements on size and format [8]. In order to improve the accuracy of structure and prediction, combining with the current developing trend of the technology of deep learning, a DNA-binding protein prediction model based on GCN and contact map was proposed. The protein graph depends on the sequence of the results of the comparison, so first introducing the preprocess of the dataset, including sequence comparison and filtering; the part of the output is used to calculate the features, and the other part as the input of Pconsc4 model [9], which is used to predict protein contact map, so the inputs of the model are feature matrix and adjacency matrix. We use them for training and prediction. The experimental results show that the prediction performance of DNA-binding proteins can be obtained by the method described. The research content of this paper is shown in Figure 1.

2. Materials and Methods

The prediction of DNA-binding proteins is divided into three parts: data preprocessing, training model, and testing. GCN differs from neural networks in that it introduces a graph structure to represent proteins, which can better represent the structure of proteins. The main purpose of protein sequence preprocessing is to obtain the features and structures of proteins. For the protein processing, the contact map is obtained by predicting the sequence through Pconsc4, and its output exactly corresponds to the adjacency matrix of GCN [10].

2.1. The Dataset. The DNA-binding protein dataset selected is the internationally common dataset. PDB14189 and PDB2272 were established by Gomes et al. [11]. Among them, the PDB14189 dataset was divided into 7129 DNA-binding protein sequences and 7060 DNA-unbinding protein sequences, and the PDB2272 dataset was divided into 1153 DNA-binding proteins and 1119 nonbinding proteins. PDB14189 was taken as the training set and PDB2272 as the test set. The dataset is detailed in Table 1 below. Among them, positive represents DNA-binding proteins, while negative represents non-DNA-binding proteins.

2.2. Protein Representation. The representation of proteins is generally divided into spatial structure and feature. The long-chain stable structure of protein also contains hydrogen bonds, hydrophobic bonds, salt bonds, and so on [12]. Each protein contains lots of atoms, if each atom is viewed as a node, then the protein graph will be very large, which will increase the pressure of training and is not easy to achieve. However, there are about hundreds of residues in a protein, and there is no other spatial information between residues, so it is more suitable to be used as nodes to represent structural features. The spatial structure of a protein can be represented by a contact map; it represents the two-dimensional structure of the protein; each element in the matrix represents the probability of contact at the corresponding position [13]; the value is between 0 and 1. Figure 2 shows a protein contact map.

Predicting the structure of a protein from its sequence is the purpose of introducing contact map. Specifically, assuming that the length of protein sequence is M, the size of its contact map is $M * M$. $M(i, j)$ represents the probability of contact between the ith residue and the jth residue. If the value is less than the threshold value, it can be considered that they are in contact. Pconsc4 is a fast and efficient method to predict contact map. Since its output is a probability value between 0 and 1, the threshold value of 0.5 was set for the obtained contact maps, and the probability value greater than or equal to 0.5 was set as 1.The rest were set as 0, so that the structural information of the protein could be well extracted, corresponding to the adjacency matrix as the input GCN network [14].

The next step is the extraction of protein features. Since residues are used as nodes, the properties of residues are selected as features. Due to the differences in the R group, different features are displayed, including aromaticity, polarity, and explicit valence [15]. Position-specific scoring matrix (PSSM) is a commonly used representation of protein features, in which the results of each element depend on the results of sequence comparison, and these results represent the feature of proteins [16]. Other features were also used, such as the primary thermal coding of the remaining symbols, whether the residue was aromatic, whether the residue was acidic charged, and whether it was extremely neutral, etc. [17], as shown in Table 2. In summary, the total number of features is 54, so the protein's feature matrix dimension is $(M, 54)$.

For PSSM, the basic position frequency matrix (PFM) [18] is calculated by the number of occurrence of residues at each position in the sequence of sequence alignment results. Equation (1) is as follows:

$$M_{k,j}^{\mathrm{PFM}} = \sum_{i=1}^{N} I\left(A_{i,j} = k\right), \tag{1}$$

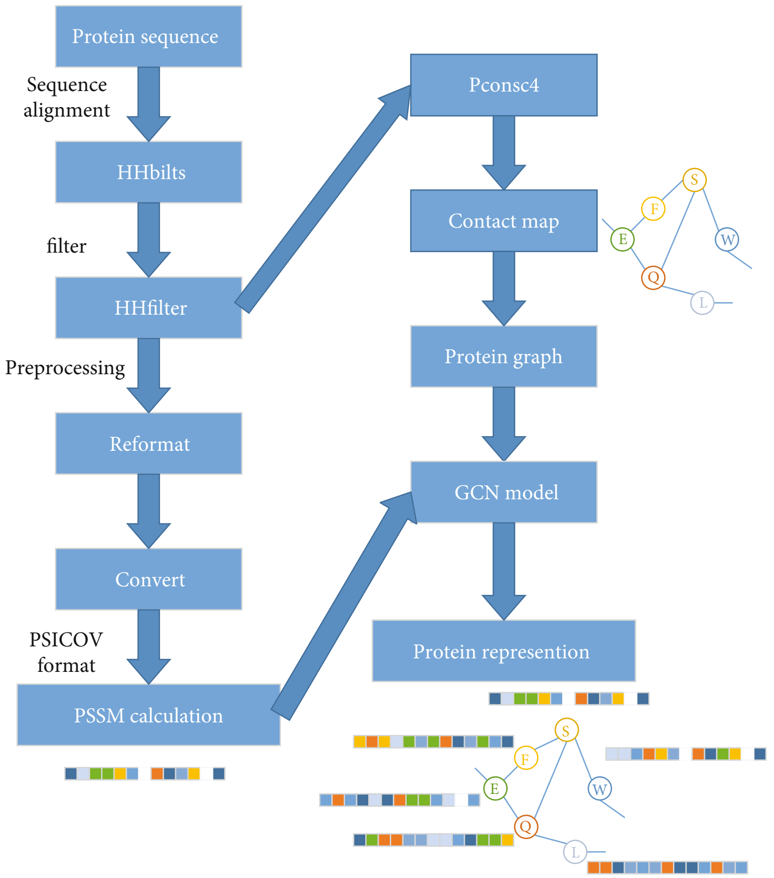


Figure 1: The processing of proteins, including the preprocessing of sequence, the generation of graph structures, and feature extraction, Pconsc4 was used to extract protein structural information. Finally, protein graph was generated higher-level feature graph through GCN.

Table 1: Introduction to the dataset.

Number\dataset	PDB14189	PDB2272
Positive	7129	1153
Negative	7060	1119
Total	14189	2272

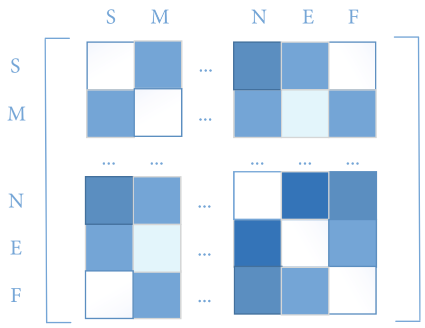


Figure 2: The contact map of protein.

where A represents a set of alignment sequences equal to the target protein length, k is the set of residues, $i = (1, 2 \cdots, N)$, $j = (1, 2, \cdots L)$, and $i(x)$ is the indicator function when the condition is met or not. Equation (2) is used to obtain the position probability matrix (PPM):

$$M_{k,j}^{\mathrm{PPM}} = \frac{M_{k,j}^{\mathrm{PFM}} + (p/4)}{N + p}. \tag{2}$$

In order to prevent the matrix entries from appearing 0, according to human experience, the pseudocount [19] p was set 0.8, so that PPM was regarded as a part of the node features.

2.3. Model Architecture.

2.3. Model Architecture. Although traditional convolution techniques perform well for Euclidean data, they perform poorly for non-Euclidean data [20]. Therefore, graph convolution technology came into being. For a graph, the edges of each node are related to other nodes and this information can be used to capture interdependencies between instances, so the node can aggregate its own features and its neighbor features to generate a new representation of the node [21]. With the continuous development of graph learning, there are many variations, like GAT, GAE, and GGN [22]. All these network models can extract the feature; for using the GCN layer, each layer convolution operation is as shown in Equation (3):

$$H^{l+1} = f\left(H^l, A\right) = \sigma\left(D\wedge^{-1/2}\hat{A}D\wedge^{-1/2}H^l W^{l+1}\right). \tag{3}$$

Among them, A is the adjacency matrix of node features,

Table 2: Node features.

Label	Feature	Size
1	One-hot encoding of the residue symbol	21
2	Position-specific scoring matrix (PSSM)	21
3	Whether the residue is aliphatic	1
4	Whether the residue is aromatic	1
5	Whether the residue is polar neutral	1
6	Whether the residue is acidic charged	1
7	Whether the residue is basic charged	1
8	Residue weight	1
9	The negative of the logarithm of the dissociation constant for the –COOH group	1
10	The negative of the logarithm of the dissociation constant for the –NH3 group	1
11	The negative of the logarithm of the dissociation constant for any other group in the molecule	1
12	The pH at the isoelectric point	1
13	Hydrophobicity of residue (pH = 2)	1
14	Hydrophobicity of residue (pH = 7)	1
	Total	54

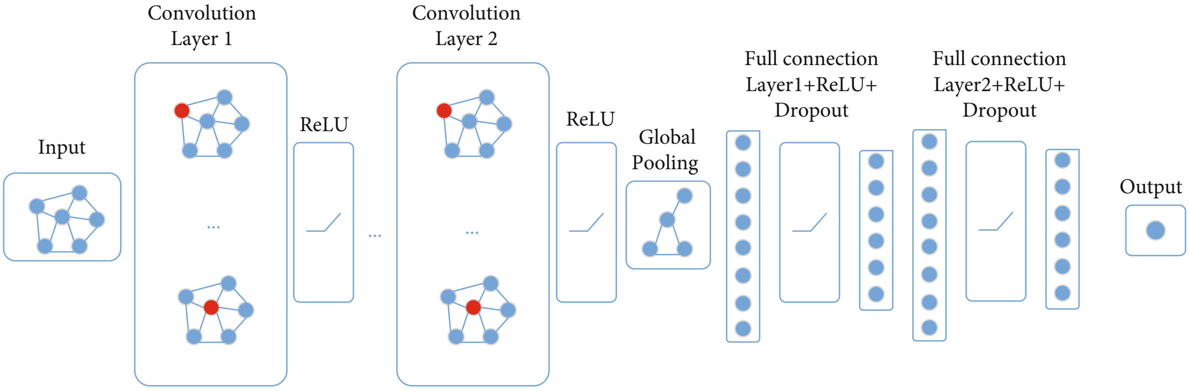

Input
Convolution Layer 1
ReLU
Convolution Layer 2
ReLU
Global Pooling
Full connection Layer1+ReLU+ Dropout
Full connection Layer2+ReLU+ Dropout
Output

Figure 3: The structure of the GCN network, graphs of DNA-binding proteins through the GCN to get their representation.

assuming that the node number is m, then its adjacency matrix is (m, m), $\widehat{D}$ is the degree of matrix (m, m), which represents the connection relationship between residues, $\widehat{D} = D + I$, I is a unit matrix, considers itself features, W^{l+1} is the first $l + 1$ layer of weighting matrix, H^l is the output of the first layer of l, and $H^0 = X$, X is the input of the feature matrix, Figure 3 shows the architecture of the model.

The protein graph contained much information about the interactions and positions of each residue pair, which was important for feature learning and predicting DNA-binding proteins. It was input into the GCN to extract the features. After convolution of multiple GCN layers, the representation of protein was effectively extracted. Then, the overall features of protein for prediction were obtained. The prediction includes two full connection layers. The results were presented as probabilities.

Using GCN to map proteins to the representation of rich features has also become a method of protein feature extraction. In addition, there were many factors affecting the experimental results, such as dropout, epoch, and batch.

Table 3: The hyperparameter settings using human experience.

Hyperparameter	Setting
Epoch	1000
Batch size	128
Learning rate	0.001
Optimizer	Adam
The number of convolution layers	3
Fully connected layers after GCN	2

Table 4: Combinations of GCN models on PDB14189.

Model	Number of layers	Layer1 (in, out)	Layer2 (in, out)	Layer3 (in, out)
GCN	1	(54,54)	—	—
GCN	2	(54,54)	(54,108)	—
GCN	3	(54,54)	(54,108)	(108,216)

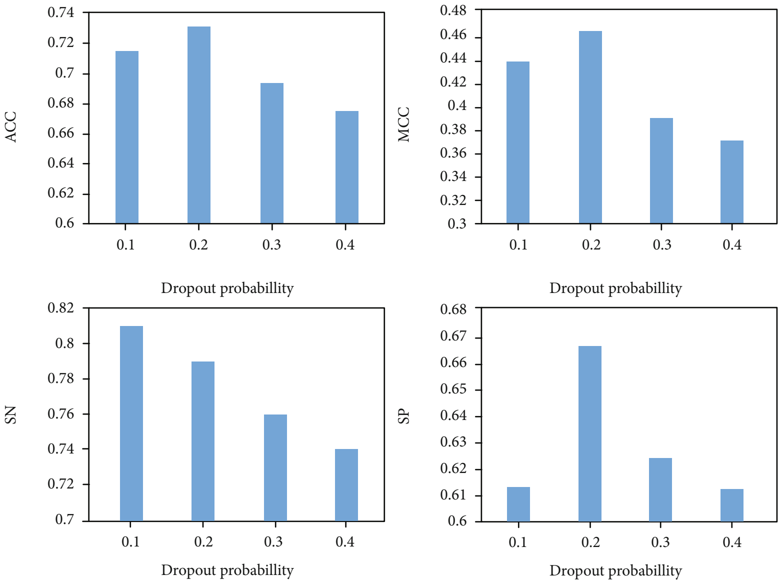


FIGURE 4: Comparison of prediction performance of different dropout probabilities.

The setting of some hyperparameters were compared and determined through experiments.

3. Results and Discussion

The experiment was built on PyTorch [23], an open source deep learning framework. The GCN model was based on its PyG implementation [24], PDB14189 was used for testing to find the optimal super parameters, and PDB2272 was used to test model performance.

3.1. The Evaluation Index. Accuracy (ACC), Matthews correlation coefficient (MCC), sensitivity (SN), and specificity (SP) were used as the evaluation indexes of the model [25], these indexes were widely used in the studies of biological sequences, as shown in

$$\begin{cases} \mathrm{SN} = \dfrac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \\ \mathrm{SP} = \dfrac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}, \\ \mathrm{ACC} = \dfrac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}}, \\ \mathrm{MCC} = \dfrac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FN})\times(\mathrm{TN}+\mathrm{FP})\times(\mathrm{TP}+\mathrm{FP})\times(\mathrm{TN}+\mathrm{FN})}}. \end{cases} \quad (4)$$

Among them, TP is the number of the correctly predicted positive samples, TN is the number of the correctly predicted negative samples, FP is the number of the wrongly predicted positive samples, and FN is the number of the wrongly predicted negative samples. SN represents the percentage of correctly predicted positive samples, SP represents the percentage of correctly predicted negative samples, ACC represents the percentage of correctly predicted samples in total samples, and MCC represents the prediction quality of the binary classification model, with a range of $[-1, 1]$. The larger the MCC is, the better the prediction quality of the model is.

3.2. The Setting of Hyperparameters. Training an optimal model requires constantly adjusting the hyperparameters of the model, which can be modified based on human experience. Some of the hyperparameters were shown in Table 3. In this model, according to human experience, the GCN layer was set to three, dimensions of input and output for each layer were shown in Table 4. Some other parameters were compared in the following experiences.

3.3. Model Performance when Selecting Different Dropouts. After protein feature extraction, in order to better improve the accuracy of classification, two full connection layers were added to the ends to improve the learning ability of the model. In the fully connected layer, in order to avoid overfitting of the model, dropout was introduced to shut down some neurons with a probability value. Different probability values will affect the performance of prediction. To evaluate the impact of different dropout values, Figure 4 shows the performance of the model according to different dropout values. When the dropout is 0.2, the model has the highest performance compared to other parameters.

3.4. Whether PSSM Is Included in Feature Selection. The selection of protein feature greatly affects the accuracy of prediction. Since the dimension of PSSM matrix constructed by features was very small, the experiment was carried out with PSSM or without PSSM. Figure 4 shows the results of various indicators under the condition. PSSM depends on the sequence correlation results, which contains much evolutionary information about the sequence, and ultimately determines the protein features. As can be seen from Figure 5, PSSM can effectively represent the features of proteins and effectively improve the prediction performance.

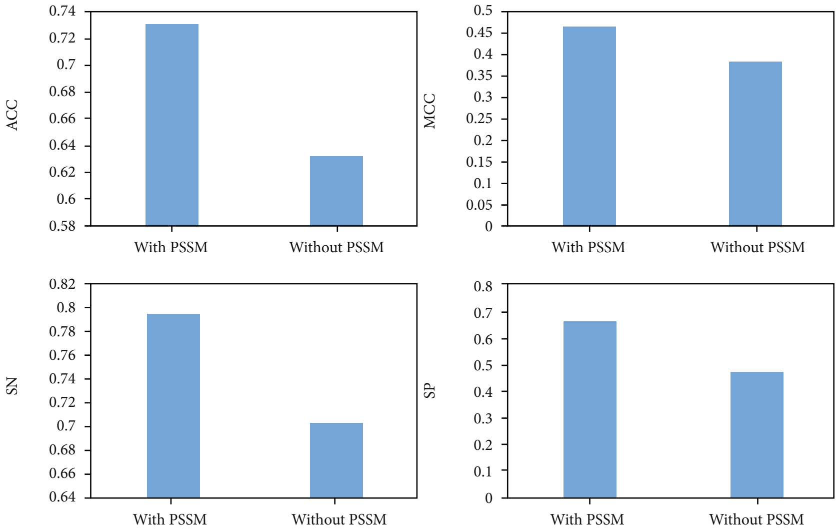


Figure 5: Comparison of performance results with or without PSSM.

Table 5: Comparison between the proposed method and existing methods on PDB2272.

Methods	ACC (%)	MCC (%)	SN (%)	SP (%)
Qu et al. [26]	48.33	3.34	48.31	48.35
Local-DPP [27]	50.57	4.56	8.76	93.66
Pse-DNA-Pro [28]	61.88	24.30	75.28	48.08
DPP-Pse-AAC [29]	58.10	16.25	56.63	59.61
Ms-DBP [30]	66.99	33.97	70.69	63.18
GCN-method	78.49	59.27	92.59	64.15

3.5. Analysis of Experimental Results. In the independent test dataset, PDB14189 was used as the training dataset to train the model, and PDB2272 was used as the test dataset. According to the optimal experimental parameters, the final DNA-binding protein classification model was constructed: the number of GCN layers were three, dropout was 0.2, PSSM was selected as the feature, the input and output dimensions of each layer were (54, 54), (54,108), and (108,216). Other methods were compared with the method, and the method reached ACC (78.49%), SN (92.59%), SP (64.15%), and MCC (59.27%). Under certain conditions, the method has certain advantages compared with the existing methods, as shown in Table 5.

4. Conclusions

DNA-binding proteins are enzymes, which can bind with DNA to produce complex proteins and play important roles in the functions of a variety of biological molecules. In order to improve the accuracy of prediction of DNA-binding protein, a DNA-binding protein prediction model based on GCN and contact map was proposed. In this model, the dataset was preprocessed by sequence alignment; then, the structural information is extracted by Pconsc4 model; PSSM and some biological characteristics are used as features. Finally, the GCN model was constructed to train and predict DNA-binding protein data. The protein graph contained information about the interactions and positions of each residue pair, which was important for feature learning and predicting binding proteins. The protein graph was input into the GCN to extract the features, and the prediction included two full connection layers. Using GCN to map proteins to the representation of rich features has also become a method of protein feature extraction. Through training and parameter tuning, the performance of GCN model was better than some existing methods. It also provides some thoughts for other fields of biological information.

In the future, we plan to carry out a research on feature extraction and network model to improve the accuracy of DNA-binding proteins and related prediction. Different biological features can be combined, and methods such as attention mechanism can be considered to improve the model, in order to achieve the goal of improving the prediction effect and other indicators.

Data Availability

The datasets can be found in the references.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This paper is supported by the National Natural Science Foundation of China (61902272, 62073231, 61772357, 62176175, 61876217, and 61902271), National Research Project (2020YFC2006602), and Provincial Key Laboratory for Computer Information Processing Technology, Soochow University (KJS2166).

References

[1] A. S. Rifaioglu, H. Atas, M. J. Martin, R. Cetin-Atalay, V. Atalay, and T. Doğan, “Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases,” *Briefings in Bioinformatics*, vol. 20, no. 5, pp. 1878–1912, 2019.

[2] M. S. Nogueira and O. Koch, “The development of target-specific machine learning models as scoring functions for docking-based target prediction,” *Journal of Chemical Information and Modeling*, 2019.

[3] Y. Wang, Y. Ding, F. Guo, L. Wei, and J. Tang, “Improved detection of DNA-binding proteins via compression technology on PSSM information,” *PLoS One*, vol. 12, no. 9, 2017.

[4] L. Zhou, X. Song, D. J. Yu, and J. Sun, “Sequence-based detection of DNA-binding proteins using Multiple-view features allied with feature selection,” *Molecular Informatics*, vol. 39, no. 8, p. 2000006, 2020.

[5] K. Liu, X. Sun, L. Jia et al., “Chemi-net: a Molecular graph convolutional network for accurate drug property prediction,” *International Journal of Molecular Sciences*, vol. 20, no. 14, p. 3389, 2019.

[6] H. Zhang, Q. Zhang, F. Ju et al., “Correction to: Predicting protein inter-residue contacts using composite likelihood maximization and deep learning,” *BMC Bioinformatics*, vol. 20, no. 1, p. 616, 2019.

[7] J. Loo, A. L. Emtage, L. Murali, S. S. Lee, A. L. W. Kueh, and S. P. H. Alexander, “Ligand discrimination during virtual screening of the CB1 cannabinoid receptor crystal structures following cross-docking and microsecond molecular dynamics simulations,” *RSC Advances*, vol. 9, no. 28, pp. 15949–15956.

[8] M. Michel, D. Menéndez Hurtado, and A. Elofsson, “PconsC4: fast, accurate, and hassle-free contact predictions,” *Bioinformatics (Oxford, England)*, vol. 35, no. 15, pp. 2677–2679, 2019.

[9] L. Jiang, S. Wang, B. Zhang et al., “"A more probable explanation" is still impossible to explain GN-z11-flash: in response to Steinhardt et al. (arXiv:2101.12738),” 2021, https://arxiv.org/abs/2102.01239.

[10] S. Wang, S. Sun, Z. Li, R. Zhang, and J. Xu, “Accurate de novo prediction of protein contact map by ultra-deep learning model,” *PLoS Computational Biology*, vol. 13, no. 1, article e1005324, 2017.

[11] J. Gomes, B. Ramsundar, E. N. Feinberg, and V. S. Pande, “Atomic convolutional networks for predicting protein-ligand binding affinity,” https://arxiv.org/abs/1703.10603, 2017.

[12] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud et al., “Automatic chemical design using a data-driven continuous representation of molecules,” *ACS Central Science*, vol. 4, no. 2, pp. 268–276, 2018.

[13] E. B. Lenselink, N. Ten Dijke, B. Bongers et al., “Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set,” *Journal of Cheminformatics*, vol. 9, no. 1, p. 45, 2017.

[14] V. Le, T. P. Quinn, T. Tran, and S. Venkatesh, “Deep in the bowel: highly interpretable neural encoder-decoder networks predict gut metabolites from gut microbiome,” *BMC Genomics*, vol. 21, no. S4, 2020.

[15] Z. Hakime, Z. Arzucan, and O. Elif, “DeepDTA: deep drug-target binding affinity prediction,” *Bioinformatics*, vol. 17, p. 17, 2018.

[16] M. Sun, S. Zhao, C. Gilvary, O. Elemento, J. Zhou, and F. Wang, “Graph convolutional networks for computational drug development and discovery,” *Briefings in Bioinformatics*, vol. 21, no. 3, pp. 919–935, 2020.

[17] T. Wen and R. B. Altman, “Graph convolutional neural networks for predicting drug-target interactions,” *Journal of Chemical Information and Modeling*, vol. 59, no. 10, pp. 4131–4149, 2019.

[18] T. Nguyen, H. Le, and S. Venkatesh, “GraphDTA: prediction of drug-target binding affinity using graph convolutional networks,” *BioRxiv*, vol. 2019, p. 684662, 2019.

[19] K. Nishida, M. C. Frith, and K. Nakai, “Pseudocounts for transcription factor binding sites,” *Nucleic Acids Research*, vol. 37, no. 3, pp. 939–944, 2009.

[20] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, “How powerful are graph neural networks?,” 2018, https://arxiv.org/abs/1810.00826.

[21] C. Shen, Y. Ding, J. Tang, J. Song, and F. Guo, “Identification of DNA–protein binding sites through multi-scale local average blocks on sequence information,” *Molecules*, vol. 22, no. 12, p. 2079, 2017.

[22] J. Hanson, T. Litfin, K. Paliwal, and Y. Zhou, “Identifying molecular recognition features in intrinsically disordered regions of proteins by transfer learning,” *Bioinformatics*, vol. 36, no. 4, 2019.

[23] A. Paszke, S. Gross, S. Chintala et al., *Automatic differentiation in PyTorch*, 2017.

[24] S. Akbar, S. Khan, F. Ali, M. Hayat, M. Qasim, and S. Gul, “iHBP-DeepPSSM: identifying hormone binding proteins using PsePSSM based evolutionary features and deep learning approach,” *Chemometrics and Intelligent Laboratory Systems*, vol. 204, article 104103, 2020.

[25] T. Song, S. Wang, D. Liu et al., “SE-OnionNet: a convolution neural network for protein–ligand binding affinity prediction,” *Frontiers in Genetics*, vol. 11, article 607824, 2021.

[26] Y. Qu, J. A. Fitzgerald, H. Rauter, and N. Farrell, “Approaches to selective DNA binding in polyfunctional dinuclear platinum chemistry. The synthesis of a trifunctional compound and its interaction with the mononucleotide 5'-guanosine monophosphate,” *Inorganic Chemistry*, vol. 40, no. 24, pp. 6324–6327, 2001.

[27] L. Wei, J. Tang, and Q. Zou, “Local-DPP: an improved DNA-binding protein prediction method by exploring local evolutionary information,” *Information Sciences*, vol. 384, pp. 135–144, 2017.

[28] B. Liu, J. Xu, S. Fan, R. Xu, J. Zhou, and X. Wang, “PseDNA-pro: DNA-binding protein identification by combining Chou's PseAAC and physicochemical distance transformation,” *Molecular Informatics*, vol. 34, no. 1, pp. 8–17, 2015.

[29] Y. D. Khan, M. Jamil, W. Hussain, N. Rasool, S. A. Khan, and K. C. Chou, “pSSbond-PseAAC: prediction of disulfide bonding sites by integration of PseAAC and statistical moments,” *Journal of Theoretical Biology*, vol. 463, pp. 47–55, 2019.

[30] X. du, Y. Diao, H. Liu, and S. Li, “MsDBP: exploring DNA-binding proteins by integrating Multiscale sequence information via Chou's Five-Step rule,” *Journal of Proteome Research*, vol. 18, no. 8, pp. 3119–3132, 2019.

Hindawi
BioMed Research International
Volume 2021, Article ID 9995073, 12 pages
https://doi.org/10.1155/2021/9995073

Research Article

Studying the Effect of Taking Statins before Infection in the Severity Reduction of COVID-19 with Machine Learning

Alireza Davoudi,[1,2] Mohsen Ahmadi,[3] Abbas Sharifi,[4] Roshina Hassantabar,[1,2] Narges Najafi,[1,2] Atefeh Tayebi,[1,2] Hamideh Abbaspour Kasgari,[5] Fatemeh Ahmadi,[1,4] and Marzieh Rabiee[1,2]

[1] *Department of Infectious Diseases, Mazandaran University of Medical Sciences, P.O. Box: 48175-866, Sari, Iran*
[2] *Antimicrobial Resistance Research Center, Communicable Diseases Institute, Mazandaran University of Medical Sciences, Sari, Iran*
[3] *Department of Industrial Engineering, Urmia University of Technology (UUT), P.O. Box: 57166-419, Urmia, Iran*
[4] *Department of Mechanical Engineering, Urmia University of Technology (UUT), P.O. Box: 57166-419, Urmia, Iran*
[5] *Department of Clinical Pharmacy, Mazandaran University of Medical Science, P.O. Box: 48175-866, Sari, Iran*

Correspondence should be addressed to Abbas Sharifi; abbas.sharifi@mee.uut.ac.ir

Received 13 March 2021; Revised 25 April 2021; Accepted 27 May 2021; Published 21 June 2021

Academic Editor: Quan Zou

Statins can help COVID-19 patients' treatment because of their involvement in angiotensin-converting enzyme-2. The main objective of this study is to evaluate the impact of statins on COVID-19 severity for people who have been taking statins before COVID-19 infection. The examined research patients include people that had taken three types of statins consisting of Atorvastatin, Simvastatin, and Rosuvastatin. The case study includes 561 patients admitted to the Razi Hospital in Ghaemshahr, Iran, during February and March 2020. The illness severity was encoded based on the respiratory rate, oxygen saturation, systolic pressure, and diastolic pressure in five categories: mild, medium, severe, critical, and death. Since 69.23% of participants were in mild severity condition, the results showed the positive effect of Simvastatin on COVID-19 severity for people that take Simvastatin before being infected by the COVID-19 virus. Also, systolic pressure for this case study is 137.31, which is higher than that of the total patients. Another result of this study is that Simvastatin takers have an average of 95.77 mmHg O_2Sat; however, the O_2Sat is 92.42, which is medium severity for evaluating the entire case study. In the rest of this paper, we used machine learning approaches to diagnose COVID-19 patients' severity based on clinical features. Results indicated that the decision tree method could predict patients' illness severity with 87.9% accuracy. Other methods, including the K-nearest neighbors (KNN) algorithm, support vector machine (SVM), Naïve Bayes classifier, and discriminant analysis, showed accuracy levels of 80%, 68.8%, 61.1%, and 85.1%, respectively.

1. Introduction

In late December 2019, a previously unidentified coronavirus, currently named the 2019 novel β-coronavirus, emerged from Wuhan, China, the provincial capital of Hubei Province. The virus was later named severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [1]. The World Health Organization (WHO) first declared the coronavirus disease (named COVID-19) as an international public health emergency and then as a pandemic [2]. The disease's incubation period is from 2 to 14 (average 4 to 7) days [1], and its initial manifestations are related to viremia. The clinical manifestations of COVID-19, which appear after an incubation period of around 5-6 days, are associated with the release of cytokines and cytokine storm syndrome in severe cases. The clinical spectrum of the disease varies from asymptomatic or mild (in more than 80%) to severe cases, which lead to acute respiratory syndrome, respiratory failure, and death. Clinical features of the disease include fever, coughing, fatigue, sweating, myalgia, sore throat, dry mouth, dry cough, shortness of breath, chest pain, hemoptysis, abdominal pain, nausea, and diarrhea [3]. According to the disease onset, the essential radiographic manifestations include scattered subpleural ground glass lesions, crazy paving lesions, and consolidation [1]. The definitive diagnosis of the disease is made by virus detection through RT-PCR. For this purpose, a sample of the pharyngeal swab, nasopharynx or oropharynx, and a sample of tracheal secretions are needed [3]. The most critical laboratory evidence of COVID-19 patients includes lymphocytopenia and increased CRP. Also, the most important risk factors include old age; diabetes; high blood pressure; chronic heart, lung, liver, and kidney diseases; cardiovascular disease; immunodeficiency; and cancers [1]. Severity criteria of the disease are $SpO_2 < 93\%$ on room air at sea level, a respiratory rate > 30 breaths/min, $PaO_2/FiO_2 < 300$ mmHg, or lung infiltrates > 50% within 48 h [3]. Oxygen therapy, using a nasal cannula or a high-flow oxygen device, should be administered immediately. So far, there has been no conclusive evidence for the effectiveness of current antiviral therapies. In this regard, chloroquine or hydroxychloroquine and ritonavir/lopinavir (Kaletra) are used in most treatment protocols, and antiviral drugs, including Favipravir, Remdesivir, Arbidol, Sofosbuvir, and Ribavirin, are currently used in clinical trials [3].

Statins are inhibitors of the enzyme hydroxyl methylglutaryl coenzyme A (HMG-CoA reductase) and are responsible for accelerating the early stages of cholesterol biosynthesis. These compounds are multivalent cardioprotective drugs increasingly recognized as mediators with direct cellular effects beyond their cardiac role. Statins can block some downstream molecules such as farnesyl pyrophosphate (FPP) and geranylgeranyl pyrophosphate (GGPP), which play a vital role in infecting viruses like influenza. They have also been discussed in terms of intercellular, intracellular, inflammatory, and proinflammatory signals in some studies. Some research has reported their anti-inflammatory and immunomodulatory properties and upregulation for ACE2 receptors and statins [4]. It appears that lipid-lowering pharmacological interventions, in particular statins, might reduce the risk of cardiovascular complications caused by COVID-19 and might potentially have an additional antiviral activity. Several studies have shown that lipid rafts are involved in the life cycle of different viruses, including coronaviruses. Evidence of the cholesterol importance for viral entry into host cells suggests a role for cholesterol-lowering therapies in reducing viral infectivity. Statins have pleiotropic impacts, including anti-inflammatory, immunomodulatory, and antithrombotic activities, in addition to their lipid reduction and plaque stability effects. In some studies that examine statin therapy in influenza infection, lower mortality rates, and intubation, statin treatment demonstrated improved blood viral clearance throughout chronic hepatitis C infection. Statins also are used to monitor critical inhibitors of SARS-CoV-2 as a potentially SARS-CoV-2 protease [5]. Nevertheless, no proper antiviral treatment has been found for this disease so far, and all medications used are based on hypotheses that do not provide adequate evidence to support them. Due to the very high prevalence of the virus and its relatively high mortality rate, finding factors that can prevent or accelerate the onset or exacerbation of the disease and its complications can provide significant help in reducing the mortality of this disease in the current pandemic. Besides, it can be helpful for the treatment of subsequent possible seasonal epidemics such as influenza.

The present study investigates the effect of using standard doses of statins in the months before infection in patients with COVID-19 admitted to the Razi Hospital in Ghaemshahr, Iran, during February and March 2020, to reduce the severity of the disease and mortality rate of COVID-19. Overall, using statins may be a good guideline in the initial months of the COVID-19 epidemic.

2. Literature Review

Virani [6] conducted a review study to assess whether ongoing statin therapy enhances the overall cardiovascular outcomes of virally infected patients, like COVID-19. According to this paper, none of the studies reported adverse effects of this therapy. Fedson et al. [7] indicated the positive effects of statin adjuvant therapy in Sierra Leone in the 2014 outbreak of Ebola treatments. Zhang et al. [8] assessed the risk of entering COVID-19 with the decrease in ACE2 expression. A retrospective analysis was presented in 13,981 COVID-19 patients, including 1219 statins, in Hubei Province, China. Based on a mixed-effect Cox model, after the tendency match, the probability of all-cause death for 28 days was 5.2% and 9.4%, with an adjusted hazard ratio of 0.58 for both the matched statin and nonstatin classes. The lower mortality risks involved with statin use were recorded in Cox's time-varying method and marginal structural model study. The possible significant improvements of statins on COVID-19 patients were addressed by Rodrigues-Diez et al. [9]. Overall, they could target infected cell virus receivers, replications, degrading, and downstream reactions by discussing central and epidemiological proof. According to their results, statins might modulate virus entrance, acting on the SARS-CoV-2 receptors, ACE2 and CD147, and the involvement of lipid rafts. Besides, statins may control viral replication or degradation and have protective effects by inducing autophagic activation.

By closing multiple molecular pathways, including NF-κB and NLRP3 inflammasomes, the well-known anti-inflammatory effects of statins might restrict the cytokine storm in extreme COVID-19 patients associated with fatal outcomes. In conclusion, statin moderation of stimulation of coagulation reaction can also help boost the results of COVID-19. According to Castiglione et al. [10], statins are low-cost, widely tested, and well-tolerated medicines. These compounds are less likely to be affected due to health emergencies such as the ongoing COVID-19 pandemic, including in low-income countries, where therapy with costly medicines is not feasible. Adjuvant therapy and further treatment of preestablished statins might enhance the clinical success of COVID-19 patients through either immunomodulatory behavior or cardiovascular damage prevention. Subir et al. [11] have shown that statin can minimize the seriousness of lung injury and mortality from extreme acute respiratory syndrome-coronavirus 2 (SARS-CoV2) infections because of its immunomodulatory, anti-inflammatory, antithrombotic, and antioxidant properties. Upregulation of statin-induced angiotensin-converting enzyme-2 (ACE2) can also minimize lung damage due to excess angiotensin II. Statins can reduce viral entry into cells by disturbing lipid rafts. Daniels et al. [12] examined the relationship using statin/angiotensin-converting enzyme inhibitors/ARB in patients hospitalized for COVID-19 in the month before hospitalization. This study incorporated factors such as the risk of the severe result and time for the extreme outcome or disease treatment. They show that obesity and diabetes are potentially severe consequences of COVID-19.

Further, the predicted effects of the male sex consistently lead to an increased risk. A relationship was obtained between COVID status and obesity in COVID-negative and COVID-positive patients as a protective and risk factor. One new research in this regard shows that a shorter recovery period is associated with a younger generation. It may represent a more robust population and the fact that younger individuals subsequently present disease over time. While current smoking was more prominent in moderate rather than serious COVID-19, this cohort was questionable for its validity due to the very low prevalence of smoking (only eight current smokers have been found). Reiner et al. [13] suggested that statins could be effective inhibitors of SARS-CoV-2 M pro, based on binding energy from pitavastatin, Rosuvastatin, Lovastatin, and Fluvastatin. This claim is supported by the fact that certain statins (especially pitavastatin) have an even more considerable binding energy than protease or polymerase inhibitors.

3. Methods and Materials

3.1. Mechanism of Statin Action with COVID-19. The primary way of COVID-19 virus infection in body cells is ACE2, which downregulates this enzyme in the cells and lowers its protection properties. The virus triggers the response of the proinflammatory host based on MYD88, TLR, and NF-κB pathway activations. Statins are widely accessible, inexpensive, healthy, fat-reducing, and immunomodulatory medicines. These compounds prevent proinflammation of the MYD88-NF-μB and facilitate the upregulation of ACE2 in experimental models. Statins can be effective in the treatment of COVID-19 patients through these pathways. Statins also counteract hyperlipidemia triggered by some therapies commonly used in antiviral and immunosuppressive COVID-19 [10].

Like avian influenza viruses, by causing an extreme proinflammatory host reaction, beta-coronaviruses cause serious respiratory diseases. Some immunomodulatory treatments have proven to be successful in SARS, MERS, and COVID-19 cases. For instance, tocilizumab, an anti-interleukin-6 receptor humanized monoclonal antibody, was beneficial as maintenance care in selected patients with COVID-19 [14]. The interaction of SARS-CoV-1 with Toll-like receptors on the host cell membrane dramatically enhances the activity of the gene MYD88, whose output stimulates the occurrence of NF-κB-causing inflammatory processes [15]. In a murine model of SARS-CoV-1 infection, inhibition of NF-κB caused a reduction in lung infection and improved the survival rate of the disease [16]. Observational models suggest that statins stabilize MYD88 following a proinflammatory stimulus, including hypoxia [17]. Also, NF-κB activation was significantly decreased within 48 h in murine cells (relating to the plasma levels obtained with a healthy human dose of 40 mg [18]). Based on this information, the use of statins can be considered an immunomodulatory treatment in patients with COVID-19.

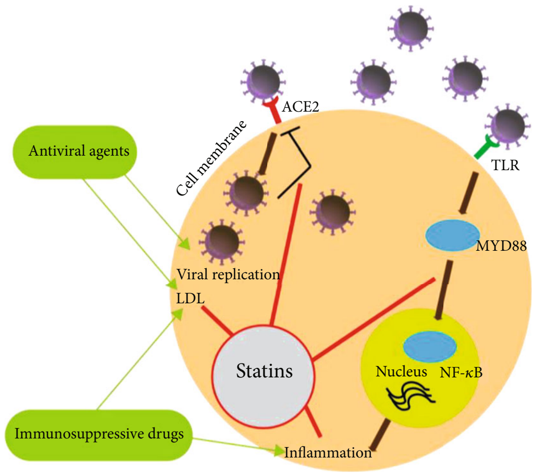


Figure 1: Statin action against COVID-19 virus.

Statins also interrupt the signaling of ACE2. After initial entry via ACE2, SARS-CoV-2 downregulates the expression of ACE2. As a result, it may foster original infiltration by innate immune cells and trigger an uncontested accumulation of angiotensin II, injuring the organ [14]. Both statins and ARBs are considered epigenetic modifications to regulate ACE2 (Figure 1) [7] experimentally. Regarding the improving effects of ACE2 on COVID-19 patients, there are currently activated RCTs with recombinant human ACE2 or ARBs1, and biological plausibility is also present in the study of statins [7].

3.2. Clinical Criteria and Variables. Indications for COVID-19 hospitalization are respiratory rate (RR) > 24, oxygen saturation (O_2Sat) < 93, significant lesion on CXR CT scan, pulmonary infiltration, and clinical judgment of a physician [9]. Criteria for severe disease include the number of breaths ≥ 30 times per minute, arterial oxygen saturation < 93 when the patient breathes in room air, and severe multifocal pulmonary involvement increases by more than 50% within 48 h [3].

4. Results and Discussion

4.1. Gathering Data. The present study investigates the effect of using standard doses of statins in the months before infection in patients with COVID-19 admitted to the Razi Hospital in Ghaemshahr (Mazandaran Province, Iran) during February and March 2020 in reducing the severity of the disease and mortality rate of COVID-19. The recorded variables and patients are illustrated in Table 1.

Table 1: Data descriptive analysis.

	N	Minimum	Maximum	Mean	Std. deviation	Unit
Age	561	1	3	1.88	0.821	Year group
<50 = 1						
50-65 = 2						
>65 = 3						
Gender	561	1	2	1.56	0.497	M/F
F = 1						
M = 2						
Duration of hospitalization	561	0	24	5.51	3.823	Days
Death	561	0	1	0.17	0.375	+/-
Diabetes	561	0	1	0.28	0.449	+/-
Hypertension	561	0	1	0.26	0.440	+/-
Heart failure	561	0	1	0.03	0.176	+/-
Chronic kidney disease	561	0	1	0.04	0.194	+/-
Chronic liver disease	561	0	1	0.01	0.119	+/-
History of transplantation	561	0	1	0.01	0.119	+/-
Solid-organ transplant = 1						
Hematological = 2						
Ischemic heart disease	561	0	1	0.13	0.335	+/-
Dyslipidemia	561	0	1	0.07	0.261	+/-
Thalassemia major	561	0	1	0.01	0.103	+/-
Allergic asthma	561	0	1	0.02	0.145	+/-
Hepatoid	561	0	1	0.06	0.232	+/-
History of radiotherapy	561	0	1	0.01	0.119	+/-
History of chemotherapy	561	0	1	0.01	0.103	+/-
Solid organs	561	0	1	0.00	0.060	+/-
Bone marrow	561	0	0	0.00	0.000	+/-
Steroid therapy	561	0	1	0.02	0.132	+/-
Steroid dosage	561	0	5	0.04	0.346	Mil
>20 mil = 1						
>20 mil = 2						
Prednisolone total > 300 = 3						
Contact history	561	0	1	0.02	0.132	+/-
Hemodialysis	561	0	1	0.02	0.151	+/-
Another underlying disease	561	0	1	0.18	0.386	+/-
Atorvastatin	560	0	1	0.17	0.379	+/-
Simvastatin	561	0	1	0.02	0.151	+/-
Rosuvastatin	561	0	1	0.01	0.073	+/-
Binary statin (statin or not)	561	0	1	0.18	0.388	+/-
History of addiction	561	0	1	0.01	0.084	+/-
Smokers	561	0	1	0.02	0.126	+/-
Fever	561	0	1	0.61	0.488	+/-
Chills	561	0	1	0.42	0.494	+/-
Rhinorrhea	561	0	1	0.08	0.266	+/-
Dry cough	561	0	1	0.49	0.500	+/-
Productive cough	561	0	1	0.16	0.362	+/-
Weakness	561	0	1	0.23	0.422	+/-
Anorexia	561	0	1	0.17	0.375	+/-

Table 1: Continued.

	N	Minimum	Maximum	Mean	Std. deviation	Unit
Sweating	561	0	1	0.08	0.269	+/-
Headaches	561	0	1	0.09	0.288	+/-
Myalgia	561	0	1	0.30	0.461	+/-
Loss of taste	561	0	1	0.06	0.245	+/-
Anosmia	561	0	1	0.00	0.042	+/-
Hematemesis	561	0	1	0.00	0.042	+/-
Diarrhea	561	0	1	0.09	0.283	+/-
Stomachache	561	0	1	0.03	0.181	+/-
Epigastric pain	561	0	1	0.01	0.103	+/-
Dizziness	561	0	1	0.02	0.156	+/-
Throat itching	561	0	1	0.01	0.073	+/-
Nausea	561	0	1	0.19	0.393	+/-
Vomiting	561	0	1	0.15	0.359	+/-
Shortness of breathing	561	0	1	0.13	0.341	+/-
Dyspnea	561	0	1	0.45	0.498	+/-
Tachypnea	561	0	1	0.12	0.325	+/-
Wheezing	561	0	1	0.00	0.060	+/-
Chest pain	561	0	1	0.04	0.190	+/-
Fatigue	561	0	1	0.06	0.232	+/-
Heart palpitations	561	0	1	0.00	0.042	+/-
Chest tightness	561	0	1	0.02	0.126	+/-
Sore throat	561	0	1	0.02	0.156	+/-
Temp	561	32.7	39.7	37.202	0.8102	˚C
Sys	561	70	220	119.22	22.879	mmHg
Dias	561	0	120	72.59	13.838	mmHg
RR	561	10	75	20.76	5.682	Br/min
HR	561	1.0	170.0	93.766	21.0701	BPM
O_2Sat on admission	561	30	100	92.42	8.117	mmHg
CT scan gloss opacity and increase in thickness between lobules or inside=1 Multiple alveolar consolidation = 2A Alveolar consolidation local = 2B Reversed halo = 3 Bronchovascular thickening in the lesion = 4 Tractional bronchiectasis = 5 Fiber tapes = 6 Acute respiratory distress syndrome = 7	561	0	7	0.32	0.920	0-7
Intensive cares 1 = primary hospitalization in ICU 2 = transfer from another part to ICU	561	0	2	0.32	0.686	0-2
Noninvasive ventilation Nasal O_2 = 1 Mask O_2 = 2 CPAP = 3 BIPAP = 4	561	0	4	0.30	0.763	0-4
Mechanical ventilation	561	0	2	0.14	0.350	0-2

Table 1: Continued.

	N	Minimum	Maximum	Mean	Std. deviation	Unit
Vasopressor						
Norepinephrine = 1						
Dopamine = 2	561	0	3	0.08	0.394	0-3
Dobutamine = 3						
Severity	561	1	5	2.48	1.592	1-5

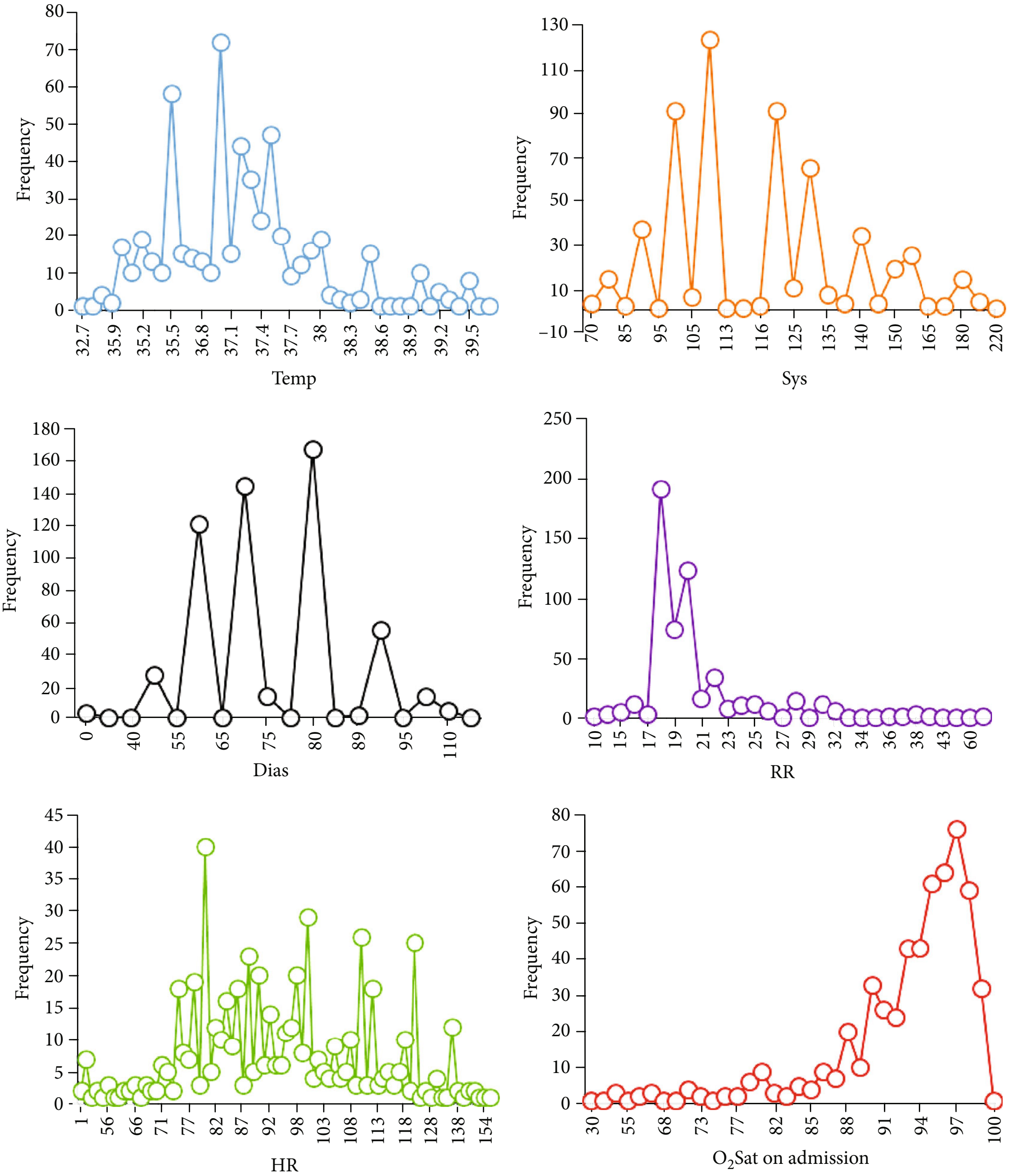


Figure 2: Frequency statistic vital signs of COVID-19 patients in admission.

This study investigates whether the severity of COVID-19 disease differs from patients who have previously taken statins due to hyperlipidemia or cardiovascular disease compared to patients who did not take statins before. In other words, the main objective is to explore if the history of taking statins has a positive effect on the COVID-19 disease process. It is of note that during the study, the patients did not use any statin during hospitalization. Table 1 shows the descriptive

Table 2: Description statistic of disease severity of the patients that take statins of Atorvastatin, Simvastatin, and Rosuvastatin.

			Severity					Total
			1	2	3	4	5	
Atorvastatin	0	Count	219	60	29	75	80	463
		% of total	47.30%	12.96%	6.26%	16.20%	17.28%	82.7%
	1	Count	36	12	9	25	15	**97**
		% of total	**37.1%**	**12.4%**	**9.3%**	**25.8%**	**15.4%**	**17.3%**
Simvastatin	0	Count	246	72	37	99	94	548
		% of total	44.89%	13.14%	6.75%	18.07%	17.15%	97.7%
	1	Count	9	1	1	1	1	**13**
		% of total	**69.23%**	**7.69%**	**7.69%**	**7.69%**	**7.69%**	**2.3%**
Rosuvastatin	0	Count	255	73	38	98	94	558
		% of total	45.70%	13.08%	6.81%	17.56%	16.85%	99.5%
	1	Count	0	0	0	2	1	**3**
		% of total	**0.00%**	**0.00%**	**0.00%**	**66.67%**	**33.33%**	**0.5%**
Total		Count	255	73	38	100	95	561
		% of total	45.5%	13.0%	6.8%	17.8%	16.9%	100.0%

Table 3: Results of Spearman correlation for the effects of statin history on COVID-19 severity.

		Atorvastatin	Simvastatin	Rosuvastatin	Severity
Atorvastatin	Correlation coefficient Sig. (1-tailed)	1.000	-0.071* (0.048)	-0.034 (0.214)	0.065 (0.063)
Simvastatin	Correlation coefficient Sig. (1-tailed)		1.000	-0.011 (0.395)	**-0.072* (0.044)**
Rosuvastatin	Correlation coefficient Sig. (1-tailed)			1.000	**0.080* (0.028)**
Severity	Correlation coefficient Sig. (1-tailed)				1.000

statistics of the patients who participated in this clinical research. The demographic data consist of age and gender, which were encoded to numerical values. Also, other criteria are past medical history; underlying diseases such as diabetes, hypertension, heart failure, chronic kidney disease (CKD), and chronic liver disease; history of transplantation; ischemic heart disease; dyslipidemia; thalassemia major; allergic asthma; hypothyroidism; history of radiotherapy; history of chemotherapy; solid organ involvement in cancer; bone marrow involvement; history of contact with COVID-19 patients; hemodialysis; and other underlying diseases such as favism, rheumatoid arthritis (RA), asthma, and stroke. Other variables are the history of steroid treatment, steroid dose, and history of addiction or smoking.

The mentioned features were encoded binary. These encoded features, along with other features, are presented in Table 1. Clinical signs for which the patient has referred to the hospital include fever, chills, rhinorrhea, dry cough, productive cough, weakness, anorexia, sweating, headaches, myalgia, loss of taste, anosmia, hematemesis, diarrhea, stomachache, epigastric pain, dizziness, throat itching, nausea, vomiting, shortness of breathing, dyspnea, tachypnea, wheezing, chest pain, fatigue, heart palpitations, chest tightness, and sore throat. Moreover, vital signs of the patient include body temperature (Temp), systolic pressure (Sys), diastolic pressure (Dias), respiratory rate (RR), heart rate (HR), and oxygen saturation (O_2Sat). Figure 2 shows the frequency statistics of vital signs of all COVID-19 patients participating in this research. Based on the results, some of the patients have fever temperatures between 36.4 and 37.7°C. Also, systolic pressure for all of the patients is between 90 and 130 mmHg. The respiratory rate for most patients is in the range of 17-21 Br/min, which is not in tachypnea condition. Regarding oxygen saturation as the essential factor of COVID-19 severity, its value is between 80 and 100 mmHg for most target patients.

4.2. Investigation of Effects of Statins on COVID-19 Severity. The most critical factor in this study is calculating the COVID-19 severity in numerical analysis. Based on the clinical sign of the patients, we encode the seriousness as follows:

$$\begin{cases} 1 \text{ else} & \text{Mild,} \\ 2\ 90\% < O_2\text{Sat} \leq 93\% & \text{Medium,} \\ 3\ 88\% < O_2\text{Sat} \leq 90\%,\ \text{RR} > 30 & \text{Severe,} \\ 4\ O_2\text{Sat} \leq 88\%,\ \text{Sys} < 90,\ \text{Dias} < 60\ \text{mmHg} & \text{Critical,} \\ 5 \text{ Death} & \text{Death.} \end{cases} \tag{1}$$

This study tried to evaluate the severity of all patients according to the history of statin taking of the patients,

Table 4: Study of patients taking Simvastatin and comparison with the total case study based on vital signs.

	Taking Simvastatin			All patients		
	N	Mean	Std. deviation	No.	Mean	Std. deviation
Temp	13	36.831	0.6033	561	37.202	0.8102
Sys	13	137.31	25.869	561	119.22	22.879
Dias	13	75.31	25.650	561	72.59	13.838
RR	13	21.92	6.538	561	20.76	5.682
HR	13	91.77	16.233	561	93.766	21.0701
O_2Sat	13	95.77	2.127	561	92.42	8.117
Severity	13	1.77	1.363	561	2.48	1.592

and the obtained results are described in Table 2. The examined research patients include people that had taken three types of statins consisting of Atorvastatin, Simvastatin, and Rosuvastatin. Of 561 patients, 17.3%, 2.3%, and 0.5% of them have used Atorvastatin, Simvastatin, and Rosuvastatin statins, respectively, for past disease treatment. Of all people who take Atorvastatin, 37.1% have mild COVID-19, and 25.8% have critical conditions. For Simvastatin, 69.23% of the patients have mild COVID-19. However, most people taking Rosuvastatin are in a critical situation.

One method for evaluating the relationship between statins' effects and severity is the statistical correlation test. Table 3 presents the correlation test findings based on Spearman's methods. The results show an indirect relationship between taking Simvastatin and severity. Regarding these findings, people who have taken Simvastatin are of lower severity than others. Moreover, most of them (69%) had mild severity. On the other hand, most patients who take Rosuvastatin are in critical condition. Furthermore, there is no significant relationship between Atorvastatin users and COVID-19 severity.

According to the results, Simvastatin reduced COVID-19 severity significantly. In Table 4, these people have been evaluated based on vital signs. For the entire case study, the average fever temperature of patients is 37.2°C. However, the number of Simvastatin users is 36.831, which is lower than the total number of patients (i.e., 372°C). Moreover, systolic pressure for this case study is 137.31, which is higher than that of total patients.

Based on the obtained results, diastolic pressure for both groups is almost equal. Also, the heart rate for Simvastatin takers is lower than the entire case study, and the respiratory rate is high in Simvastatin takers. The most critical parameter of patients for this comparison is oxygen saturation. In this respect, Simvastatin takers have a 95.77 mmHg O_2Sat, which puts them in the mild group. However, in evaluating the complete case study, the O_2Sat is 92.42, putting them in the category of patients with medium severity.

In conclusion, we can estimate the positive influence of Simvastatin on COVID-19 severity for people that take Simvastatin before infection to the COVID-19 virus. The results of studying clinical symptoms are illustrated in Figure 3. The vertical axis shows the percentage of people with particular symptoms or historical illnesses for both case studies. Based on these results, 28% of all patients have diabetes, while only 15.38% of Simvastatin takers are involved in diabetes. Moreover, 61.14% of all patients have a fever in admission, while 100% of Simvastatin takers have a fever. None of the patients who have taken Simvastatin statin had a dry cough, while 49.20% (almost half) showed dry cough symptoms. In addition, no one has weakness, headache, anosmia, vomiting blood, diarrhea, epigastric pain, dizziness, throat itching, nausea, wheezing, chest pain, heart palpitations, chest tightness, and sore throat, among Simvastatin takers.

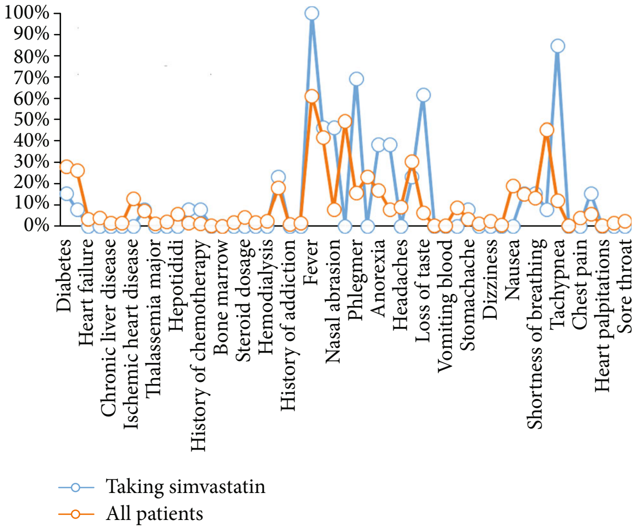


Figure 3: Study of patients taking Simvastatin and comparison with the total case study based on clinical symptoms.

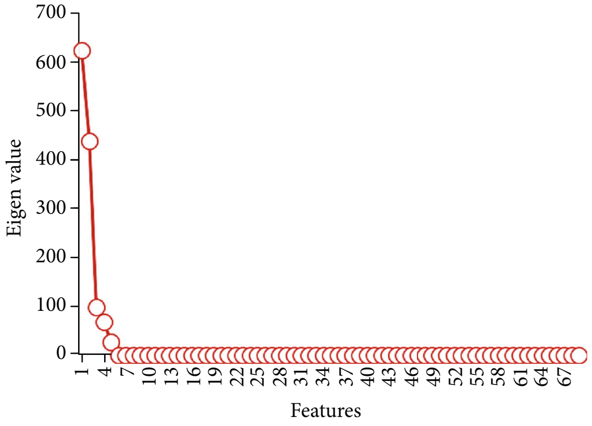
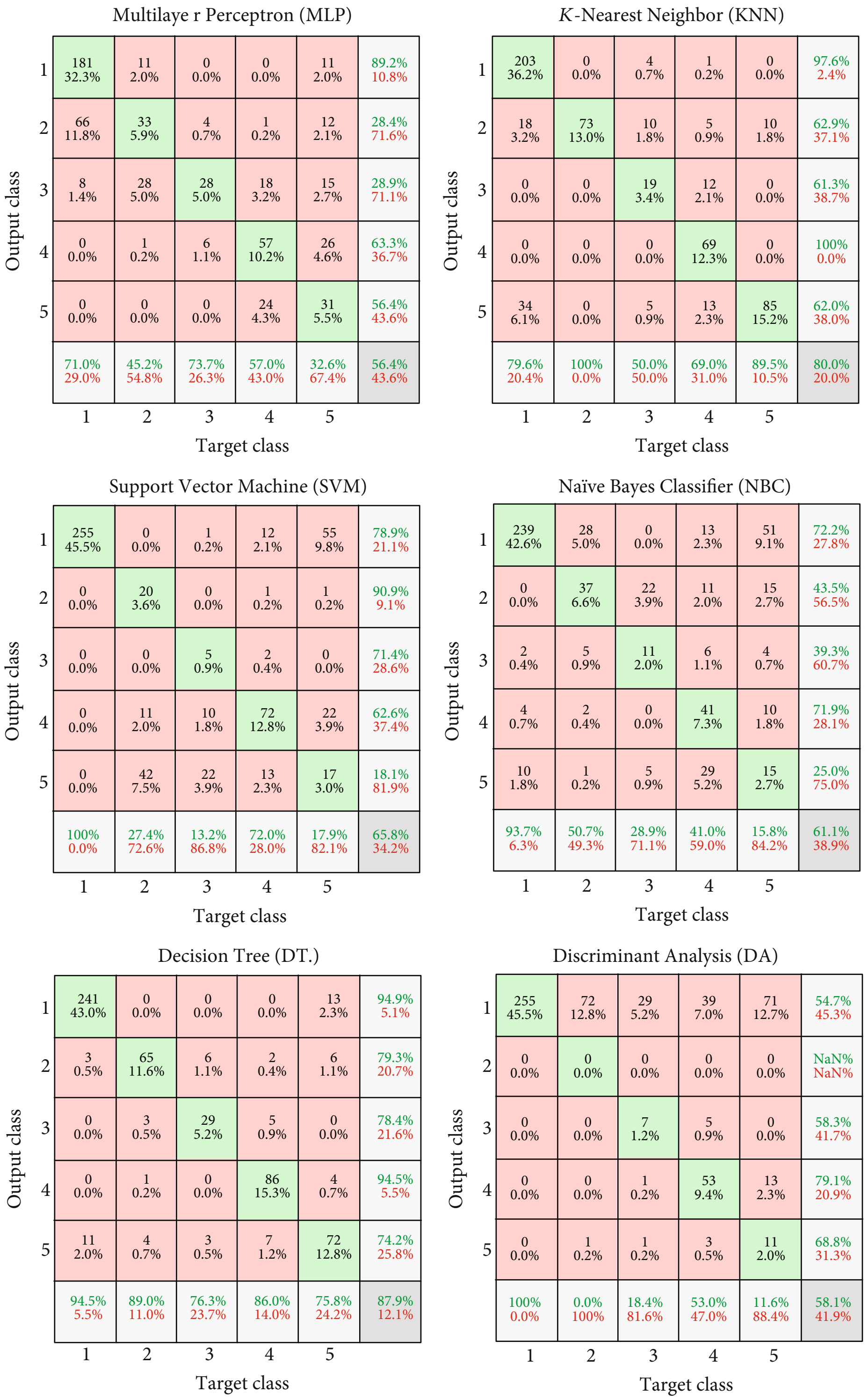


Figure 4: Results of feature reduction using PCA.

The significant signs of this subgroup are tachypnea or respiratory rate higher than 20 breaths per minute, given that 84.62% have tachypnea. Besides, 61.54% of Simvastatin takers lost their taste ability. Moreover, 69.23% of this case study has a productive cough in admission.

4.3. Diagnosis of COVID-19 Severity Based on Machine Learning Methods. Computer-aided diagnosis (CAD) tools have been recently used to study various features' impact and identify various diseases from the patient data [19]. Computationally efficient artificial neural networks (ANNs) [20, 21] have been utilized to monitor the patients' health



Multilaye r Perceptron (MLP)

	1	2	3	4	5	
1	181 32.3%	11 2.0%	0 0.0%	0 0.0%	11 2.0%	89.2% 10.8%
2	66 11.8%	33 5.9%	4 0.7%	1 0.2%	12 2.1%	28.4% 71.6%
3	8 1.4%	28 5.0%	28 5.0%	18 3.2%	15 2.7%	28.9% 71.1%
4	0 0.0%	1 0.2%	6 1.1%	57 10.2%	26 4.6%	63.3% 36.7%
5	0 0.0%	0 0.0%	0 0.0%	24 4.3%	31 5.5%	56.4% 43.6%
	71.0% 29.0%	45.2% 54.8%	73.7% 26.3%	57.0% 43.0%	32.6% 67.4%	56.4% 43.6%

K-Nearest Neighbor (KNN)

	1	2	3	4	5	
1	203 36.2%	0 0.0%	4 0.7%	1 0.2%	0 0.0%	97.6% 2.4%
2	18 3.2%	73 13.0%	10 1.8%	5 0.9%	10 1.8%	62.9% 37.1%
3	0 0.0%	0 0.0%	19 3.4%	12 2.1%	0 0.0%	61.3% 38.7%
4	0 0.0%	0 0.0%	0 0.0%	69 12.3%	0 0.0%	100% 0.0%
5	34 6.1%	0 0.0%	5 0.9%	13 2.3%	85 15.2%	62.0% 38.0%
	79.6% 20.4%	100% 0.0%	50.0% 50.0%	69.0% 31.0%	89.5% 10.5%	80.0% 20.0%

Support Vector Machine (SVM)

	1	2	3	4	5	
1	255 45.5%	0 0.0%	1 0.2%	12 2.1%	55 9.8%	78.9% 21.1%
2	0 0.0%	20 3.6%	0 0.0%	1 0.2%	1 0.2%	90.9% 9.1%
3	0 0.0%	0 0.0%	5 0.9%	2 0.4%	0 0.0%	71.4% 28.6%
4	0 0.0%	11 2.0%	10 1.8%	72 12.8%	22 3.9%	62.6% 37.4%
5	0 0.0%	42 7.5%	22 3.9%	13 2.3%	17 3.0%	18.1% 81.9%
	100% 0.0%	27.4% 72.6%	13.2% 86.8%	72.0% 28.0%	17.9% 82.1%	65.8% 34.2%

Naïve Bayes Classifier (NBC)

	1	2	3	4	5	
1	239 42.6%	28 5.0%	0 0.0%	13 2.3%	51 9.1%	72.2% 27.8%
2	0 0.0%	37 6.6%	22 3.9%	11 2.0%	15 2.7%	43.5% 56.5%
3	2 0.4%	5 0.9%	11 2.0%	6 1.1%	4 0.7%	39.3% 60.7%
4	4 0.7%	2 0.4%	0 0.0%	41 7.3%	10 1.8%	71.9% 28.1%
5	10 1.8%	1 0.2%	5 0.9%	29 5.2%	15 2.7%	25.0% 75.0%
	93.7% 6.3%	50.7% 49.3%	28.9% 71.1%	41.0% 59.0%	15.8% 84.2%	61.1% 38.9%

Decision Tree (DT.)

	1	2	3	4	5	
1	241 43.0%	0 0.0%	0 0.0%	0 0.0%	13 2.3%	94.9% 5.1%
2	3 0.5%	65 11.6%	6 1.1%	2 0.4%	6 1.1%	79.3% 20.7%
3	0 0.0%	3 0.5%	29 5.2%	5 0.9%	0 0.0%	78.4% 21.6%
4	0 0.0%	1 0.2%	0 0.0%	86 15.3%	4 0.7%	94.5% 5.5%
5	11 2.0%	4 0.7%	3 0.5%	7 1.2%	72 12.8%	74.2% 25.8%
	94.5% 5.5%	89.0% 11.0%	76.3% 23.7%	86.0% 14.0%	75.8% 24.2%	87.9% 12.1%

Discriminant Analysis (DA)

	1	2	3	4	5	
1	255 45.5%	72 12.8%	29 5.2%	39 7.0%	71 12.7%	54.7% 45.3%
2	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	NaN% NaN%
3	0 0.0%	0 0.0%	7 1.2%	5 0.9%	0 0.0%	58.3% 41.7%
4	0 0.0%	0 0.0%	1 0.2%	53 9.4%	13 2.3%	79.1% 20.9%
5	0 0.0%	1 0.2%	1 0.2%	3 0.5%	11 2.0%	68.8% 31.3%
	100% 0.0%	0.0% 100%	18.4% 81.6%	53.0% 47.0%	11.6% 88.4%	58.1% 41.9%

Figure 5: Confusion matrixes of classification using machine learning approaches.

status and diagnose various diseases such as COVID-19 and mental health disorders [21] using smartphones and smartwatches. Machine learning methods, particularly ANNs, have also been used on lung X-ray images to detect COVID-19 in the lung tissue and detect the infected areas.

Inspired by machine learning applications in intelligent healthcare and investigating various aspects of COVID-19 disease, we designed machine learning networks to diagnose the severity of COVID-19 patients based on the variables (features) mentioned before. In this regard, initially, there are 69 features as independent variables. However, to obtain the best and uncomplex nonparametric classification, we should reduce this number. Therefore, principal component analysis (PCA) was used to reduce the number of initial features. The results of the PCA method are shown in Figure 4. Based on eigenvalues resulting from PCA, the number of features is reduced to 5, suggesting that we should use five features to classify and diagnose the patients' severity.

Besides, the severity factor consists of categorical labels from 1 to 5 according to Equation (1). It is also assigned as a dependent variable for diagnosis. Here, it is aimed to find machine learning architectures to diagnose COVID-19 patients' severity based on clinical signs.

To diagnose the patients' severity, we used six types of machine learning classifiers, including multilayer perceptron (MLP), K-nearest neighbors (KNN), support vector machine (SVM), Naïve Bayes classifier (NBC), decision tree (DT), and discriminant analysis (DA). The confusion matrixes of the classification methods are illustrated in Figure 5. These matrixes consist of 5×5 class matrixes (red) that its orthogonal elements are true values (green), and red elements are false detection values. In these matrixes, gray elements show the method's sensitivity (horizontal) and precision (vertical).

The low corner element indicates the classification accuracy. For example, in Figure 5, at MLP matrix, from 255 patients with mild severity condition, 181 are diagnosed correctly. In other words, the sensitivity for this class is 71%. However, 66 of them are diagnosed as medium severity. The MLP architecture consists of three hidden layers with 20, 10, and 1 neuron(s), in the order of their appearance. The absolute accuracy for the MLP approach is 58.8% (41.2% loss). In the DT method, the final accuracy is 87.9%, which is higher than that of other KNN, SVM, NBC, and DA (i.e., 80%, 68.8%, 61.1%, and 85.1%, respectively). In the DT method, the highest sensitivity belongs to mild patients. In other words, 94.5% of mild patients are diagnosed correctly. Regarding other severity groups, the sensitivity is 89%, 76.3%, 86%, and 75.8% for medium, severe, critical, and death people, respectively. Finally, it can be concluded that the DT methods are the best classifier among machine learning methods for diagnosing COVID-19 patients from clinical features.

5. Limitations

In this study, medical information may face limitations that can prevent some of the use or disclosure. For example, there are certain restrictions on using specific categories of information (i.e., HIV testing or treatment of mental illness). Also, government medical insurances restrict the disclosure of beneficiary information for purposes not related to these insurances. These limitations have made it very difficult to access all COVID-19 patient information in this study. To deal with this shortcoming, we chose patients with complete health information. The other limitation is the lack of personal information from the patient to our specialist doctors.

6. Conclusion

Statins are multivalent cardioprotective drugs increasingly recognized as mediators with direct cellular effects beyond their cardiac role. These drugs inhibit the enzyme hydroxyl methylglutaryl coenzyme A (HMG-CoA reductase) and are responsible for accelerating the early stages of cholesterol biosynthesis. In this study, the role and possible anti-inflammatory effects of this drug are investigated. Statins that are commonly prescribed in Iran include Atorvastatin and Simvastatin. This investigation is a retrospective descriptive-analytical cross-sectional study based on the medical records of patients. According to the preliminary information of the project implementers, more than 1500 patients with COVID-19 have been hospitalized at this center from February and March 2020. In this study, the medical records of the patients were examined. Next, their clinical and laboratory characteristics, including the history of taking statins before the onset of the disease, were entered into a previously prepared and reproduced form of information. Only patients who make a definitive diagnosis based on virus isolation by RT-PCR with a swab of the throat, nasopharynx, or oropharynx and a sample of tracheal secretions or typical radiological findings were included. Severity criteria include the number of breaths equal to or more than 30 beats per minute, arterial oxygen saturation less than 93 (when the patient breathes in-room air), severe multifocal pulmonary involvement (which increases by more than 50% within 48 h), and the need for intubation and mechanical ventilation, CPAP, and BIPAP. This paper evaluated the effects of statin taking before infection on COVID-19 severity. Moreover, machine learning methods were used to diagnose COVID-19 severity based on clinical features. Overall, the results can be summarized as follows:

(i) There is an indirect (positive) relationship between taking Simvastatin and COVID-19 severity

(ii) People who have taken Simvastatin are of lower severity than others

(iii) About 69% of Simvastatin takers are of mild severity

(iv) There is no significant relationship between Atorvastatin users and COVID-19 severity

(v) Most patients who take Rosuvastatin are in critical condition

(vi) The average fever temperature of all case studies is 37.2°C

(vii) The average fever temperature of Simvastatin takers is 36.8°C

(viii) The systolic pressure for Simvastatin takers is 137.31 mmHg

(ix) The heart rate for Simvastatin takers is lower than the entire case study

(x) The respiratory rate is high in Simvastatin takers

(xi) Simvastatin takers have a 95.77 mmHg oxygen saturation, placing them in mild severity conditions

(xii) The average oxygen saturation of all case studies is 92.42 mmHg, which puts them in mild severity conditions

(xiii) About 84.62% of Simvastatin takers have tachypnea

(xiv) About 61.54% of Simvastatin takers lost their taste ability

(xv) Principle component analysis (PCA) was used to reduce initial features from 71 to 5

(xvi) The accuracy of the decision tree method is 87.9%, which is higher than that of other approaches

(xvii) The accuracy of KNN, SVM, NBC, and DA is 80%, 68.8%, 61.1%, and 85.1%, respectively

(xviii) The sensitivity of the DT method for patient diagnosis is 89%, 76.3%, 86%, and 75.8% for medium, severe, critical, and dead people, respectively

In conclusion, we can estimate the positive influence of Simvastatin on COVID-19 severity for people that take Simvastatin before infection to the COVID-19 virus. Furthermore, it was found that the decision tree method is an effective tool to predict the patients' severity based on clinical symptoms.

Data Availability

The present study investigates the effect of using standard doses of statins in the months before infection in patients with COVID-19 admitted to Razi Hospital in Ghaemshahr (Mazandaran Province, Iran), and the data of the article is unpublishable due to the preservation of patients' information.

Conflicts of Interest

The authors declare that they have received no financial support or have no conflicts of interest in this research and its publication.

References

[1] W. Guan, Z. Y. Ni, Y. Hu et al., "Clinical characteristics of coronavirus disease 2019 in China," *The New England Journal of Medicine*, vol. 382, no. 18, pp. 1708–1720, 2020.

[2] World Health Organization, *Events as they happen*, World Health Organization, 2020, September 2020, https://www.who.int/emergencies/diseases/novel-coronavirus-2019/events-as-they-happen.

[3] T. Li, "Diagnosis and clinical management of severe acute respiratory syndrome Coronavirus 2 (SARS-CoV-2) infection: an operational recommendation of Peking Union Medical College Hospital (V2.0): Working Group of 2019 Novel Coronavirus Peking Union Medical College Hospital," *Emerging Microbes & Infections*, vol. 9, no. 1, pp. 582–585, 2020.

[4] P. Mehrbod, A. R. Omar, M. Hair-Bejo, A. Haghani, and A. Ideris, "Mechanisms of action and efficacy of statins against influenza," *BioMed Research International*, vol. 2014, Article ID 872370, 8 pages, 2014.

[5] D. Radenkovic, S. Chawla, M. Pirro, A. Sahebkar, and M. Banach, "Cholesterol in relation to COVID-19: should we care about it?," *Journal of Clinical Medicine*, vol. 9, no. 6, article 1909, 2020.

[6] S. S. Virani, *Is there a role for statin therapy in acute viral infections?*, American College of Cardiology, 2020.

[7] D. S. Fedson, S. M. Opal, and O. M. Rordam, "Hiding in plain sight: an approach to treating patients with severe covid-19 infection," *MBio*, vol. 11, no. 2, 2020.

[8] X.-J. Zhang, J. J. Qin, X. Cheng et al., "In-hospital use of statins is associated with a reduced risk of mortality among individuals with COVID-19," *Cell Metabolism*, vol. 32, no. 2, pp. 176–187.e4, 2020.

[9] R. R. Rodrigues-Diez, A. Tejera-Muñoz, L. Marquez-Exposito et al., "Statins: could an old friend help in the fight against COVID-19?," *British Journal of Pharmacology*, vol. 177, no. 21, pp. 4873–4886, 2020.

[10] V. Castiglione, M. Chiriacò, M. Emdin, S. Taddei, and G. Vergaro, "Statin therapy in COVID-19 infection," *European Heart Journal - Cardiovascular Pharmacotherapy*, vol. 6, no. 4, pp. 258-259, 2020.

[11] R. Subir, J. Mukherjee Jagat, and K. Gangopadhyay Kalyan, "Pros and cons for use of statins in people with coronavirus disease-19 (COVID-19)," *Diabetes and Metabolic Syndrome: Clinical Research & Reviews*, vol. 14, no. 5, pp. 1225–1229, 2020.

[12] L. B. Daniels, A. M. Sitapati, J. Zhang et al., "Relation of statin use prior to admission to severity and recovery among COVID-19 inpatients," *The American Journal of Cardiology*, vol. 136, pp. 149–155, 2020.

[13] Ž. Reiner, M. Hatamipour, M. Banach et al., "Statins and the Covid-19 main protease: in silico evidence on direct interaction," *Archives of Medical Science*, vol. 16, no. 3, pp. 490–496, 2020.

[14] M. Madjid, P. Safavi-Naeini, S. D. Solomon, and O. Vardeny, "Potential effects of coronaviruses on the cardiovascular system: a review," *JAMA Cardiology*, vol. 5, no. 7, pp. 831–840, 2020.

[15] A. L. Totura, A. Whitmore, S. Agnihothram et al., "Toll-like receptor 3 signaling via TRIF contributes to a protective innate immune response to severe acute respiratory syndrome coronavirus infection," *MBio*, vol. 6, no. 3, pp. 1–14, 2015.

[16] M. L. DeDiego, J. L. Nieto-Torres, J. A. Regla-Nava et al., "Inhibition of NF-κB-Mediated inflammation in severe acute respiratory syndrome coronavirus-infected mice increases survival," *Journal of Virology*, vol. 88, no. 2, pp. 913–924, 2014.

[17] X. Yuan, Y. Deng, X. Guo, J. Shang, D. Zhu, and H. Liu, "Atorvastatin attenuates myocardial remodeling induced by chronic intermittent hypoxia in rats: partly involvement of TLR-4/MYD88 pathway," *Biochemical and Biophysical Research Communications*, vol. 446, no. 1, pp. 292–297, 2014.

[18] P. Chansrichavala, U. Chantharaksri, P. Sritara, and S. C. Chaiyaroj, "Atorvastatin attenuates TLR4-mediated NF-κB activation in a MyD88-dependent pathway," *Asian Pacific Journal of Allergy and Immunology*, vol. 27, no. 1, pp. 49–57, 2009.

[19] M. Ahmadi, A. Sharifi, S. Hassantabar, and S. Enayati, "QAIS-DSNN: tumor area segmentation of MRI image with optimized quantum matched-filter technique and deep spiking neural network," *BioMed Research International*, vol. 2021, Article ID 6653879, 16 pages, 2021.

[20] S. Hassantabar, N. Stefano, V. Ghanakota et al., “Coviddeep: Sars-cov-2/covid-19 test based on wearable medical sensors and efficient neural networks,” 2020, http://arxiv.org/abs/2007.10497.

[21] S. Hassantabar, M. Ahmadi, and A. Sharifi, “Diagnosis and detection of infected tissue of COVID-19 patients based on lung X-ray image using convolutional neural network approaches,” *Chaos, Solitons & Fractals*, vol. 140, article 110170, 2020.

Hindawi
BioMed Research International
Volume 2021, Article ID 5515342, 11 pages
https://doi.org/10.1155/2021/5515342

Research Article

i4mC-EL: Identifying DNA N4-Methylcytosine Sites in the Mouse Genome Using Ensemble Learning

Yanjuan Li, Zhengnan Zhao, and Zhixia Teng

College of Information and Computer Engineering, Northeast Forestry University, Harbin 150040, China

Correspondence should be addressed to Zhixia Teng; tengzhixia@nefu.edu.cn

Received 9 February 2021; Accepted 21 May 2021; Published 1 June 2021

Academic Editor: Andrea Scribante

As one of important epigenetic modifications, DNA N4-methylcytosine (4mC) plays a crucial role in controlling gene replication, expression, cell cycle, DNA replication, and differentiation. The accurate identification of 4mC sites is necessary to understand biological functions. In the paper, we use ensemble learning to develop a model named i4mC-EL to identify 4mC sites in the mouse genome. Firstly, a multifeature encoding scheme consisting of Kmer and EIIP was adopted to describe the DNA sequences. Secondly, on the basis of the multifeature encoding scheme, we developed a stacked ensemble model, in which four machine learning algorithms, namely, BayesNet, NaiveBayes, LibSVM, and Voted Perceptron, were utilized to implement an ensemble of base classifiers that produce intermediate results as input of the metaclassifier, Logistic. The experimental results on the independent test dataset demonstrate that the overall rate of predictive accurate of i4mC-EL is 82.19%, which is better than the existing methods. The user-friendly website implementing i4mC-EL can be accessed freely at the following.

1. Introduction

As a chemical modification occurring on DNA sequences, DNA methylation can change genetic properties under the condition that the order of DNA sequences remains unchanged. DNA methylation has many manifestations, such as 5-methylcytosine (5mC for short), N6-methyladenine (6 mA for short), and N4-methylcytosine (4mC for short) [1]. Among them, the 5mCs are widely present in prokaryotes and eukaryotes and are of great significance for controlling gene differentiation and gene expression, maintaining chromosome stability and cell structure [2, 3]. They also can cause some diseases such as cancer [4–6]. The 6mAs are also widely distributed in prokaryotes and eukaryotes, which play a crucial role in replication, expression, and transcription of gene [7]. The 4mCs which were found in 1983 mainly exist in prokaryotes, and they can control DNA replication, gene expression, and cell cycle [8]. However, compared with 5mCs and 6mAs, the current research on 4mCs is still insufficient. To make up for this defect and further understand 4mCs' biological properties and functions, the first thing we need to do is to identify 4mCs from various DNA sequences, which is still a hot research topic so far.

In order to identify 4mCs, many biology-based approaches have been explored. Single molecule real-time sequencing technology (SMRT for short) [9, 10] detects optical signals of bases matching the template at the single-molecule level to identify 4mCs. 4mC-Tet-assisted-bisulfite-sequencing technology (4mC-Tet for short) [11] identifies 4mCs by using bisulfite to convert unmethylated cytosine in the DNA sequences into uracil while to keep methylated cytosine unchanged. However, this kind of technologies is time-consuming and resource-intensive. Moreover, the explosive growth of DNA sequences also makes it more difficult to achieve whole-genome sequencing through these technologies. Therefore, using machine learning (ML for short) to identify 4mCs shows more advantages. Up to now, there are many models using machine learning to identify 4mCs. iDNA4mC [12], the earliest model for 4mC identification, is primarily used to identify 4mCs from the genomes of six species, A.thaliana, C.elegans, D.melanogaster, E.coli, G.pickeringii, and G.subterraneus, and its positive data containing 4mCs were obtained from a reliable database called MethSMRT [13]. Soon afterwards, several other models, 4mcPred [14], 4mcPred-SVM [15], 4mcPred-IFL [16], and Meta-4mcPred [17], were proposed successively, which used the same dataset as iDNA4mC [12] for 4mCs identification of the genomes of these six species. i4mC-Rose [18] is the first and the only model for 4mCs identification in the genome of Rosaceae, and it derived positive dataset from the MDR [19] and the other reliable database for storing 4mC data. For the mouse genome we wanted to study, there have been currently two models, 4mCpred-EL [20] and i4mC-Mouse [21]. Among them, their samples containing 4mCs were also obtained from the MethSMRT database. In addition, 4mCpred-EL selected 4 ML algorithms and 7 feature encoding schemes to generate 28 sets of results as the final coding. Subsequently, 4mCpred-EL trained 4 submodels through the final coding and these 4 ML algorithms and then combined the 4 submodels into the final model by majority voting. i4mC-Mouse trained 6 submodels using 6 feature encoding schemes and random forest (RF for short) algorithm, and then the 6 submodels were combined into the final model by weighted voting. Compared with 4mCpred-EL, i4mC-mouse has better performance according to the indicators, ACC and MCC. Although exciting results have been achieved in 4mCpred-EL and i4mC-Mouse, the performance is able to be further increased. In this paper, to further improve the prediction capability, we propose a new mouse's 4mCs predictor, i4mC-EL.

2. Materials and Methods

2.1. Framework of i4mC-EL. In the present study, a novel model named i4mC-EL is proposed to indentify mouse's 4mCs, and we can see the framework of it in Figure 1. First, using two different feature encoding schemes, Kmer and EIIP, each DNA sequence was encoded into a 1364-dimensional vector and a 41-dimensional vector, respectively. Next, the 1364-dimensional vector and the 41-dimensional vector of each DNA sequence were combined to form a 1405-dimensional multifeature vector. Finally, a two-stage stacked ensemble learning classifier with these multifeature vectors as input was constructed. The ensemble classifier used BayesNet, NavieBayes Multinomial, LibSVM, and Voted Perceptron as base classifiers and used Logistic as metaclassifier. i4mC-EL's datasets and feature encoding schemes and classifiers will be described detailly below.

2.2. Dataset. This paper adopted the benchmark dataset constructed by Hasan'steam [21]. In this dataset, the positive samples containing mouse's 4mCs were obtained from the MethSMRT [17] database, and the negative samples were taken from chromosome DNA sequences. They were all fragments of DNA sequences consisting of 41 nucleotides with a "C" in the middle. Only the sequences whose modQV value greater than or equal to 20 were considered to obtain the high-quality dataset. To prevent the predictor from overfitting, the threshold of CD-HIT [22] was set to 70% to remove redundant sequences [23]. The dataset contained 1,812 DNA sequences, 906 of which were 4mCs and 906 were non-4mCs. About 80% of the dataset was randomly selected as the training dataset, and the remaining about 20% was used as the independent test dataset. The training dataset (train-1492) consisted of 746 4mCs and 746 non-4mCs. And the independent test dataset (test-320) included 160 4mCs and 160 non-4mCs.

2.3. Feature Encoding. Transforming DNA sequences into vectors that can make a distinction between 4mCs and non-4mCs availably is the first step to build an ensemble learning-based predictor to identify 4mCs [24–29]. Here, a multifeature encoding scheme composed of Kmer [30–33] and EIIP [34] was used to encode DNA sequences. Kmer represented the DNA sequences as the occurrence frequencies of k adjacent nucleotides. EIIP encoded each nucleotide in DNA sequences with its corresponding electron-ion energy. In the experiment of Section 3, we will find that this multifeature is able to encode DNA sequences availably. The following parts are detailed descriptions of Kmer and EIIP.

2.3.1. Kmer. This encoding scheme refers to the frequency of k-nucleotides composed of k continuous nucleotides in each sequence. For sequence $D = d_1 d_2 d_3 \cdots d_{L-2} d_{L-1} d_L$, each element of each feature vector is calculated by Equation (1):

$$f(X) = \frac{F(X)}{L - k + 1}, \tag{1}$$

where X is one of the k-nucleotide, $F(X)$and $f(X)$ are the count and frequency of X in D, respectively, and L is D's length. After Kmer, sequences are transformed into 4^k -dimensional vectors. For example, when the k-mer parameter $k = 2$, the value of AA in the 16-dimensional (4^2) feature vector of sequence D_1 = AAACTAGTC is 0.25.

In the present study, we choose the values of the parameter k to be 1, 2, 3, 4, and 5, generating 1364-dimensional ($4^1 + 4^2 + 4^3 + 4^4 + 4^5$) feature vectors.

2.3.2. EIIP. EIIP is the short name of electron-ion interaction pseudopotential. The encoding scheme based on EIIP was proposed by Nair and Sreenadhan in 2006. Through it, each nucleotide in each sequence is replaced by its corresponding electron-ion interaction pseud potential value (Table 1). For example, the result of sequence D_2 = AACTG after EIIP encoding is (0.1260, 0.1260, 0.1340, 0.1335, 0.0806). In the present study, each sequence is transformed into a 41-dimensional feature vector.

2.4. Classifier. As an open data mining platform, Weka has assembled a large number of machine learning algorithms that can undertake data mining tasks. In the present paper, the classifiers we used were all implemented by Weka, such as BayesNet, NaiveBayes, SGD, SimpleLogistic, SMO, IBk, JRip, J48, and ensemble learning. Finally, we chose the ensemble learning, and the results of related experiment will be presented in section 3.

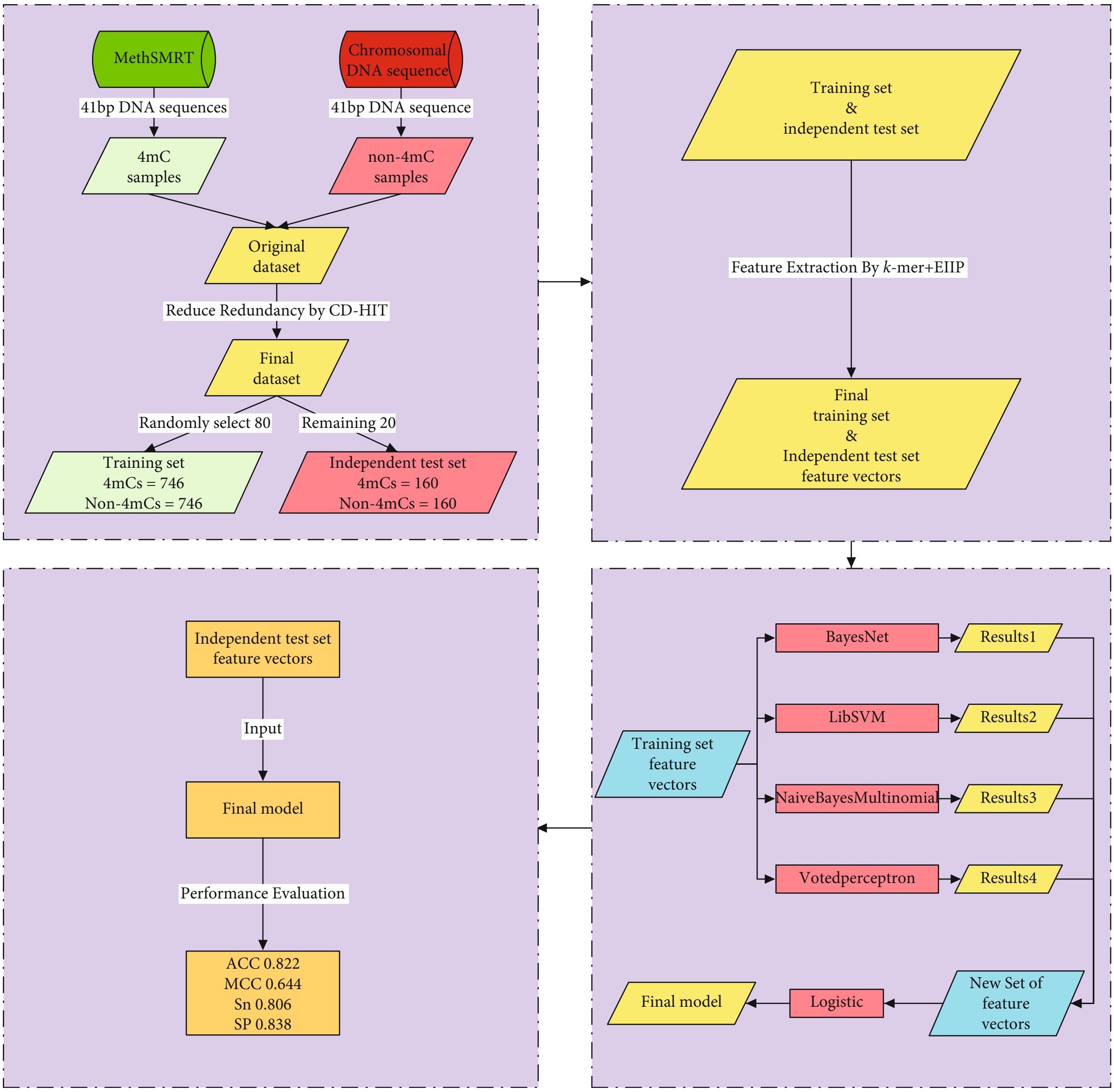

MethSMRT
Chromosomal DNA sequence
41bp DNA sequences
41bp DNA sequence
4mC samples
non-4mC samples
Original dataset
Reduce Redundancy by CD-HIT
Final dataset
Randomly select 80
Remaining 20
Training set 4mCs = 746 Non-4mCs = 746
Independent test set 4mCs = 160 Non-4mCs = 160
Training set & independent test set
Feature Extraction By k-mer+EIIP
Final training set & Independent test set feature vectors
Training set feature vectors
BayesNet
LibSVM
NaiveBayesMultinomial
Votedperceptron
Results1
Results2
Results3
Results4
New Set of feature vectors
Logistic
Final model
Independent test set feature vectors
Input
Final model
Performance Evaluation
ACC 0.822
MCC 0.644
Sn 0.806
SP 0.838

FIGURE 1: The framework of i4mC-EL.

TABLE 1: The electron-ion interaction pseudopotential values for DNA nucleotides.

NT	A	C	G	T
EIIP	0.1260	0.1340	0.0806	0.1335

According to different combination strategies, bagging, boosting, and stacking are the three main types of ensemble learning. Ensemble learning is widely used in bioinformatics because it can improve the prediction performance of classifiers, such as protein-protein interaction [35], disease prediction [36], type III secreted effectors prediction [37], and protein subcellular location prediction [38]. In detail, we used two-stage stacked ensemble learning.

In the two-stage stacked ensemble learning, the base classifiers used in this paper include BayesNet [39], Voted Perceptron [40], Naive Bayes Multinomial [41], and LibSVM [42], and the metaclassifier was Logistic. At the first stage of the ensemble learning classifier, based on the multifeature vectors proposed in this paper, four base classifiers are, respectively, trained to relabel the training dataset and the independent test dataset. At the second stage, the outputs of base classifiers are utilized as input for the metaclassifier.

Figure 2 gives the detailed process of model generation and result output, the steps are as follows.

Step 1. Partition dataset. Divide the training dataset into ten parts and mark them as train 1, train 2, …, train 10. The independent test dataset remains unchanged.

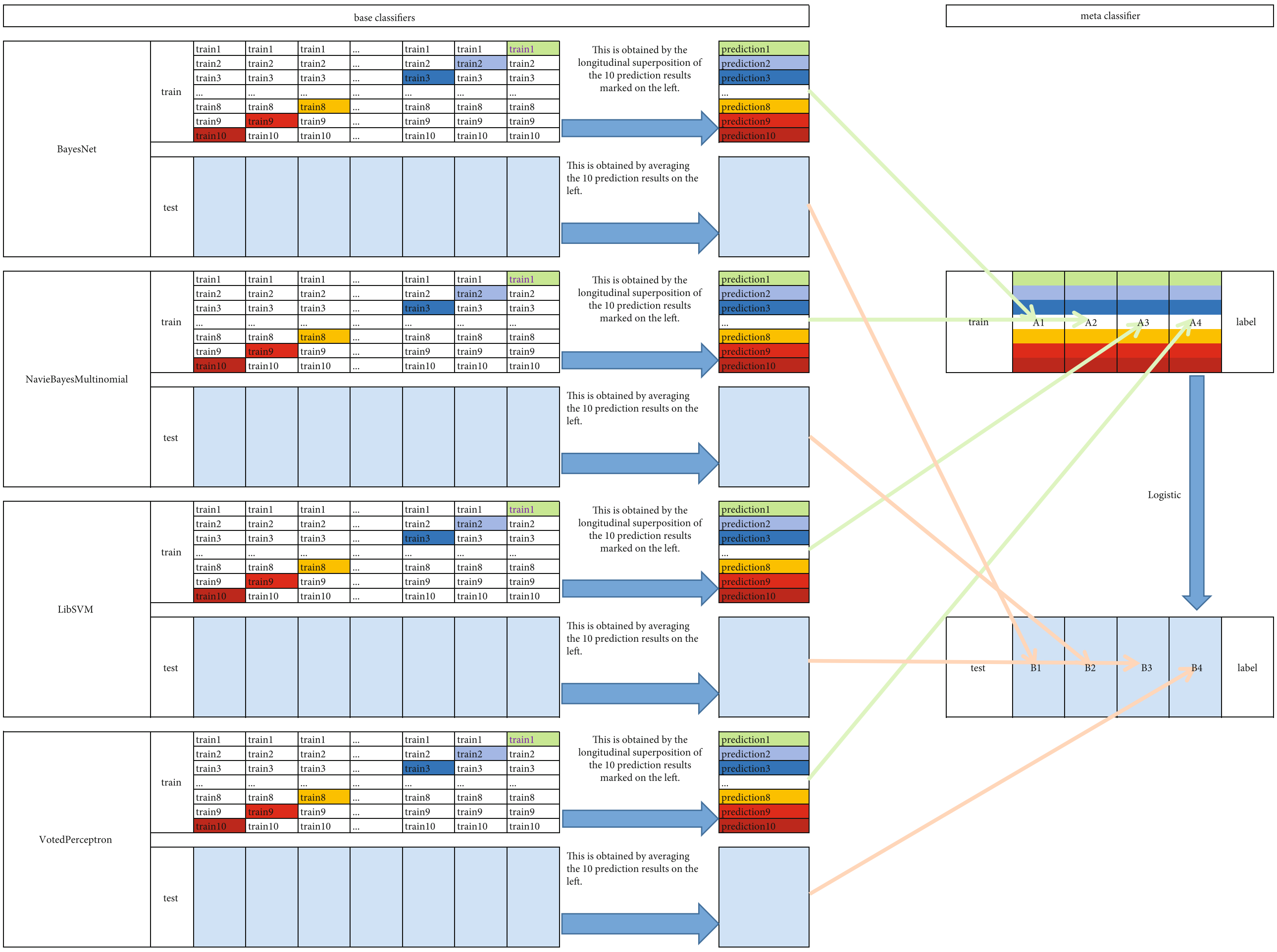


Figure 2: Working diagram of ensemble learning.

Step 2. Train base classifiers. In the present paper, we chose BayesNet, Voted Perceptron, Naive Bayes Multinomial, and LibSVM as base classifiers. For one base classifier such as BayesNet, 10-fold crossvalidation is performed. In detail, train 1, train 2, …, train 10 are used as validation dataset in turn, the other nine parts are used as the training dataset, and prediction is made on the independent test dataset. This would get 10 predictions from the training dataset together with another 10 predictions on the independent test dataset. Combine the 10 predictions on the training dataset vertically to get A1 and take the average of the 10 predictions on the independent test dataset to get B1. Similarly, we could get A2, B2 from NavieBayes Multinomial, A3, B3 from LibSVM, and A4, B4 from Voted Perceptron.

Step 3. Train metaclassifiers. Use the predictive values of the 4 base classifiers on the training dataset, A1, A2, A3, and A4, as 4 features to train the logistic classifier.

Step 4. Predict new data. Use the trained model to make predictions on the 4 features, B1, B2, B3, and B4, constructed from the predicted values of the independent test dataset of the 4 base classifiers, and then the final prediction results are obtained.

Table 2: The contrast of performance for dissimilar feature encoding schemes under 10-fold crossvalidation.

Schemes	ACC	MCC	Sn	Sp
BPF	0.668	0.335	0.665	0.670
DPE	0.614	0.228	0.619	0.609
RFHC	0.658	0.316	0.669	0.647
RevKmer	0.755	0.511	0.745	0.765
PseKNC	0.794	0.589	0.786	0.803
k-mer + BPF	0.724	0.448	0.729	0.718
k-mer + RFHC	0.747	0.493	0.744	0.749
RevKmer+DBE	0.738	0.476	0.723	0.753
RevKmer+EIIP	0.779	0.558	0.764	0.794
k-mer + BPF + DPE	0.732	0.464	0.741	0.723
Our method	0.803	0.606	0.784	0.822

2.5. Performance Evaluation. For the sake of validating the quality of our classification predictor, we used four indicators widely adopted in the field of bioinformatics for evaluation [43–53]. These indicators can be calculated using the

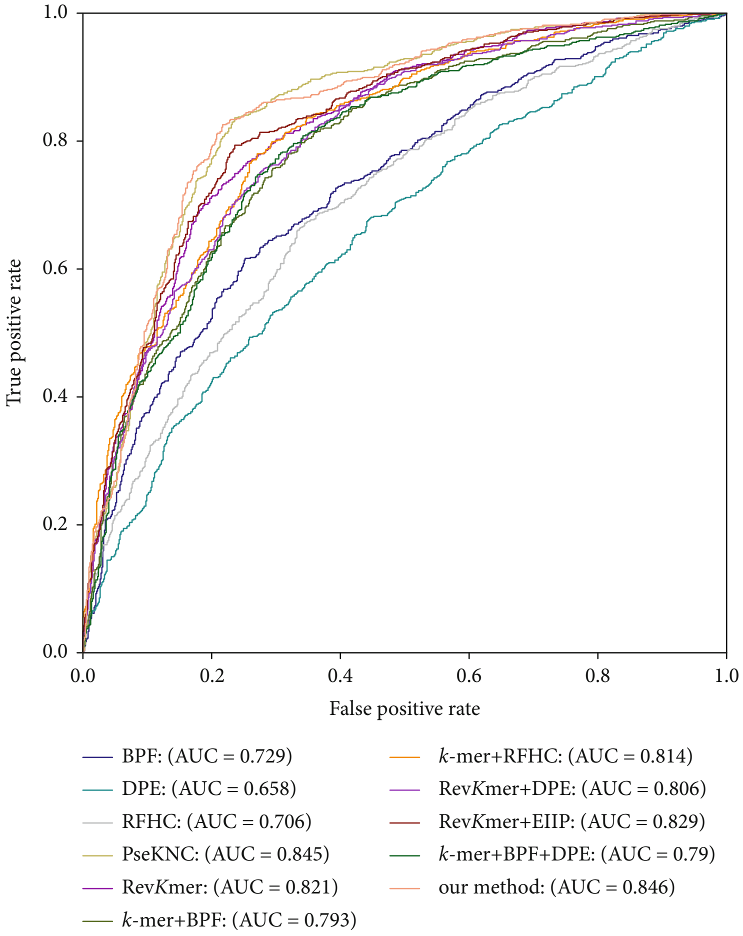


Figure 3: ROC curves for dissimilar feature encoding schemes under 10-fold crossvalidation.

formulas below:

$$\begin{aligned}
\mathrm{ACC} &= \frac{\mathrm{TN}+\mathrm{TP}}{\mathrm{TN}+\mathrm{FN}+\mathrm{FP}+\mathrm{TP}},\\
\mathrm{MCC} &= \frac{\mathrm{TN}\times\mathrm{TP}-\mathrm{FN}\times\mathrm{FP}}{\sqrt{(\mathrm{TN}+\mathrm{FN})\times(\mathrm{FN}+\mathrm{TP})\times(\mathrm{TP}+\mathrm{FP})\times(\mathrm{FP}+\mathrm{TN})}},\\
\mathrm{Sn} &= \frac{\mathrm{TP}}{\mathrm{FN}+\mathrm{TP}},\\
\mathrm{Sp} &= \frac{\mathrm{TN}}{\mathrm{FP}+\mathrm{TN}},
\end{aligned} \tag{2}$$

where TP indicates the number of the sequences that they are actually 4mCs, and that they are identified as 4mCs by the model, FP indicates the number of the sequences that they are actually non-4mCs but that they are identified as 4mCs by the model, TN indicates the number of the sequences that they are actually non-4mCs, and that they are identified as non-4mCs by the model, FN indicates the number of the sequences that they are actually 4mCs but that they are identified as non-4mCs by the model. The Sn refers to the prediction accuracy of 4mCs. The Sp refers to the prediction accuracy of non-4mCs. ACC refers to the prediction accuracy of both 4mCs and non-4mCs. MCC represents the reliability of the prediction results. The higher the values of the above four indicators have, the more superior the capability of the predictor would be.

Table 3: The contrast of performance for dissimilar classifiers under 10-fold crossvalidation.

Classifiers	ACC	MCC	Sn	Sp
BayesNet	0.727	0.453	0.739	0.714
NaiveBayes	0.752	0.504	0.751	0.753
SGD	0.712	0.424	0.710	0.713
SimpleLogistic	0.761	0.522	0.753	0.768
SMO	0.702	0.405	0.706	0.698
IBk	0.637	0.276	0.584	0.690
JRip	0.707	0.414	0.692	0.723
J48	0.665	0.330	0.674	0.655
RandomForest	0.770	0.541	0.753	0.787
AdaBoostM1	0.713	0.427	0.739	0.688
Bagging	0.729	0.459	0.744	0.714
Our method	0.803	0.606	0.784	0.822

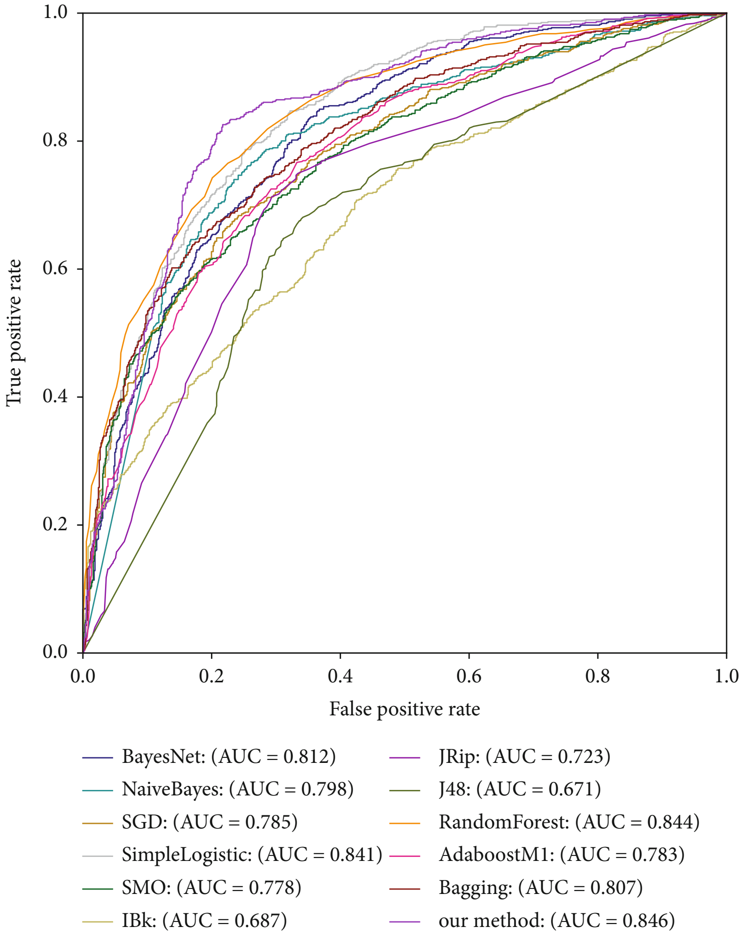


Figure 4: ROC curves for dissimilar classifiers under 10-fold crossvalidation.

3. Results and Discussion

3.1. Crossvalidation Results of TRAIN-1492. To find the features that can adequately represent the structure and function of the DNA sequences, we attempted to contrast numerous feature encoding schemes. And to achieve the optimal accuracy, we also tried to train the model using several different classification algorithms. The results of relevant comparative experiments are as below.

3.1.1. Feature Encoding Comparison on Crossvalidation. AS shown in section of "feature encoding," we encode the DNA sequences with a multifeature, which combines k-mer and EIIP feature encoding method. To verify the validity of the proposed multifeature, we compare the proposed multifeature with BPF, DPE, RFHC, RevKmer, and PseKNC feature encoding schemes and their combinations using ensemble learning classification. Among them, BPF and DPE are encoding schemes based on nucleotide positions, in which BPF takes mononucleotides as its encoding targets, while DPE takes dinucleotides as its encoding targets. RFHC is an encoding scheme based on the physicochemical properties of nucleotides. RevKmer is a variant of Kmer that considers not only the current k-nucleotides themselves, but also their reverse complementary nucleotides. PseKNC is a method to integrate continuous local and global k-tuple nucleotide information into the feature vectors of DNA sequences.

Table 4: The contrast of performance for dissimilar feature encoding schemes on TEST-320.

Schemes	ACC	MCC	Sn	Sp
BPF	0.753	0.530	0.606	0.900
DPE	0.697	0.401	0.600	0.794
RFHC	0.716	0.438	0.631	0.800
RevKmer	0.666	0.335	0.744	0.588
PseKNC	0.781	0.563	0.788	0.775
k-mer + BPF	0.772	0.553	0.681	0.863
k-mer + RFHC	0.800	0.614	0.694	0.906
RevKmer+DBE	0.756	0.516	0.700	0.813
RevKmer+EIIP	0.713	0.427	0.763	0.663
k-mer + BPF + DPE	0.772	0.553	0.681	0.863
Ourmethod	0.822	0.644	0.806	0.838

Table 2 displays experimental results, in which "our method" denotes the multifeature mentioned in the section "feature encoding." As shown in Table 2, from the perspective of ACC and MCC, the index values of our method are

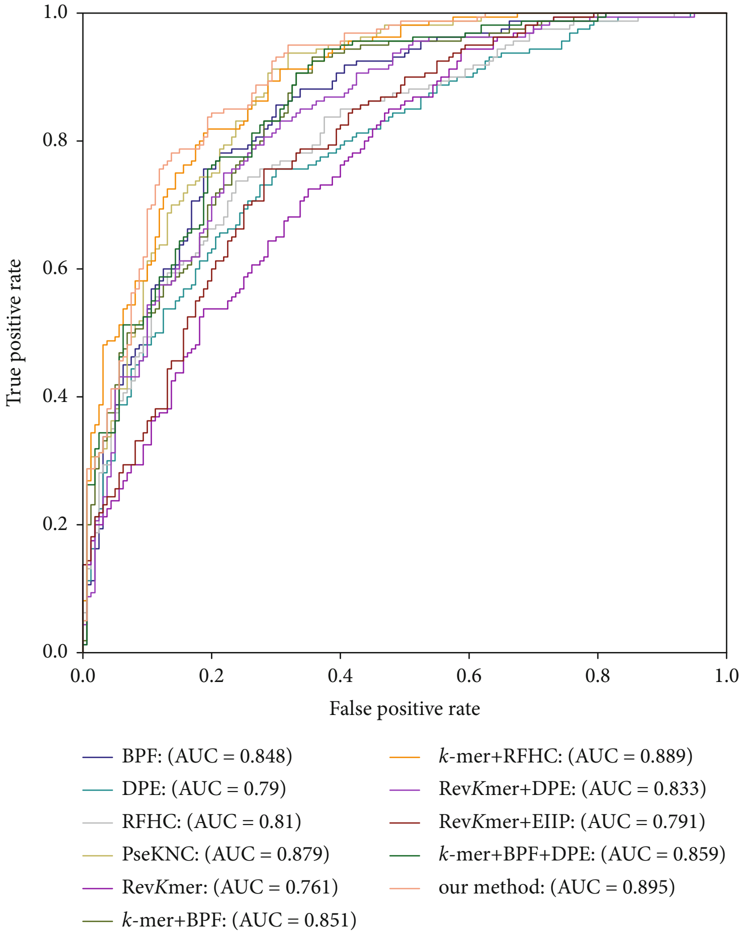


Figure 5: ROC curves for dissimilar feature encoding schemes on TEST-320.

higher than those of all other feature encoding schemes, which indicates that our method has a better overall performance. From the perspective of Sp, the index value of our method is still the highest, which indicates that it is more dominant to identify non-4mC from negative samples. These conclusions demonstrate that our method has good validity.

To further illustrate the prediction capability of our selected multifeature encoding scheme, the ROC curves for dissimilar feature encoding schemes under 10-fold crossvalidation are displayed in Figure 3. From Figure 3, we can see that our method has the largest area under ROC curve (AUC), which demonstrates that our method can represent mouse's DNA sequences better than others.

Table 5: The contrast of performance for dissimilar classifiers on TEST-320.

Classifiers	ACC	MCC	Sn	Sp
BayesNet	0.769	0.547	0.675	0.863
NaiveBayes	0.788	0.577	0.744	0.831
SGD	0.688	0.379	0.756	0.619
Simple Logistic	0.728	0.456	0.738	0.719
SMO	0.675	0.353	0.744	0.606
IBk	0.600	0.201	0.563	0.638
JRip	0.769	0.541	0.713	0.825
J48	0.663	0.325	0.656	0.669
Random Forest	0.778	0.558	0.738	0.819
AdaBoostM1	0.791	0.581	0.794	0.788
Bagging	0.781	0.564	0.744	0.819
Our method	0.822	0.644	0.806	0.838

3.1.2. Classifier Comparison on Crossvalidation. As shown in the section "classifier," we inputted the multifeature composed of *k*-mer and EIIP into an ensemble learning classifier called stacking, then obtained a predictor which is used for identifying mouse's 4mCs. To verify the validity of stacking used in this paper, on the basis of the multifeature used in this paper, we compared stacking with eleven commonly used classifiers, BayesNet, Naive Bayes, SGD, Simple Logistic, SMO, IBK, JRip, J48, Random Forest, AdaBoostM1, and Bagging. Among them, BayesNet characterizes the dependencies among attributes with the aid of directed acyclic graphs and uses conditional probability tables to describe the joint probability distribution of attributes. NaiveBayes is a simple probabilistic classifier based on Bayes' theorem under the assumption that each attribute is independent of each other. SGD implements a regularized linear support vector machine classifier with stochastic gradient descent learning. Simple

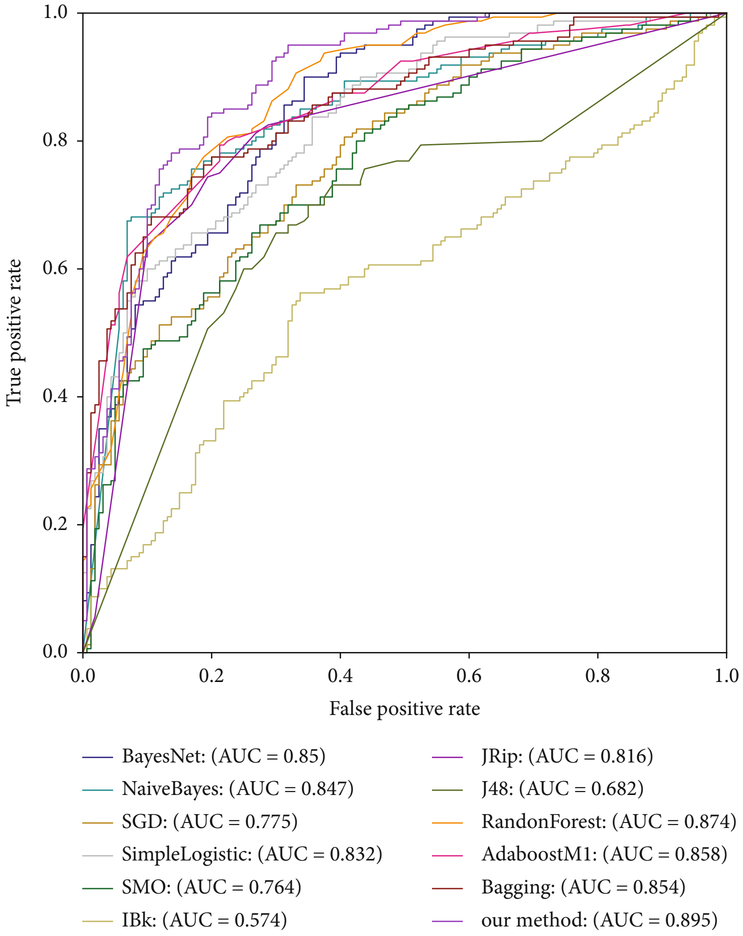


Figure 6: ROC curves for dissimilar classifiers on TEST-320.

Logistic is a linear logistic regression classifier with only one independent variable. SMO is a support vector machine classifier using a continuous minimum optimization algorithm. IBk classifies the data point by determining the category of k data points closest to it. JRip is a classifier based on rule induction. J48 is a decision tree classifier that uses information gain rate to select attributes for partitioning. Random Forest refers to a classifier that utilizes multiple trees to train and predict a sample. AdaBoostM1 is a classifier that enables the previously incorrectly predicted training samples to receive more attention at follow-up by adjusting their distribution. Bagging uses bootstrap sampling to obtain m (m is the predetermined number of base classifiers) sample datasets from the original dataset, which are used to train m base classifiers that are then integrated by voting.

The results of these comparative experiments are displayed in Table 3, where “our method” refers to the stacking classifier. From Table 3, we can see that our method outperforms the other classifiers in all indicators.

To further illustrate the classification capability of our selected stacking classifier, the ROC curves for dissimilar classifiers under 10-fold crossvalidation are displayed in Figure 4. From Figure 4, we can see that the area under ROC curve (AUC) of our method is the largest, which proves that our proposed method has better prediction performance for identifying 4mCs in the mouse genome than other methods.

3.2. Independent Validation Results of TEST-320. In this section, a comparative experiment on the independent test dataset (TEST-320) will be conducted to show the generalization capability of our selected multifeature and stacking classifier. The rationale for this is that this model is trained and tested on two different datasets, which is the equivalent of performing a real prediction task with the generated model.

3.2.1. Feature Encoding Comparison on Independent Validation. Using the stacking classifier, we, respectively, evaluate the generalization capability of various feature encoding schemes described in Section 3.1.1 on TEST-320. Table 4 displays these comparison experimental results. From Table 4, among the compared feature encoding schemes, our method performed best in ACC, Sn, and MCC, which were 82.19%, 0.806, and 0.644, respectively. Although the Sp of our method is lower than that of BPF, k-mer + BPF, k-mer + RFHC, k-mer + BPF + DPE, and PseKNC+EIIP+RFHC, the other three indicators of our method are higher than theirs.

For the sake of further describing the generalization capability of our selected multifeature encoding scheme, Figure 5

Table 6: The contrast of performance for dissimilar models on TEST-320.

Models	ACC	MCC	Sn	Sp
4mcPred-EL	0.791	0.584	0.757	0.825
i4mC-Mouse	0.816	0.633	0.807	0.825
i4mC-EL	0.822	0.644	0.806	0.838

displays the ROC curves for dissimilar feature encoding schemes on TEST-320. From Figure 5, we can see that the AUC of our method is the largest, and the ROC curve of our method is closer to the upper left, which demonstrates that our selected multifeature is more suitable than other schemes to encode the DNA sequences used to recognize mouse's 4mC.

3.2.2. Classifier Comparison on Independent Validation. We compared stacking classifier used in this paper with other eleven classifiers on TEST-320 under the condition of using the multifeature combing *k*-mer and EIIP as the input of the stacking. The results of these comparative experiments are displayed in Table 5, from which we can see that although the Sp of BayesNet is a little higher than that of our method, our method outperforms other classifiers in ACC, Sn, and MCC. Overall, our selected stacking classifier performs better than the others, indicating that it is effective for identifying mouse's 4mC.

For the sake of further describing the generalization capability of our selected stacking classifier, the ROC curves for dissimilar classifiers on TEST-320 are displayed in Figure 6, where we can get the conclusion that the AUC of our method is the largest too, which proves that our proposed stacking-based ensemble classifier method is more suitable for the identification of mouse's 4mCs than other classifiers.

3.3. Contrast with Extant Models on TEST-320. Here, we contrasted i4mC-EL with 4mCpred-EL and i4mC-Mouse on TEST-320 for the sake of further evaluating its performance. Table 6 displays these contrast experimental results, in which the data of 4mCpred-EL and i4mC-Mouse are from reference. From Table 6, we can see that i4mC-EL is superior to 4mcPred-EL and i4mC-Mouse in three indexes which are ACC, Sp, and MCC. Although the Sn of i4mC-Mouse is a little higher than that of our method, our method outperforms i4mC-Mouse in the other three indexes. All in all, i4mC-EL performs better than extant methods.

4. Conclusions

In the present paper, an ensemble learning model called i4mC-EL which was able to identify mouse's 4mC sites was designed. In the process of constructing i4mC-EL, to determine the optimal combination of feature encoding schemes and classifiers, we conducted abundant comparative experiments on dissimilar features and classifiers. Finally, we encoded DNA sequences with multifeatures combing *k*-mer and EIIP, then used two-stage stacked ensemble learning as classifier. We used BayesNet, NavieBayes Multinomial, LibSVM, and VotedPerceptron as base classifiers and Logistic as metaclassifier.

In addition, we contrasted i4mC-EL with existing models for the sake of proving its effectiveness. The results show that i4mC-EL is better than the existing models and has better generalization capability. In summary, i4mC-EL is effective in predicting the 4mC sites in the mouse genome, which helps us to understand the biochemical properties of 4mC.

We will use adaptive feature vectors to donate DNA sequences to optimize the feature encoding scheme [54, 55] in the future work. Furthermore, other improvements, encoding schemes, classifier algorithms, and intelligent computing models to identify 4mC sites will also be considered.

Data Availability

The datasets used during the present study are available from the corresponding author upon reasonable request, or can be downloaded from http://106.12.83.135:8080/i4mC-EL/.

Conflicts of Interest

The authors declare that they have no competing interests.

Acknowledgments

This work is supported by the Natural Science Foundation of Heilongjiang Province (LH2019F002), National Natural Science Foundation of China (61901103, 61671189), and the Fundamental Research Funds for the Central Universities (2572018BH05, 2572017CB33).

References

[1] H. Lv, F. Y. Dao, D. Zhang et al., "iDNA-MS: An integrated computational tool for detecting DNA modification sites in multiple genomes," *iScience*, vol. 23, no. 4, article 100991, 2020.

[2] S. M. Irene, S. Maria, M. M. Rosaria, and D. E. Maurizio, "DNA methylation 40 years later: its role in human health and disease," *Journal of Cellular Physiology*, vol. 204, no. 1, pp. 21–35, 2005.

[3] Y. Zuo, M. Song, H. Li et al., "Analysis of the epigenetic signature of cell reprogramming by computational DNA methylation profiles," *Current Bioinformatics*, vol. 15, no. 6, pp. 589–599, 2020.

[4] C. Ling and L. Groop, "Epigenetics: a molecular link between environmental factors and type 2 Diabetes," *Diabetes*, vol. 58, no. 12, pp. 2718–2725, 2009.

[5] Y. Zhang, C. Kou, S. Wang, and Y. Zhang, "Genome-wide differential-based analysis of the relationship between DNA methylation and gene expression in cancer," *Current Bioinformatics*, vol. 14, no. 8, pp. 783–792, 2019.

[6] W. Tang, S. Wan, Z. Yang, A. E. Teschendorff, and Q. Zou, "Tumor origin detection with tissue-specific miRNA and DNA methylation markers," *Bioinformatics*, vol. 34, no. 3, pp. 398–406, 2018.

[7] Y. Fu, G.-Z. Luo, K. Chen et al., "N^6-Methyldeoxyadenosine Marks Active Transcription Start Sites in *Chlamydomonas*," *Cell*, vol. 161, no. 4, pp. 879–892, 2015.

[8] K. Chen, B. S. Zhao, and C. He, "Nucleic acid modifications in regulation of gene expression," *Cell Chemical Biology*, vol. 23, no. 1, pp. 74–85, 2016.

[9] B. A. Flusberg, D. R. Webster, J. H. Lee et al., "Direct detection of DNA methylation during single-molecule, real-time sequencing," *Nature Methods*, vol. 7, no. 6, pp. 461–465, 2010.

[10] T. Zhu, J. Guan, H. Liu, and S. Zhou, "RMDB: an integrated database of single-cytosine-resolution DNA methylation in Oryza sativa," *Current Bioinformatics*, vol. 14, no. 6, pp. 524–531, 2019.

[11] Y. Huang, H.-T. Ren, Q. Zou, Y.-Q. Wang, J.-L. Zhang, and X.-L. Yu, "Computational identification and characterization of miRNAs and their target genes from five cyprinidae fishes," *Saudi Journal of Biological Sciences*, vol. 246, pp. 1126–1135, 2017.

[12] C. Wei, Y. Hui, F. Pengmian, D. Hui, and L. Hao, "iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties," *Bioinformatics*, vol. 33, no. 22, pp. 3518–3523, 2017.

[13] P. Ye, Y. Luan, K. Chen, Y. Liu, C. Xiao, and Z. Xie, "MethSMRT: an integrative database for DNA N6-methyladenine and N4-methylcytosine generated by single-molecular real-time sequencing," *Nucleic Acids Research*, vol. 45, no. D1, pp. D85–D89, 2017.

[14] H. Wenying, J. Cangzhi, and Z. Quan, "4mCPred: machine learning methods for DNA N4-methylcytosine sites prediction," *Bioinformatics*, vol. 35, no. 4, pp. 593–601, 2019.

[15] W. Leyi, L. Shasha, N. L. A. Eijy, S. Ran, and Z. Quan, "Exploring sequence-based features for the improved prediction of DNA N4-methylcytosine sites in multiple species," *Bioinformatics*, vol. 35, no. 8, pp. 1326–1333, 2019.

[16] W. Leyi, S. Ran, L. Shasha et al., "Iterative feature representations improve N4-methylcytosine site prediction," *Bioinformatics*, vol. 35, no. 23, pp. 4930–4937, 2019.

[17] B. Manavalan, S. Basith, T. H. Shin, L. Wei, and G. Lee, "Meta-4mCpred: a sequence-based meta-predictor for accurate DNA 4mC site prediction using effective feature representation," *Molecular Therapy - Nucleic Acids*, vol. 16, pp. 733–744, 2019.

[18] M. M. Hasan, B. Manavalan, M. S. Khatun, and H. Kurata, "i4mC-ROSE, a bioinformatics tool for the identification of DNA N4-methylcytosine sites in the Rosaceae genome," *International Journal of Biological Macromolecules*, vol. 157, pp. 752–758, 2020.

[19] Z.-Y. Liu, J.-F. Xing, W. Chen et al., "MDR: an integrative DNA N6-methyladenine and N4-methylcytosine modification database for Rosaceae," *Horticulture Research*, vol. 6, no. 1, 2019.

[20] B. Manavalan, S. Basith, T. H. Shin, D. Y. Lee, L. Wei, and G. Lee, "4mCpred-EL: An Ensemble Learning Framework for Identification of DNA N4-Methylcytosine Sites in the Mouse Genome," *Cell*, vol. 8, no. 11, p. 1332, 2019.

[21] M. M. Hasan, B. Manavalan, W. Shoombuatong, M. S. Khatun, and H. Kurata, "i4mC-mouse: improved identification of DNA N4-methylcytosine sites in the mouse genome using multiple encoding schemes," *Computational and Structural Biotechnology Journal*, vol. 18, pp. 906–912, 2020.

[22] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, "CD-HIT: accelerated for clustering the next-generation sequencing data," *Bioinformatics*, vol. 28, no. 23, pp. 3150–3152, 2012.

[23] Q. Zou, G. Lin, X. Jiang, X. Liu, and X. Zeng, "Sequence clustering in bioinformatics: an empirical study," *Briefings in Bioinformatics*, vol. 21, no. 1, pp. 1–10, 2020.

[24] Y. Zhang and Q. Zou, "PPTPP: a novel therapeutic peptide prediction method using physicochemical property encoding and adaptive feature representation learning," *Bioinformatics*, vol. 36, no. 13, pp. 3982–3987, 2020.

[25] J. Zhang and B. Liu, "A review on the recent developments of sequence-based protein feature extraction methods," *Current Bioinformatics*, vol. 14, no. 3, pp. 190–199, 2019.

[26] Q. Zou, P. Xing, L. Wei, and B. Liu, "Gene2vec: gene subsequence embedding for prediction of mammalianN6-methyladenosine sites from mRNA," *RNA*, vol. 25, no. 2, pp. 205–218, 2019.

[27] B. Liu, X. Gao, and H. Zhang, "BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA, and protein sequences at sequence level and residue level based on machine learning approaches," *Nucleic Acids Research*, vol. 47, no. 20, article e127, 2019.

[28] J. Shao, K. Yan, and B. Liu, "FoldRec-C2C: protein fold recognition by combining cluster-to-cluster model and protein similarity network," *Briefings in Bioinformatics*, vol. 22, no. 3, 2021.

[29] Y.-J. Tang, Y.-H. Pang, and B. Liu, "IDP-Seq2Seq: identification of intrinsically disordered regions based on sequence to sequence learning," *Bioinformaitcs*, vol. 36, no. 21, pp. 5177–5186, 2021.

[30] B. Liu, H. Wu, and K.-C. Chou, "Pse-in-One 2.0: an improved package of web servers for generating various modes of pseudo components of DNA, RNA, and protein sequences," *Natural Science*, vol. 9, no. 4, pp. 67–91, 2017.

[31] C.-C. Li and B. Liu, "MotifCNN-fold: protein fold recognition based on fold-specific features extracted by motif-based convolutional neural networks," *Briefings in Bioinformatics*, vol. 21, no. 6, pp. 2133–2141, 2020.

[32] Z. Y. Zhang, Y. H. Yang, H. Ding, D. Wang, W. Chen, and H. Lin, "Design powerful predictor for mRNA subcellular location prediction in Homo sapiens," *Briefings in Bioinformatics*, vol. 22, no. 1, pp. 526–535, 2021.

[33] H. Yang, W. Yang, F. Y. Dao et al., "A comparison and assessment of computational method for identifying recombination hotspots in Saccharomyces cerevisiae," *Briefings in Bioinformatics*, vol. 21, no. 5, pp. 1568–1580, 2020.

[34] A. S. Nair and S. S. Pillai, "A coding measure scheme employing electron-ion interaction pseudopotential (EIIP)," *Bioinformation*, vol. 1, no. 6, 2006.

[35] H. Zhu, X. Du, and Y. Yao, "ConvsPPIS: identifying protein-protein interaction sites by an ensemble convolutional neural network with feature graph," *Current Bioinformatics*, vol. 15, no. 4, pp. 368–378, 2020.

[36] N. Sultana, N. Sharma, K. P. Sharma, and S. Verma, "A sequential ensemble model for communicable disease forecasting," *Current Bioinformatics*, vol. 15, no. 4, pp. 309–317, 2020.

[37] J. Li, L. Wei, F. Guo, and Q. Zou, "EP3: An ensemble predictor that accurately identifies type III secreted effectors," *Briefings in Bioinformatics*, vol. 22, no. 2, pp. 1918–1928, 2020.

[38] S. Wan, Y. Duan, and Q. Zou, "HPSLPred: an ensemble multi-label classifier for human protein subcellular location prediction with imbalanced source," *Proteomics*, vol. 17, no. 17-18, p. 1700262, 2017.

[39] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian Network Classifiers," *Machine Learning*, vol. 29, no. 2/3, pp. 131–163, 1997.

[40] Y. Freund and R. E. Schapire, "Large margin classification using the perceptron algorithm," *Machine Learning*, vol. 37, no. 3, 1999.

[41] M. C. Wang, "A prior-valued estimator applied to multinomial classification," *Communications in Statistics - Theory and Methods*, vol. 15, no. 2, pp. 405–427, 1986.

[42] C.-C. Chang and C.-J. Lin, "LIBSVM," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1–27, 2011.

[43] Z. B. Lv, D. H. Wang, H. Ding, B. N. Zhong, and L. Xu, "Escherichia Coli DNA N-4-methycytosine site prediction accuracy improved by light gradient boosting machine feature selection technology," *IEEE Access*, vol. 8, pp. 14851–14859, 2020.

[44] J. Wang, Y. Shi, X. Wang, and H. Chang, "A drug target interaction prediction based on LINE-RF learning," *Current Bioinformatics*, vol. 15, no. 7, pp. 750–757, 2020.

[45] T. Smolarczyk, I. Roterman-Konieczna, and K. Stapor, "Protein secondary structure prediction: a review of progress and directions," *Current Bioinformatics*, vol. 15, no. 2, pp. 90–107, 2020.

[46] Y. Liu, X.-H. Ouyang, Z.-X. Xiao, L. Zhang, and Y. Cao, "A review on the methods of peptide-MHC binding prediction," *Current Bioinformatics*, vol. 15, no. 8, pp. 878–888, 2020.

[47] H. Huang and X. Gong, "A review of protein inter-residue distance prediction," *Current Bioinformatics*, vol. 15, no. 8, pp. 821–830, 2020.

[48] J. Shao and B. Liu, "ProtFold-DFG: protein fold recognition by combining directed fusion graph and PageRank algorithm," *Briefings in Bioinformatics*, vol. 22, no. 3, 2021.

[49] Y. Pang and B. Liu, "SelfAT-fold: protein fold recognition based on residue-based and motif-based self-attention networks," *IEEE/ACM Transactions on Compuational Biology and Bioinformatics*, p. 1, 2020.

[50] H. Wang, Y. Ding, J. Tang, and F. Guo, "Identification of membrane protein types via multivariate information fusion with Hilbert-Schmidt independence criterion," *Neurocomputing*, vol. 383, pp. 257–269, 2020.

[51] Y. Ding, J. Tang, and F. Guo, "Identification of drug-target interactions via dual Laplacian regularized least squares with multiple kernel fusion," *Knowledge-Based Systems*, vol. 204, 2020.

[52] Y. Ding, J. Tang, and F. Guo, "Identification of drug-target interactions via fuzzy bipartite local model," *Neural Computing & Applications*, vol. 32, pp. 10303–10319, 2020.

[53] D. Zhang, H.-D. Chen, H. Zulfiqar et al., "iBLP: an XGBoost-based predictor for identifying bioluminescent proteins," *Computational and Mathematical Methods in Medicine*, vol. 2021, Article ID 6664362, 15 pages, 2021.

[54] D. Wang, Z. Zhang, Y. Jiang et al., "DM3Loc: multi-label mRNA subcellular localization prediction and analysis based on multi-head self-attention mechanism," *Nucleic Acids Research*, vol. 49, no. 8, p. e46, 2021.

[55] F. Y. Dao, H. Lv, D. Zhang, Z. M. Zhang, L. Liu, and H. Lin, "DeepYY1: a deep learning approach to identify YY1-mediated chromatin loops," *Briefings in Bioinformatics*, 2020.

Hindawi
BioMed Research International
Volume 2021, Article ID 9923112, 10 pages
https://doi.org/10.1155/2021/9923112

Research Article

LSTMCNNsucc: A Bidirectional LSTM and CNN-Based Deep Learning Method for Predicting Lysine Succinylation Sites

Guohua Huang,[1] Qingfeng Shen,[1] Guiyang Zhang,[1] Pan Wang,[1] and Zu-Guo Yu[2]

[1] *School of Information Engineering, Shaoyang University, Shaoyang 42200, China*
[2] *Key Laboratory of Intelligent Computing and Information Processing of Ministry of Education and Hunan Key Laboratory for Computation and Simulation in Science and Engineering, Xiangtan University, Xiangtan, Hunan 411105, China*

Correspondence should be addressed to Guohua Huang; guohuahhn@163.com

Received 15 March 2021; Revised 25 April 2021; Accepted 3 May 2021; Published 29 May 2021

Academic Editor: Quan Zou

Lysine succinylation is a typical protein post-translational modification and plays a crucial role of regulation in the cellular process. Identifying succinylation sites is fundamental to explore its functions. Although many computational methods were developed to deal with this challenge, few considered semantic relationship between residues. We combined long short-term memory (LSTM) and convolutional neural network (CNN) into a deep learning method for predicting succinylation site. The proposed method obtained a Matthews correlation coefficient of 0.2508 on the independent test, outperforming state of the art methods. We also performed the enrichment analysis of succinylation proteins. The results showed that functions of succinylation were conserved across species but differed to a certain extent with species. On basis of the proposed method, we developed a user-friendly web server for predicting succinylation sites.

1. Introduction

Protein post-translational modification (PTM) refers to the chemical interaction occurring prior to protein biosynthesis and after mRNAs are translated into polypeptide chains. PTM has different categories and is very prevalent in the cells. More than 450 categories of PTMs were discovered to date, such as phosphorylation, methylation, and acetylation [1–3]. PTM increases diversity of protein structures and functions, viewed as one of most regulating mechanisms in the cellular process. Lysine succinylation is a type of protein TPMs, in which a succinyl group (-CO-CH2-CH2-CO2H) is attached to lysine residue of proteins [4]. Succinylation is reversible, dynamic, and evolutionarily conserved, widely existing in the prokaryote and the eukaryotes cells [5, 6]. The succinylation of proteins induces shift in the charge and the structural alteration and thus would yield effects on functions of proteins [6]. Growing evidences also showed aberrant succinylations were involved in the pathogenesis of some diseases including cancers [7], metabolism disease [8, 9], and nervous system diseases [10]. Thus, identifying succinylation sites and understanding its mechanism are crucial to develop drugs for related diseases.

Identifying succinylation sites has two main routes: experimental and computational methods. The experimental methods were represented by mass spectrometry, which contributed to the validation of succinylation and collection of first-hand data. On the other hand, the experimental methods are labor-intensive and time-consuming without assist of the computational methods. The computational methods are based on data yielded by the experimental methods and build machine learning-based models to predict new succinylations. Therefore, identifying succinylation is a cyclic iterative process from experiment to computation and again from computation to experiment. We focused on the computational methods to predict succinylation. In the past decades, more than ten computational methods have been developed for identifying succinylation [11–29]. Most of these computational methods extracted features directly from protein sequences, which were subsequently used for training model. For example, Zhao et al. [11] used the autocorrelation functions, the group weight-based encoding, the normalized van der Waals volume, and the position weight amino acid composition. Kao et al. [25] exploited the amino acid composition and informative *k*-spaced amino acid pairs. Xu et al. [12] and Jia et al. [13, 19] employed pseudo amino acid composition. Dehzangi et al. [23] exploited the structure information. Hasan et al. [28] compared 12 types of feature as well as two learning methods: random forest and support vector machine for succinylation prediction. Different features have different performance with species. So does the learning methods. The best performance was no more than 0.83 AUC (area under receiver operating characteristic curve) for independent test. Like sentences of language, the protein sequences should have semantic. However, all the methods above failed to seize semantic relationship hidden among residues. Thapa et al. [29] presented a convolutional neural network- (CNN-) based deep learning method DeepSuccinylSite for predicting succinylation. Different from traditional methods, the DeepSuccinylSite exploited word embedding which translated word into vector, which was an extensively used method in the field of natural language process. The CNN is a widely used method to extract local features especially in the field of image processing. Inspired by the DeepSuccinylSite and loss of semantic relationship between residues, we fused long short-term memory (LSTM) and CNN into a deep learning method for succinylation prediction.

2. Data

All the succinylated proteins were downloaded from the PLMD (Protein Lysine Modifications Database) database which is dedicated to specifically collect protein lysine modification [30–32]. The PLMD has evolved to version 3.0, housing 284780 modification events in 53501 proteins for 20 types of lysine modification. We extracted 6377 proteins containing 18593 succinylation sites. To remove dependency of the proposed method on the homology, we used the software CD-Hit [33, 34] to cluster 6377 protein sequences. The sequence identify cut-off was set to 0.4, and we obtained 3560 protein sequences, of which any two kept sequence similarity less than 0.4. We randomly divided these 3560 proteins into the training and the testing samples at the ratio of training to testing 4 : 1, resulting in 712 testing and 2848 training sequences. For each protein sequence, we extracted all the peptides which centered the lysine residue with 15 amino acid residues in the downstream/upstream of it. For peptides less than 15 amino acid residues, we prefixed or suffixed "*X*" to supply it. The length of the amino acids is influential in prediction of succinylation sites. The short amino acid peptides would miss key information, while the long peptides would include noise or redundancy. Whether the short or the long peptides would cause low accuracy of prediction. Among methods to predict succinylation sites, iSuc-PseAAC [12] adopted the shorter peptides of 15 amino acid residues; SuccinSite2.0 [20] and GPSuc [22] adopted the longer 41 amino acid residues, while the most methods including SSEvol-Suc [23], Success [24], iSuc-PseOpt [13], pSuc-Lys [19], SucStruct [18], and PSSM-Suc [17] adopted peptides of 31 amino acid residues, which is of moderate length. Thus, we chose 31 amino acid residues as basic peptides. The peptides with succinylation sites were viewed positive samples and the others as negative ones. For the training set, the negative samples extremely outnumbered the positive ones. Unbalanced training set would cause preference to negative samples in the process of prediction. Therefore, we randomly sampled the same size of negative examples as the positive ones. Finally, the training set comprised 6512 positive and 6512 negative samples, while the testing set 1479 positive and 16457 negative samples. All the experimental data are freely available to scientific communities.

3. Method

As shown in Figure 1, the proposed deep learning network consisted mainly of embedding, 1D convolution, pooling, bidirectional LTSM, dropout, flatten, and fully connected layers. Peptides with 31 amino acid residues were entered to the embedding layer and were translated into vectors with shape of (31, 64). Then, two different network structures, respectively, took the embedding as input, and their outputs were concatenated as input to the fully connected layer. One structure was the convolution neural network, and another was the bidirectional LSTM neural network. The final output was a neuron representing probability of belonging to the positive sample. The parameters and the shape of output of each layers in the deep neural network are listed in Table 1. The total number of trainable parameters is 336,897.

3.1. Embedding Layer. Most machine learning-based methods for predicting protein post-translational modification generally required an encoding step which translated sequences into vector representation. For example, the frequently used encoding schemes included position specific scoring matrix [35], amino acid composition, composition of *k*-space amino acid pair [14], and pseudo amino acid composition [36]. For sequences of text, these methods might lose hidden semantic. The word2vec [37, 38] is different from the above methods, embedding word into vector. The word2vec is capable of extracting semantic of word. An interesting example is that King − Man + Woman = Queen. Similar to the word2vec [37, 38], the embedding layer translated words into vector representations. In this method, the character of amino acid corresponds to word.

3.2. 1D Convolution Layer. The convolution neural network (CNN) proposed by LeCun et al. [39, 40] is a feed forward network. Compared with the conventional neural network, the CNN has two notable properties: local connectivity and parameter sharing. The local connectivity lies that two neighboring layers are not fully connected but locally connected. That is to say, the neuron in a layer is not connected to all neurons in the neighboring layers. The CNN implemented the parameter sharing via the filter (also called convolution kernel). The filter slides on the image and convoluted with all sections in image. The filter is shared by the image. In the last ten years, many deep convolution neural networks such as AlexNet [41], VGG [42], GoogleNet [43], and ResNet [44] have been proposed and applied to computer vision. The

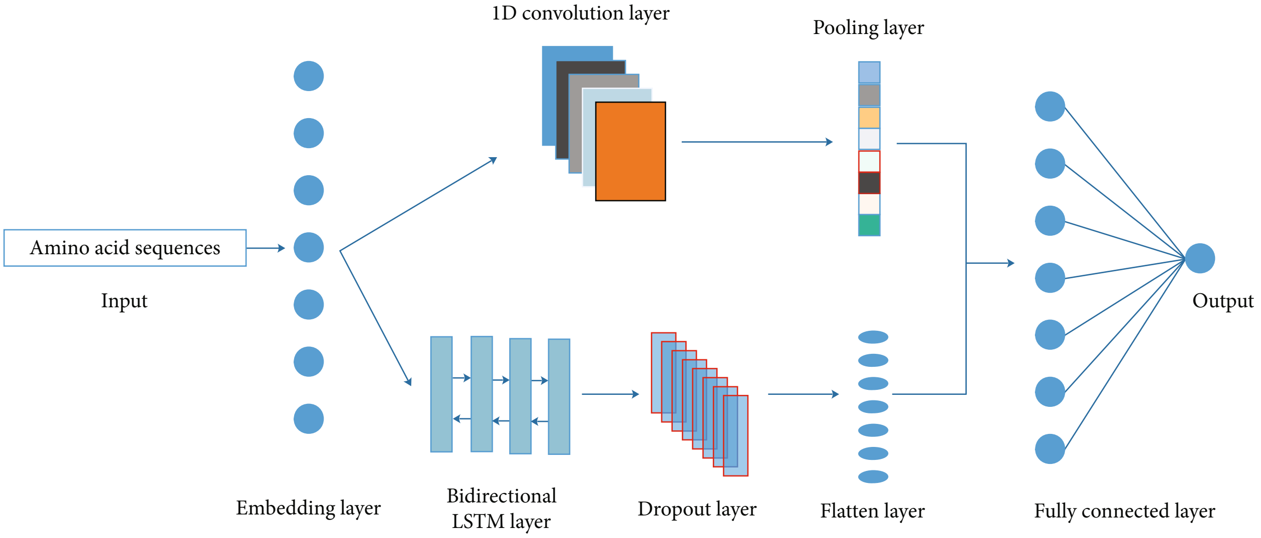


FIGURE 1: Flowchart of the proposed method.

TABLE 1: Number of parameters and shape of output in the LSTMCNNsucc.

Layers	Parameters	Output
Embedding	1472	(None, 31, 64)
Bidirectional LSTM	197632	(None, 31, 256)
Dropout	0	(None, 31, 256)
Flatten	0	(None, 7936)
1D convolution	10272	(None, 27, 32)
Pooling	0	(None, 32)
Dense (16)	127504	(None, 16)
Dense (1)	17	(None, 1)

CNN achieved significant advance in terms of classification error in comparison with the previous deep neural network. The convolution is of 1-dimension, 2-dimension, or more than 2 dimensions. Here, we used 1D convolution. Suppose a discrete sequence was $\alpha = [a_1, a_2, \cdots, a_n]$, and the convolution kernel was $\beta = [b_1, b_2, \cdots b_m]$. The 1D convolution product of α and β was expressed by

$$\alpha * \beta = \left[\sum_{i=1}^{m} a_{jd+i-1} b_i\right], j = 1, 2, \cdots, k, \tag{1}$$

where d was the stride of convolution and k was the length of the output sequence. Generally, k was the most integer less than or equal to $(n - m)/d + 1$.

3.3. Pooling Layer. The pooling operation firstly appeared in the AlexNet [41] and is increasingly becoming one of components of the deep CNN architecture. The pooling operation has such categories as max pooling, min pooling, and mean pooling. The role of pooling operation included removal of redundancy information and reduction of overfitting. Here, we used the max pooling operation. Given an n-channel input $A = (a_{i,j,k})$, the max pooling operation was defined by

$$\max_j \left\{ a_{i,j,k} \right\}. \tag{2}$$

3.4. Bidirectional LSTM Layer. Recurrent neural network (RNN) [45, 46] is a different framework of neural network from multiple layer perception. The RNN shares weights and is especially suitable to the field of sequence analysis such as language translation and semantic understanding. An unfolded RNN model was shown in Figure 2(a). The hidden state H_t at the time step t was not only dependent on the current input but also on the previous hidden state, which was computed by

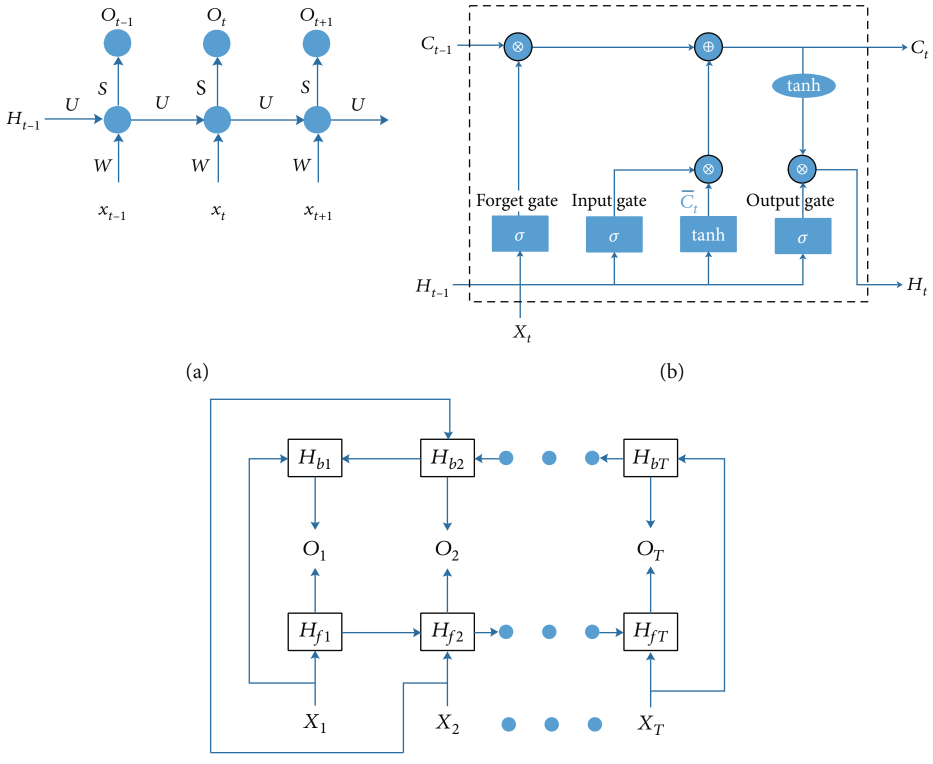
$$H_t = f(X_t W + H_{t-1} U + \alpha), \tag{3}$$

where f was an activation function and α was a bias. The output O_t at the time step t was computed by

$$O_t = g(H_t S + \beta), \tag{4}$$

where g was also an activation function and β was a bias. For long sequences, there was a fatal question with the RNN, i.e., vanishing gradient. Among all the solutions to the vanishing gradient, the LSTM [47] is one of the better. The LSTM contains a candidate memory cell and three gates: forget gate, input gate, and output gate, as shown in Figure 2(b). The forget gate F_t, the input gate I_t, and the output gate P_t at the time step t were computed, respectively, by

$$\begin{aligned} F_t &= \sigma\left(X_t W_{x,f} + H_{t-1} W_{h,f} + b_f\right), \\ I_t &= \sigma\left(X_t W_{x,i} + H_{t-1} W_{h,i} + b_i\right), \\ P_t &= \sigma\left(X_t W_{x,o} + H_{t-1} W_{h,o} + b_o\right), \end{aligned} \tag{5}$$



(a)

(b)

(c)

FIGURE 2: The structure of neural networks: (a) for RNN, (b) for LSTM, and (c) for directional LSTM.

TABLE 2: Comparison with state of the art methods.

Method	SN	SP	ACC	MCC
LSTMCNNsucc	0.5916	0.7957	0.7789	0.2508
SuccinSite [15]	0.3977	0.8635	0.8272	0.1925
iSuc-PseAAC [12]	0.1258	0.8929	0.8296	0.0165
DeepSuccinylSite [29]	0.7438	0.6879	0.6923	0.2438

where $W_{x,f}$ and $W_{h,f}$ were weights of the LSTM from input to forget gate and from the hidden state to the forget gate, respectively. $W_{x,i}$ and $W_{h,i}$ were link weights from input to input gate and from the hidden state to the input gate, respectively. $W_{x,o}$ and $W_{h,o}$ were link weights from input to output gate and from the hidden state to the output gate, respectively. b_f, b_i, and b_o were the bias of the forget and the input and the output gate, respectively. σ was the activation function. The candidate memory cell was calculated by

$$\bar{C}_t = \tanh\left(X_t W_{x,c} + H_{t-1} W_{h,c} + b_c\right), \tag{6}$$

where $W_{x,c}$ and $W_{h,c}$ were weights of the LSTM from input to the candidate memory and from the hidden state to the candidate memory, respectively, and b_c was the bias. The memory cell at the time step t was computed by

$$C_t = F_t \otimes C_{t-1} + I_t \otimes \bar{C}_t, \tag{7}$$

where $\otimes$ was defined as element-wise multiplication. The hidden state was updated by

$$H_t = I_t \otimes \tanh\left(C_t\right). \tag{8}$$

The previous RNN was forward. The output at the time step t was only dependent on the preceding inputs and the hidden state. In fact, the output might be relevant to the latter input and the hidden state. Schuster et al. [48] proposed a bidirectional RNN to model this relationship, showed in Figure 2(c). The forward hidden state at the time step t was computed by

$$H_t^f = \sigma\left(X_t W^f_{x,h} + H^f_{t-1} W^f_{h,h} + b^f_h\right), \tag{9}$$

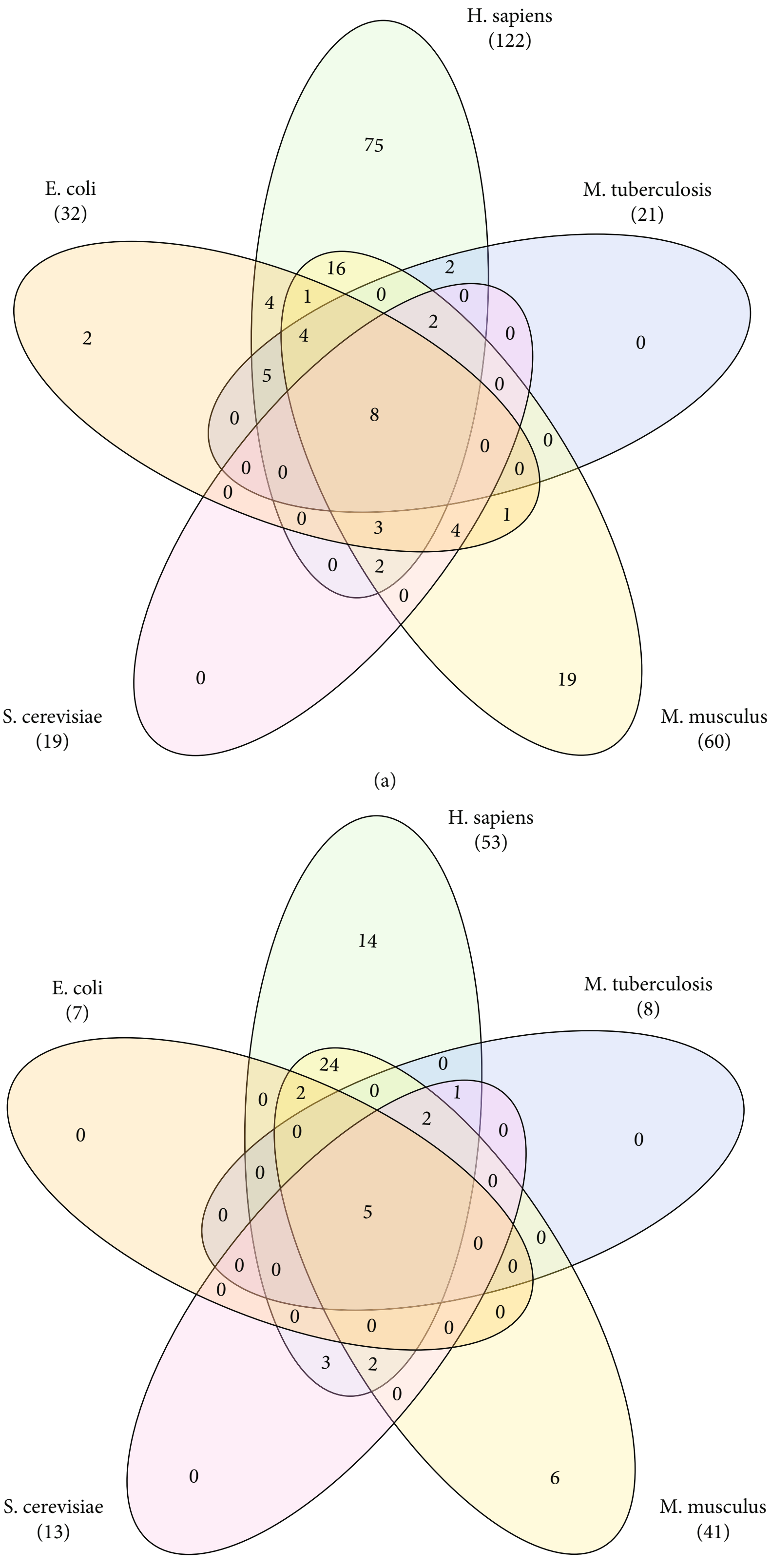
while the backward hidden state was computed by

$$H_t^b = \sigma\left(X_t W^b_{x,h} + H^b_{t+1} W^b_{h,h} + b^b_h\right). \tag{10}$$

The output at the time step t was computed by

$$O_t = \left[H_t^f, H_t^b\right] W_{h,o} + b_o. \tag{11}$$

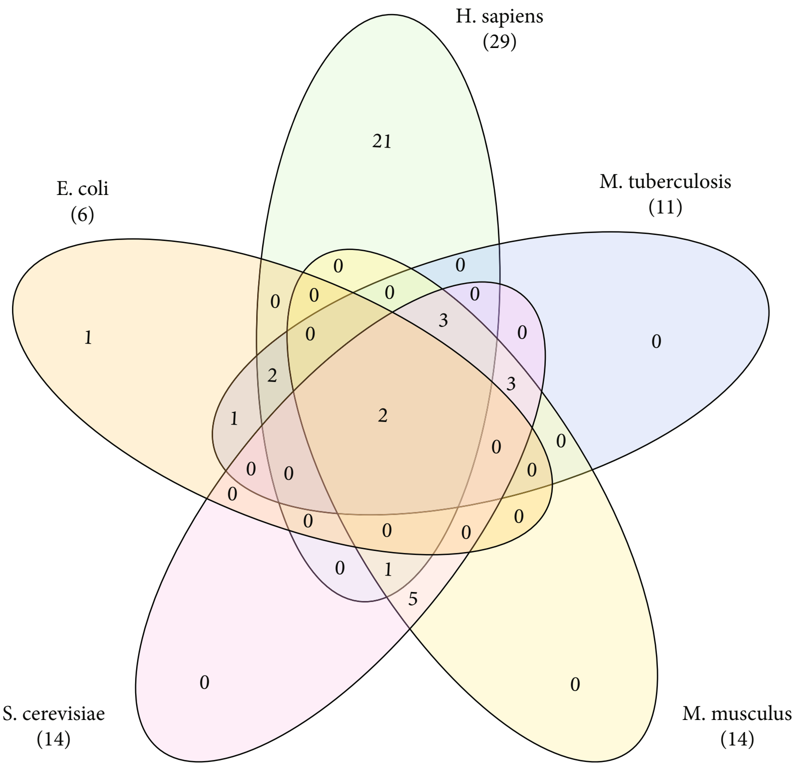
3.5. Dropout Layer. The deep neural network is prone to lead to overfitting when the number of training samples was too less. To deal with this issue, Hinton et al. [49] proposed the dropout concept. Due to its effect and efficiency, the dropout is increasingly becoming the frequently used trick in the deep



(a)

(b)

Figure 3: Continued.



(c)

Figure 3: The numbers of shared terms (a) for biological process, (b) cellular component, and (c) molecular function.

learning area [41, 50–53]. The neurons were dropped out at a certain rate of dropout, and parameters of only preserved neurons were updated in the training stage, while all the neurons were used in the predicting stage.

3.6. Flatten Layer and Fully Connected Layer. The role of flatten layer was only to convert the data into one-dimension and then facilitated connection of the fully connected layer. No parameters were trainable in the flatten layer. The fully connected layer was similar to hidden layer in the MLP, each neuron connected to the neurons in the preceding layer.

4. Metrics

We adopted to evaluate the predicted result these frequently used metrics in the binary classification questions such as sensitivity (SN), specificity (SP), accuracy (ACC), and Matthews correlation coefficient (MCC), which were defined by

$$\begin{aligned} \mathrm{SN} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \\ \mathrm{SP} &= \frac{\mathrm{TN}}{\mathrm{FP}+\mathrm{TN}}, \\ \mathrm{ACC} &= \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FN}+\mathrm{FP}+\mathrm{TN}}, \\ \mathrm{MCC} &= \frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FN})(\mathrm{TP}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})}}, \end{aligned} \tag{12}$$

where TP and TN were defined as numbers of the true positive and the true negative samples, respectively, FP and FN, respectively, as numbers of the false positive and the false negative samples in the prediction. SN reflected the accuracy of the correctly predicted positive samples, SP accuracy of the correctly predicted negative samples, and ACC the average accuracy of the correctly predicted samples. SN, SP, and ACC ranged from 0 to 1, larger meaning better performance. MCC was Matthews correlation coefficient, representing correlation between the true class and the predicted class. MCC ranged from -1 to 1. 1 meant perfect prediction, 0 random prediction, and -1 meant that the prediction was completely opposite to the true.

5. Results

Table 2 showed the predicting performance of the trained model on the 712 testing sequences. Although more than ten approaches or tools for predicting succinylation have been proposed in the past ten years, either they did not provide online predicting server or the web server could not work. We compared the proposed method to three methods whose web predicting server still can work [28]: SuccinSite [15], iSuc-PseAAC [12], and DeepSuccinylSite [29]. 712 testing sequences were used to examine three approaches. Among 712 testing sequences, at least 225 sequences repeated in the training set of the SuccinSite, and at least 223 repeated in the training set of DeepSuccinylSite. These minus 225 sequences were used to examine the SuccinSite and these minus 223 sequences to test the DeepSuccinylSite. iSuc-PseAAC [12] obtained best SP and

Table 3: Significant KEGG pathway terms.

Species	KEGG terms	Benjamini
E. coli	Metabolic pathways	3.30*E*-08
	Biosynthesis of amino acids	1.00*E*-06
	Biosynthesis of secondary metabolites	2.40*E*-04
	Biosynthesis of antibiotics	7.40*E*-04
	Lysine biosynthesis	3.30*E*-03
H. sapiens	Biosynthesis of antibiotics	3.70*E*-10
	Metabolic pathways	2.80*E*-09
	Ribosome	3.40*E*-08
	Valine, leucine, and isoleucine degradation	1.30*E*-06
	Carbon metabolism	6.20*E*-06
	Oxidative phosphorylation	1.10*E*-05
	Parkinson's disease	2.60*E*-05
	Citrate cycle (TCA cycle)	1.00*E*-04
	Huntington's disease	4.10*E*-04
	Alzheimer's disease	7.80*E*-04
	Aminoacyl-tRNA biosynthesis	1.00*E*-03
	Butanoate metabolism	3.40*E*-03
	Proteasome	8.20*E*-03
M. musculus	Metabolic pathways	6.20*E*-26
	Parkinson's disease	8.50*E*-11
	Oxidative phosphorylation	3.40*E*-10
	Nonalcoholic fatty liver disease (NAFLD)	1.00*E*-09
	Huntington's disease	2.80*E*-09
	Alzheimer's disease	1.40*E*-08
	Ribosome	3.30*E*-07
	Peroxisome	1.80*E*-06
	Glycine, serine, and threonine metabolism	1.50*E*-05
	Pyruvate metabolism	9.00*E*-05
	Propanoate metabolism	2.40*E*-04
	Valine, leucine, and isoleucine degradation	1.90*E*-03
	Glyoxylate and dicarboxylate metabolism	3.10*E*-03
	Biosynthesis of antibiotics	5.60*E*-03
M. tuberculosis	Metabolic pathways	1.00*E*-04
	Microbial metabolism in diverse environments	2.50*E*-04
	Biosynthesis of antibiotics	4.40*E*-04
	Biosynthesis of secondary metabolites	1.00*E*-02
	Propanoate metabolism	1.00*E*-02
S. cerevisiae	Metabolic pathways	5.20*E*-05
	Biosynthesis of amino acids	3.30*E*-04
	2-Oxocarboxylic acid metabolism	7.90*E*-04
	Biosynthesis of antibiotics	3.50*E*-03
	Oxidative phosphorylation	3.50*E*-03

best ACC but worst SN and worst MCC. The SuccinSite [15] reached better SP and better ACC but worse MCC and worse SN. The iSuc-PseAAC [12] and the SuccinSite [15] were in favor of predicting the negative samples. The DeepSuccinylSite [29] was better than the LSTMCNNsucc in terms of SN, worse than the LSTMCNNsucc in terms of sp. The overall performance of the LSTMCNNsucc was slightly better than that of the DeepSuccinylSite.

5.1. Functional Analysis. We used the statistical overrepresentation test of gene list analysis in the PANTHER classification system [54, 55] to perform function enrichment analysis of the succinylated proteins. The significant biological process, the molecular function, and the cellular component terms (p value≤0.01) were listed in the supplementary materials 1 and 2. For five species, Escherichia coli (E. coli), Homo sapiens (H. sapiens), Mus musculus (M. musculus), Mycobacterium tuberculosis (M. tuberculosis), and Saccharomyces cerevisiae (S. cerevisiae), they shared some common functions, but they had also own specific functions. The numbers of shared terms among five species are shown in Figure 3. H. sapiens and M. musculus shared 36 significant biological process terms and 35 cellular component terms, much more than the numbers of shared terms between any other two species (Figures 3(a) and 3(b)). Five species shared eight biological process GO terms: "biosynthetic process (GO:0009058)", "carboxylic acid metabolic process (GO:0019752)", "organic acid metabolic process (GO:0006082)", "organic substance biosynthetic process (GO:1901576)", "organonitrogen compound biosynthetic process (GO:1901566)", "organonitrogen compound metabolic process (GO:1901564)", "oxoacid metabolic process (GO:0043436)", and "small molecule metabolic process (GO:0044281)"; 5 cellular component GO terms: "cytoplasm (GO:0005737)", "cytoplasmic part (GO:0044444)", "cytosol (GO:0005829)", "intracellular (GO:0005622)", and "intracellular part (GO:0044424)"; and two molecular function GO terms: "catalytic activity (GO:0003824)", and "molecular_ function (GO:0003674)". H. sapiens had much more own specific functions than other species, with 75 specific biological process GO terms, 14 GO cellular component terms, and 21 molecular function GO terms. No specific functions existed in both M. tuberculosis and S. cerevisiae whether for biological process, cellular component, or molecular functions.

We also performed enrichment analysis of Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway by functional annotation in the DAVID tool [56, 57] to investigate in which pathway the succinylated proteins were involved. The statistically significant KEGG terms (Benjamini ≤ 0.01) are listed in Table 3. Different species were involved in some identical pathways. For example, both metabolic pathways and biosynthesis of antibiotics were enriched in the succinylated proteins for five species, implying the universal role of succinylation. On the other hand, different pathways were involved in different species. H. sapiens and M. musculus shared more pathway and had more pathways than other three species, implying species-specific role of the succinylation.

5.2. LSTMCNNsucc Web Server. We built a web server of the proposed LSTMCNNsucc at http://8.129.111.5/. Users either directly input protein sequences in a fasta format or upload a file of fasta format to perform prediction. When both protein sequences and files were submitted, the file was given to priority of prediction.

6. Conclusion

We presented a bidirectional LSTM and CNN-based deep learning method for predicting succinylation sites. The method absorbed semantic relationship hidden in the succinylation sequences, outperforming state-of-the-art method. The functions of succinylation proteins were conserved to a certain extent across species but also were species-specific. We also implemented the proposed method into a user-friendly web server which is available at http://8.129.111.5/.

Data Availability

The experimental succinylation and nonsuccinylation sites used to support the findings of this study have been deposited in the website http://8.129.111.5/ and are freely available to all scientific communities.

Conflicts of Interest

The authors declare that no competing interest exists.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (11871061, 61672356), by the Scientific Research Fund of Hunan Provincial Education Department (18A253), by the open project of Hunan Key Laboratory for Computation and Simulation in Science and Engineering (2019LCESE03), and by Shaoyang University Graduate Research Innovation Project (CX2020SY060).

Supplementary Materials

Supplementary 1. KEGG enrichmeent analysis of KEGG pathway for 5 species.

Supplementary 2. GO enrichment analysis of GO terms for 5 species.

References

[1] E. S. Witze, W. M. Old, K. A. Resing, and N. G. Ahn, "Mapping protein post-translational modifications with mass spectrometry," *Nature Methods*, vol. 4, no. 10, pp. 798–806, 2007.

[2] B. S. Sharma, "Post-translational modifications (PTMs), from a cancer perspective: an overview," *Oncogen Journal*, vol. 2, no. 3, p. 12, 2019.

[3] G. Huang and X. Li, "A review of computational identification of protein post-translational modifications," *Mini-Reviews in Organic Chemistry*, vol. 12, no. 6, pp. 468–480, 2015.

[4] Z. Zhang, M. Tan, Z. Xie, L. Dai, Y. Chen, and Y. Zhao, "Identification of lysine succinylation as a new post-translational modification," *Nature Chemical Biology*, vol. 7, no. 1, pp. 58–63, 2011.

[5] J. Du, Y. Zhou, X. Su et al., “Sirt5 is a NAD-dependent protein lysine demalonylase and desuccinylase,” *Science*, vol. 334, no. 6057, pp. 806–809, 2011.

[6] A. Sreedhar, E. K. Wiese, and T. Hitosugi, “Enzymatic and metabolic regulation of lysine succinylation,” *Genes & Diseases*, vol. 7, no. 2, pp. 166–171, 2020.

[7] Y. Xiangyun and N. Xiaomin, “Desuccinylation of pyruvate kinase M2 by SIRT5 contributes to antioxidant response and tumor growth,” *Oncotarget*, vol. 8, no. 4, pp. 6984–6993, 2017.

[8] M. Yang, Y. Wang, Y. Chen et al., “Succinylome analysis reveals the involvement of lysine succinylation in metabolism in pathogenic Mycobacterium tuberculosis,” *Molecular & Cellular Proteomics*, vol. 14, no. 4, pp. 796–811, 2015.

[9] Y. Yang and G. E. Gibson, “Succinylation links metabolism to protein functions,” *Neurochemical Research*, vol. 44, no. 10, pp. 2346–2359, 2019.

[10] G. E. Gibson, H. Xu, H. L. Chen, W. Chen, T. T. Denton, and S. Zhang, “Alpha-ketoglutarate dehydrogenase complex-dependent succinylation of proteins in neurons and neuronal cell lines,” *Journal of Neurochemistry*, vol. 134, no. 1, pp. 86–96, 2015.

[11] X. Zhao, Q. Ning, H. Chai, and Z. Ma, “Accurate in silico identification of protein succinylation sites using an iterative semi-supervised learning technique,” *Journal of Theoretical Biology*, vol. 374, pp. 60–65, 2015.

[12] Y. Xu, Y. X. Ding, J. Ding, Y. H. Lei, L. Y. Wu, and N. Y. Deng, “iSuc-PseAAC: predicting lysine succinylation in proteins by incorporating peptide position-specific propensity,” *Scientific Reports*, vol. 5, no. 1, article 10184, 2015.

[13] J. Jia, Z. Liu, X. Xiao, B. Liu, and K. C. Chou, “iSuc-PseOpt: identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset,” *Analytical Biochemistry*, vol. 497, pp. 48–56, 2016.

[14] H. D. Xu, S. P. Shi, P. P. Wen, and J. D. Qiu, “SuccFind: a novel succinylation sites online prediction tool via enhanced characteristic strategy,” *Bioinformatics*, vol. 31, no. 23, pp. 3748–3750, 2015.

[15] M. M. Hasan, S. Yang, Y. Zhou, and M. N. Mollah, “SuccinSite: a computational tool for the prediction of protein succinylation sites by exploiting the amino acid patterns and properties,” *Molecular BioSystems*, vol. 12, no. 3, pp. 786–795, 2016.

[16] H. Ai, R. Wu, L. Zhang et al., “pSuc-PseRat: predicting lysine succinylation in proteins by exploiting the ratios of sequence coupling and properties,” *Journal of Computational Biology*, vol. 24, no. 10, pp. 1050–1059, 2017.

[17] A. Dehzangi, Y. López, S. P. Lal et al., “PSSM-Suc: accurately predicting succinylation using position specific scoring matrix into bigram for feature extraction,” *Journal of Theoretical Biology*, vol. 425, pp. 97–102, 2017.

[18] Y. López, A. Dehzangi, S. P. Lal et al., “SucStruct: prediction of succinylated lysine residues by using structural properties of amino acids,” *Analytical Biochemistry*, vol. 527, pp. 24–32, 2017.

[19] J. Jia, Z. Liu, X. Xiao, B. Liu, and K. C. Chou, “pSuc-Lys: predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach,” *Journal of Theoretical Biology*, vol. 394, pp. 223–230, 2016.

[20] M. M. Hasan, M. S. Khatun, M. N. H. Mollah, C. Yong, and D. Guo, “A systematic identification of species-specific protein succinylation sites using joint element features information,” *International Journal of Nanomedicine*, vol. Volume 12, pp. 6303–6315, 2017.

[21] Q. Ning, X. Zhao, L. Bao, Z. Ma, and X. Zhao, “Detecting succinylation sites from protein sequences using ensemble support vector machine,” *BMC Bioinformatics*, vol. 19, no. 1, p. 237, 2018.

[22] A. G. de Brevern, M. M. Hasan, and H. Kurata, “GPSuc: Global Prediction of Generic and Species-specific Succinylation Sites by aggregating multiple sequence features,” *PLoS One*, vol. 13, no. 10, 2018.

[23] A. Dehzangi, Y. López, S. P. Lal et al., “Improving succinylation prediction accuracy by incorporating the secondary structure via helix, strand and coil, and evolutionary information from profile bigrams,” *PLoS One*, vol. 13, no. 2, article e0191900, 2018.

[24] Y. López, A. Sharma, A. Dehzangi et al., “Success: evolutionary and structural properties of amino acids prove effective for succinylation site prediction,” *BMC Genomics*, vol. 19, no. S1, p. 923, 2018.

[25] H.-J. Kao, V.-N. Nguyen, K.-Y. Huang, W.-C. Chang, and T.-Y. Lee, “SuccSite: incorporating amino acid composition and informative k-spaced amino acid pairs to identify protein succinylation sites,” *Genomics, Proteomics & Bioinformatics*, vol. 18, no. 2, pp. 208–219, 2020.

[26] K.-Y. Huang, J. B.-K. Hsu, and T.-Y. Lee, “Characterization and identification of lysine succinylation sites based on deep learning method,” *Scientific Reports*, vol. 9, no. 1, article 16175, 2019.

[27] L. Zhang, M. Liu, X. Qin, G. Liu, and H. Ding, “Succinylation Site Prediction Based on Protein Sequences Using the IFS-LightGBM (BO) Model,” *Computational and Mathematical Methods in Medicine*, vol. 2020, Article ID 8858489, 15 pages, 2020.

[28] M. M. Hasan, M. S. Khatun, and H. Kurata, “Large-scale assessment of bioinformatics tools for lysine succinylation sites,” *Cell*, vol. 8, no. 2, p. 95, 2019.

[29] N. Thapa, M. Chaudhari, S. McManus et al.DeepSuccinylSite: a deep learning based approach for protein succinylation site prediction,” *BMC Bioinformatics*, vol. 21, no. S3, p. 63, 2020.

[30] Z. Liu, J. Cao, X. Gao et al., “CPLA 1.0: an integrated database of protein lysine acetylation,” *Nucleic Acids Research*, vol. 39, suppl_1, pp. D1029–D1034, 2011.

[31] Z. Liu, Y. Wang, T. Gao et al., “CPLM: a database of protein lysine modifications,” *Nucleic Acids Research*, vol. 42, no. D1, pp. D531–D536, 2014.

[32] H. Xu, J. Zhou, S. Lin, W. Deng, Y. Zhang, and Y. Xue, “PLMD: an updated data resource of protein lysine modifications,” *Journal of Genetics and Genomics*, vol. 44, no. 5, pp. 243–250, 2017.

[33] Y. Huang, B. Niu, Y. Gao, L. Fu, and W. Li, “CD-HIT suite: a web server for clustering and comparing biological sequences,” *Bioinformatics*, vol. 26, no. 5, pp. 680–682, 2010.

[34] W. Li and A. Godzik, “Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences,” *Bioinformatics*, vol. 22, no. 13, pp. 1658-1659, 2006.

[35] G. Huang, L. Lu, K. Feng et al., “Prediction of S-nitrosylation modification sites based on kernel sparse representation classification and mRMR algorithm,” *BioMed Research International*, vol. 2014, Article ID 438341, 10 pages, 2014.

[36] Q. Xiang, K. Feng, B. Liao, Y. Liu, and G. Huang, “Prediction of lysine malonylation sites based on pseudo amino acid,”

Combinatorial Chemistry & High Throughput Screening, vol. 20, no. 7, pp. 622–628, 2017.

[37] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” *Advances in neural information processing systems*, pp. 3111–3119, Curran Associates Inc., 2013.

[38] T. Mikolov, K. Chen, G. Corrado, and D. JJapa, “Efficient estimation of word representations in vector space,” 2013, https://arxiv.org/abs/1301.3781.

[39] Y. LeCun, B. Boser, J. S. Denker et al., “Backpropagation applied to handwritten zip code recognition,” *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989.

[40] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[41] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[42] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, https://arxiv.org/abs/1409.1556.

[43] C. Szegedy, W. Liu, Y. Jia et al., “Going deeper with convolutions,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, 2015.

[44] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 778-779, 2016.

[45] B. A. Pearlmutter, “Learning state space trajectories in recurrent neural networks,” *Neural Computation*, vol. 1, no. 2, pp. 263–269, 1989.

[46] C. L. Giles, G. M. Kuhn, and R. J. Williams, “Dynamic recurrent neural networks: theory and applications,” *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 153–156, 1994.

[47] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[48] M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks,” *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[49] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, “Improving neural networks by preventing co-adaptation of feature detectors,” 2012, https://arxiv.org/abs/1207.0580.

[50] N. Srivastava, “Improving neural networks with dropout,” *University of Toronto*, vol. 182, no. 566, p. 7, 2013.

[51] X. Bouthillier, K. Konda, P. Vincent, and R. Memisevic, “Dropout as data augmentation,” 2015, https://arxiv.org/abs/1506.08700.

[52] P. Baldi and P. Sadowski, “The dropout learning algorithm,” *Artificial Intelligence*, vol. 210, pp. 78–122, 2014.

[53] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[54] H. Mi, A. Muruganujan, D. Ebert, X. Huang, and P. D. Thomas, “PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools,” *Nucleic Acids Research*, vol. 47, no. D1, pp. D419–D426, 2018.

[55] H. Mi, A. Muruganujan, X. Huang et al., “Protocol update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0),” *Nature Protocols*, vol. 14, no. 3, pp. 703–721, 2019.

[56] D. W. Huang, B. T. Sherman, and R. A. Lempicki, “Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists,” *Nucleic Acids Research*, vol. 37, no. 1, pp. 1–13, 2009.

[57] D. W. Huang, B. T. Sherman, and R. A. Lempicki, “Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources,” *Nature Protocols*, vol. 4, no. 1, pp. 44–57, 2009.

Hindawi
BioMed Research International
Volume 2021, Article ID 9933873, 12 pages
https://doi.org/10.1155/2021/9933873

Research Article

An Ensemble Learning-Based Method for Inferring Drug-Target Interactions Combining Protein Sequences and Drug Fingerprints

Zheng-Yang Zhao, Wen-Zhun Huang, Xin-Ke Zhan, 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
and Xin-Ke Zhan; przhanxinke@gmail.com

Received 15 March 2021; Revised 14 April 2021; Accepted 16 April 2021; Published 26 April 2021

Academic Editor: Quan Zou

Identifying the interactions of the drug-target is central to the cognate areas including drug discovery and drug reposition. Although the high-throughput biotechnologies have made tremendous progress, the indispensable clinical trials remain to be expensive, laborious, and intricate. Therefore, a convenient and reliable computer-aided method has become the focus on inferring drug-target interactions (DTIs). In this research, we propose a novel computational model integrating a pyramid histogram of oriented gradients (PHOG), Position-Specific Scoring Matrix (PSSM), and rotation forest (RF) classifier for identifying DTIs. Specifically, protein primary sequences are first converted into PSSMs to describe the potential biological evolution information. After that, PHOG is employed to mine the highly representative features of PSSM from multiple pyramid levels, and the complete describers of drug-target pairs are generated by combining the molecular substructure fingerprints and PHOG features. Finally, we feed the complete describers into the RF classifier for effective prediction. The experiments of 5-fold Cross-Validations (CV) yield mean accuracies of 88.96%, 86.37%, 82.88%, and 76.92% on four golden standard data sets (*enzyme*, *ion channel*, *G protein-coupled receptors* (*GPCRs*), and *nuclear receptor*, respectively). Moreover, the paper also conducts the state-of-art light gradient boosting machine (LGBM) and support vector machine (SVM) to further verify the performance of the proposed model. The experimental outcomes substantiate that the established model is feasible and reliable to predict DTIs. There is an excellent prospect that our model is capable of predicting DTIs as an efficient tool on a large scale.

1. Introduction

The identification of interacting drug-target pairs is of cardinal significance in pharmaceutical science. Previous development of genomics, protein engineering, and molecular biology dynamically helps researchers in finding the potential therapeutic drugs and explaining the by-effect of a trial. In past decades, the Food and Drug Administration (FDA) declared that the demand for new drugs is hard to meet due to the adverse clinical outcomes of some candidate drugs [1]. Classifying DTIs remains to be a critical step for better developing and applying novel molecule-targeted drugs. Previously, researchers utilized clinical experiments as the main approach to discover DTIs. Nevertheless, the traditional experiments are still cumbersome, costly, and time-consuming. Meanwhile, it also has to confront the contingency and inefficiency of the results. Therefore, novel computer-aided drug development (CADD) methods need to be advanced for effectively avoiding these drawbacks [2].

With the progress of protein primary sequence detection technologies and spectral techniques in determination of the chemical composition structure of drugs, the public database has had an explosive growth in size. These databases which provide multiple download formats comprehensively construct a reliable data platform for researchers. Different kinds of databases, such as the Therapeutic Target Database (TTD) [3], DrugBank [4], ChEMBL [5], and KEGG [6], collect the information of the protein primary structure, drug molecular structure, and drug-target pairs with known interactions to assist establishing the prediction model of DTIs. In the past years, researchers have made many achievements in predicting DTIs by combining traditional computing methods and bioinformatics. The most widespread applications are based on molecular docking, genome, and pharmacophore [7]. Molecular docking simulation is utilized to detect the optimal binding position between drug molecules and targets based on energy matching. This method also requires complete three-dimensional (3D) substructures of proteins, but they are hard to explore by Nuclear Magnetic Resonance (NMR), electron microscopy, and X-ray crystallography [8]. Pharmacophores are a characteristic element of drug-active molecules that play a pivotal role in the prediction of DTIs [9]. Researches suggested that the pharmacophore method can effectively inspect the multitarget drug design and reduce the blindness of screening. The difficulty of matching molecular pharmacophores is determined by the number of pharmacophore characteristics. In addition, whether the molecule can match the pharmacophore is also related to the conformations of the molecules [10]. When the conformation changes, the molecule will not match the existing pharmacophore model. Therefore, the establishment of the pharmacophore model is still not comprehensive for further bioassay. At the same time, this method does not take 3D structures of targets into account, which declines the accuracy of the pharmacophore model [11]. In general, it is exceedingly urgent to develop more robust and universal methods for the prediction of DTIs without a ligand and 3D target structure.

Up to now, many learning-based models are developed to detect potential DTIs. For instance, Ding et al. [12] developed a fuzzy bipartite local model (FBLM) based on fuzzy least square support vector machine and multiple kernel learning (MKL) for predicting DTIs. Specifically, MKL is employed to fuse multiple kernels of drugs and targets, and FBLM is adopted to infer the unknown DTIs. Krisztian et al. [13] utilized the Modified Linear Regression (MOLIERE) to predict the potential DTIs based on asymmetric loss models (ALM). Ye et al. [14] proposed a new prediction framework based on Adversarial Bayesian Personalized Ranking (AdvB). More specially, the latent factor matrices of drugs and targets are trained by partial order relationships. Then, the scores of inner products of factors are trained to predict DTIs. Maryam et al. [15] developed an effective model named the Coupled Tensor-Matrix Completion (CTMC) to repurpose drug molecules by constructing drug-drug and target-target tensors. Pliakos and Vens [16] proposed to address DTI prediction as a multioutput prediction task by learning ensembles of multioutput biclustering trees (eBICT) on reconstructed networks. An et al. [17] combined Weighted Extreme Learning Machine (WELM) and Speed Up Robot Features (SURF) to predict DTIs. Laarhoven et al. [18] proposed the Kronecker Regularized Least Square- (Kron-RLS-) based predictive models, which employed the Kronecker product to fuse drug and target feature spaces. Gönen [19] proposed a joint Bayesian formulation of projecting drug compounds and target proteins into a unified subspace, and this formulation combines dimensionality reduction, matrix factorization, and binary classification for predicting drug-target interaction networks. Zheng et al. [20] proposed a factor model, named Multiple Similarity Collaborative Matrix Factorization (MSCMF) which is an extension of weighted low-rank approximation for one-class collaborative filtering. In this model, drugs and proteins are projected onto low dimensional feature space, and the weights of low-rank matrix and similarity matrix are estimated by alternating least square method to predict DTIs.

In this study, we present a novel computational method which exploits protein primary sequence and molecular fingerprints of drug compounds. More specially, this model numerically characterizes different amino acids as PSSMs to carry biological evolution information. Then, the proposed model employs the PHOG approach to extract the 680-dimensional local features of PSSM from different pyramid levels. Finally, the RF classifier is employed to effectively predict DTIs based on the fusion which contains PHOG descriptors of PSSMs and drug fingerprints. This experiment also evaluates the prediction performance by conducting 5-fold Cross-Validation (CV) on *enzyme*, *ion channel*, *G protein-coupled receptors* (*GPCRs*), and *nuclear receptor* data sets. For the sake of verifying the reliability of the model, we also carried out the state-of-the-art LGBM and SVM on benchmark data sets. The overall results of the experiments illustrate that the established model is practicable in providing accurate candidates for clinical experiments by predicting DTIs. Figure 1 depicts the workflow of the proposed model.

2. Materials and Methods

2.1. Data Sets. In this paper, entire experiments were performed on benchmark data sets, viz., *enzyme*, *ion channel*, *GPCRs*, and *nuclear receptor*. All data sets originate from the databases of DrugBank [4], SuperTarget [21], BRENDA [22], and KEGG BRITE [6]. The statistical quantities of existing drugs are 445, 233, 210, and 54, respectively. The numbers of known proteins are 664, 95, 204, and 26, respectively. The counts of the DTIs which have been proven are 2926, 635, 1476, and 90, respectively. The number of known DTIs which were regarded as positive sample data set is 5127. Table 1 fully lists the statistical amounts of drugs, target proteins, and DTIs.

In this section, the bipartite graph is employed to display the DTI network. The nodes of the graph denote drugs and proteins, the edges which connect the nodes denote the relationships between drugs and targets. The interacting drug-target pairs are considered as positive samples; the others are regarded as negative samples in the sparse network. Taking the *ion channel* data set as an instance, there are 42840 (210 × 204) edges existing in the graph. The verified 1476 real drug-target interactions construct the positive sample set, and the residual 41364 (42840-1476) pairs represent the negative samples. It is obvious that there is a big quantity gap between positive samples and negative samples. For attaining sample balance, a downsampling algorithm is adopted in uncorrelated pairs to form the negative set which contains the same number of samples as the positive one. In consideration of the scale of the sparse network and the large ratio of

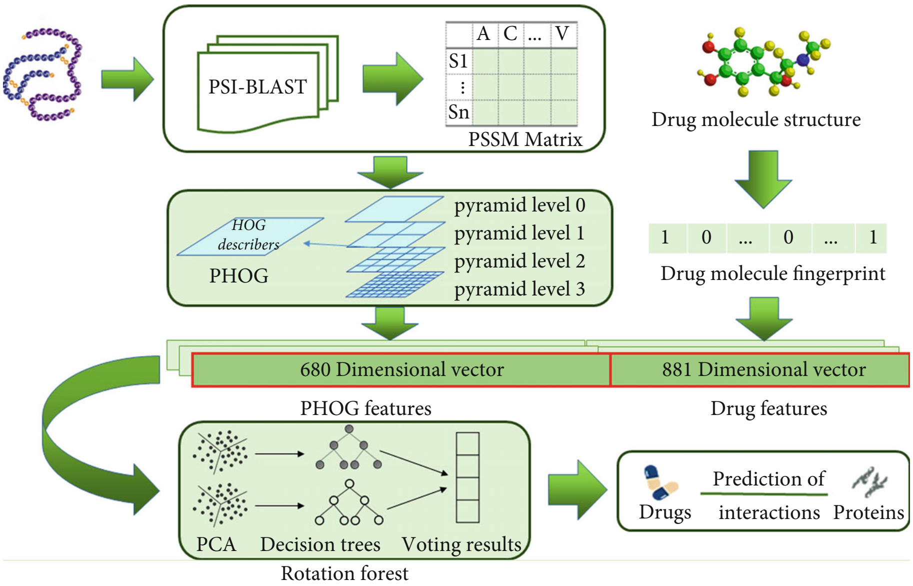


Figure 1: Workflow for the proposed model to predict DTIs.

Table 1: The statistical quantities for drugs, target proteins, and DTIs.

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

differences, the possibility that the drug-target pairs with real interactions are collected in the negative data set can be ignored. Therefore, the sample quantities of four negative data sets are 2926, 1476, 635, and 90, respectively.

2.2. Drug Substructure Characterization. In recent years, many physical and chemical properties are utilized to describe the drug compound information including geometry, topology, and quantum chemistry [23, 24]. At present, the researchers demonstrate that molecular fingerprints can effectively characterize the drug substructure. The fingerprints of structural bonds represent the drugs as Boolean substructure vectors by separating the drug molecular structure into a variety of segments. Although the molecule is sliced into individual segments, it still retains the entire structure information of the drug [25, 26]. These printers reduce the information loss and error accumulation in the process of description and screening. Specifically, the predefined dictionary which contains all substructures matches all fragments of the given drug molecule. If the fragment exists in the dictionary, the corresponding position in the carrier is set to 1; otherwise, it is set to 0. The complete fingerprint database provides an effective way to describe the molecular structure of drugs as binary fingerprint vectors. We utilized the chemical structure map from the PubChem system in the website https://pubchem.ncbi.nlm.nih.gov/, and the map contains 881 molecular substructures [27]. Hence, the feature describers of the drug molecular structure take the form of an 881-dimensional binary vector.

2.3. Position-Specific Scoring Matrix (PSSM). In general, researchers took many physicochemical approaches to numerically describe target proteins [28]. The effective descriptors will differentially convert proteins to enhance the performance of the classifier. Within the experiment, the Position-Specific Scoring Matrix (PSSM) is utilized to represent the biological evolution of proteins [29], and this matrix contains the probability information of 20 amino acids at each position in the original protein sequence. In the practical process, the Position-Specific Iterated Basic Local Alignment Search Tool (PSI-BLAST) is employed to generate the corresponding PSSM for different sorts of amino acids. The matrix is as follows:

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

where the PSSM is expressed as a matrix of $L \times 20$ and L denotes the length of the amino acid. $\ell_{i,j}$ denotes the evolutionary score that the i_{th} residue mutates into the j_{th} amino acid in the evolutionary process. The experiments also optimized the parameters of PSI-BLAST to obtain more reliable homologous sequences. In summary, parameter e which represents the noise of protein matching is assigned to 0.001, and the frequency of iterations is set to 3.

2.4. Pyramid Histogram of Oriented Gradients. The pyramid histogram of oriented gradients (PHOG) is a feature extraction method which describes the local features by counting the distribution of the gradient direction histogram from

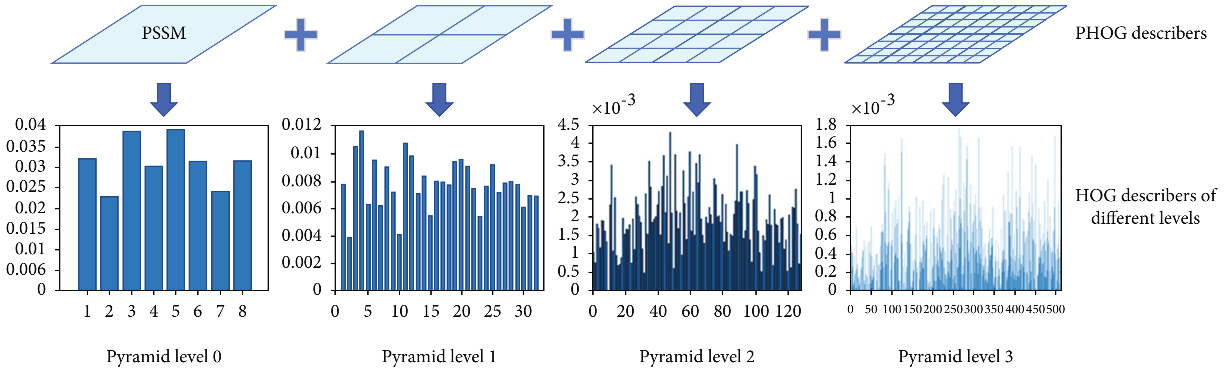


Figure 2: The example of merging HOG describers into PHOG describers.

different pyramid levels [30]. Meanwhile, this method has strong antinoise performance and antirotation ability [31]. Firstly, the given original image F is segmented into $i \times i$ spatial grids in the i_{th} pyramid level. Then, the histogram of oriented gradient (HOG) vectors of each grid should be calculated. Herein, we adopted Sobel operators to detect the edges and reduce the noise of the image. The Sobel operators can be defined as follows:

$$\mathrm{Sobel}_x = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \times [1\,2\,1] = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}, \tag{2}$$

$$\mathrm{Sobel}_y = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} \times [\,1 \quad 0 \quad -1\,] = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix}, \tag{3}$$

where Sobel_x and Sobel_y represent the horizontal operator and the vertical operator individually [32]. Then, the first-order differential Sobel operator is utilized to convolute the given image as follows:

$$G_x = F * \mathrm{Sobel}_x, \tag{4}$$

$$G_y = F * \mathrm{Sobel}_y, \tag{5}$$

where G_x denotes the convolution of picture F in the x-axis direction, where G_y denotes the convolution of image F in the y-axis direction. After convolution, the image F is converted into I which can be obtained as follow:

$$I = \sqrt{G_x^2 + G_y^2}. \tag{6}$$

The gradient magnitude g and direction θ of pixels in grids can be obtained by the following formulas:

$$g(\varphi, \omega) = \sqrt{\mathrm{g}_x(\varphi, \omega)^2 + g_y(\varphi, \omega)^2}, \tag{7}$$

$$\theta(\varphi, \omega) = \arctan \frac{g_y(\varphi, \omega)}{g_x(\varphi, \omega)}, \tag{8}$$

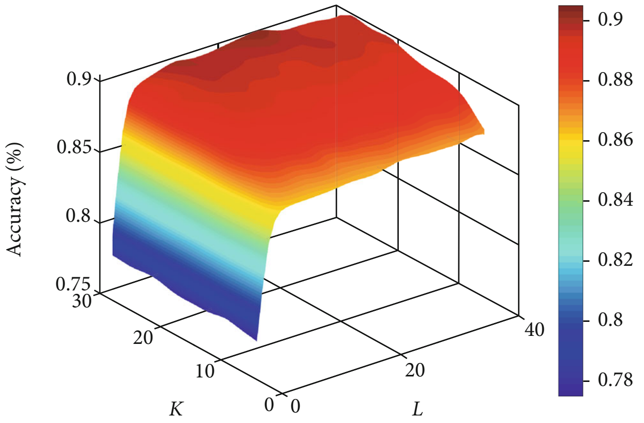


Figure 3: Accuracy surface of RF classifier influenced by parameters K and L.

where g_x and g_y can be computed as follows:

$$g_x(\varphi, \omega) = I(\varphi + 1, \omega) - I(\varphi - 1, \omega), \tag{9}$$

$$g_y(\varphi, \omega) = I(\varphi, \omega + 1) - I(\varphi, \omega - 1), \tag{10}$$

where φ and ω represent the coordinate position of a pixel in the picture. The [0-360] orientation is divided into m regions, and the pixels are divided into m regions to count HOG by gradient direction. Then, the HOG eigenvectors which contain m values have to be normalized by the following formula:

$$V = \frac{V}{\sqrt{\|V\|_2^2 + \varepsilon^2}}, \tag{11}$$

where V represents the HOG feature vector and ε is a small constant. Finally, HOG features of each spatial grid from all pyramid levels are concatenated to be PHOG feature descriptors. In this experiment, we set parameters $L = 3$ and $m = 8$. The number of grids in four levels is 85 $(1 + 2 \times 2 + 4 \times 4 + 8 \times 8)$, and converted the PSSM into a 680 (85×8) dimensional vector. Figure 2 gives an example of merging HOG describers into PHOG describers.

Table 2: 5-fold CV performance of our approach on the *enzyme* data set.

Test set	Acc. (%)	Pre. (%)	Sen. (%)	Spec. (%)	MCC (%)
1	88.89	88.31	88.93	88.85	77.77
2	90.34	91.85	89.27	91.49	80.71
3	89.83	91.61	88.07	91.65	79.73
4	87.69	88.19	86.99	88.40	75.39
5	88.03	88.83	86.34	89.65	76.07
Average	88.96 ± 1.13	89.76 ± 1.82	87.92 ± 1.25	90.01 ± 1.50	77.93 ± 2.29

Table 3: 5-fold CV performance of our approach on the *ion channel* data set.

Test set	Acc. (%)	Pre. (%)	Sen. (%)	Spec. (%)	MCC (%)
1	86.44	86.13	87.83	84.97	72.85
2	84.74	84.67	85.23	84.25	69.49
3	88.64	87.95	89.46	87.84	77.30
4	84.24	83.88	82.37	85.90	68.35
5	87.80	88.55	87.38	88.24	75.60
Average	86.37 ± 1.90	86.24 ± 2.02	86.45 ± 2.74	86.24 ± 1.75	72.72 ± 3.84

Table 4: 5-fold CV performance of our approach on the *GPCR* data set.

Test set	Acc. (%)	Pre. (%)	Sen. (%)	Spec. (%)	MCC (%)
1	85.04	82.96	88.19	81.89	70.22
2	81.89	84.85	81.16	82.76	63.73
3	81.50	77.50	82.30	80.85	62.86
4	83.79	84.35	80.83	86.47	67.49
5	82.21	85.83	80.15	84.62	64.58
Average	82.88 ± 1.49	83.10 ± 3.30	82.53 ± 3.26	83.32 ± 2.24	65.78 ± 3.03

Table 5: 5-fold CV performance of our approach on the *nuclear receptor* data set.

Test set	Acc. (%)	Pre. (%)	Sen. (%)	Spec. (%)	MCC (%)
1	75.00	75.12	70.59	78.95	49.77
2	77.78	77.78	77.89	77.65	55.56
3	86.11	85.71	90.00	81.25	71.81
4	77.14	72.73	88.89	64.71	55.44
5	68.57	60.90	87.50	52.63	42.12
Average	76.92 ± 6.30	74.45 ± 9.01	82.97 ± 8.43	71.04 ± 12.14	54.94 ± 10.91

2.5. Rotation Forest (RF). Rodriguez et al. developed the early integrated forest into the rotation forest (RF) [33]. Rotation forest works well on difference promotions and classifications of small sample data sets [34]. In particular, the RF classifier has outstanding performance on balancing the diversity and accuracy of the base classifier by rotating the subsets. Meanwhile, the model also preserves the efficiency, interpretability, and simplicity of the decision tree. In this paper, we employ RF to predict DTIs. The detailed process is shown as follows.

In practical terms, the data is randomly separated into K subsets containing disjoint features. Afterward, the bootstrap and Principal Component Analysis (PCA) method are applied in subsets to obtain rotation matrices with high diversity. Finally, these matrixes are fed into the corresponding base classifier, and the scores of each decision tree are counted. The matrix X of $n \times m$ is treated as a training feature set which contains m features of n samples, and $T = (t_1, t_2, \cdots, t_n)^T$ stores the corresponding labels of n samples. RF has L base classifiers D_i. The detailed training process of the base classifier is as follows.

(I) After optimizing the model, the data set M is separated into K disjoint subsets at random, and each subset has $C = m/k$ features

(II) Let $M_{i,j}$ represent the j_{th} subset of M, and $X_{i,j}$ is the corresponding feature set of $M_{i,j}$. Then, calculate

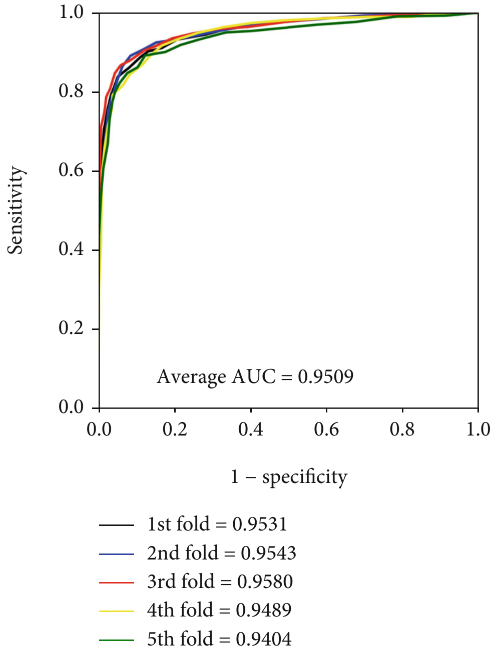


Figure 4: ROC curves performed by our approach on the *enzyme* data set.

the new training feature set $X'_{i,j}$ by bootstrap sampling on 75% of $X_{i,j}$

(III) Perform PCA on $X'_{i,j}$ to get the principal component coefficients which can be represented as $a_{i,j}^{(1)}, a_{i,j}^{(2)}, \cdots a_{i,j}^{(C_j)}$

(IV) These coefficients construct the sparse rotation matrix Z_i as follows:

$$Z_i = \begin{bmatrix} a_{i,1}^{(1)}, a_{i,1}^{(2)}, \cdots a_{i,1}^{(C_1)} & 0 & \cdots & 0 \\ 0 & a_{i,2}^{(1)}, a_{i,2}^{(2)}, \cdots a_{i,2}^{(C_2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{i,K}^{(1)}, a_{i,K}^{(2)}, \cdots a_{i,K}^{(C_k)} \end{bmatrix}. \tag{12}$$

In the process of classification, the possibility that sample x belongs to category y_i is $d_{i,j}(xZ_i^a)$ generated by base classifier D_i. Subsequently, count the confidence degrees that x belongs to each class by mean combination as follows:

$$\mu_j(x) = \frac{1}{L}\sum_{i=1}^{L} d_{i,j}(xZ_i^a). \tag{13}$$

The sample x will be distributed into the most possible class in accordance with the degree.

3. Results and Discussion

3.1. Evaluation Criteria. Throughout the experiments, accuracy (Acc.), sensitivity (Sen.), precision (Pre.), specificity (Spec.), and Matthews correlation coefficient (MCC) comprehensively appraise the prediction performance. These criteria can be defined as follows:

$$\text{Acc.} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}, \tag{14}$$

$$\text{Sen.} = \frac{\text{TP}}{\text{TP} + \text{TN}}, \tag{15}$$

$$\text{Pre.} = \frac{\text{TP}}{\text{FP} + \text{TP}}, \tag{16}$$

$$\text{Spec.} = \frac{\text{TN}}{\text{TP} + \text{FP}}, \tag{17}$$

$$\text{MCC} = \frac{\text{TN} \times \text{TP} - \text{FN} \times \text{FP}}{\sqrt{(\text{TN} + \text{FN}) \times (\text{TP} \times \text{FP}) \times (\text{TN} + \text{FP}) \times (\text{TP} \times \text{FN})}}, \tag{18}$$

where true positive (TP) represents the sum of interacting drug-target pairs with correct predictions, true negative (TN) reflects the aggregate of noninteracting drug-target pairs with correct predictions, false positive (FP) denotes the count of noninteracting drug-target pairs with incorrect classifications, and false negative (FN) represents the count of interacting drug-target pairs with incorrect classifications. Furthermore, receiver operating characteristic (ROC) curves are employed to depict results [35], and the area under the curve (AUC) is calculated to justify the prediction feasibility [36].

3.2. Parameter Discussion. In this experiment, parameters K and L are relevant to the results of the model. The K value and L value represent the numbers of the feature subsets and decision trees of RF, respectively. We applied the grid search algorithm to get the optimum parameters [37]. The method indicates that the accuracy ascends with the growth of the L value. When $K = 28$ and $L = 26$, the model has the best performance. Hence, we set the K value and the L value as 28 and 26, respectively. Figure 3 shows the accuracy surface of the RF classifier influenced by parameters K and L.

3.3. Fivefold CV Results on Four Data Sets. This section applied 5-fold CV on *enzyme*, *ion channel*, *GPCR*, and *nuclear receptor* data sets to obtain evaluation results for further verifying the reliability of our model. During the validation, the data set was broken into five subsets on average. Specifically, each subset took turns to be regarded as the testing part; the other four subsets merged into the training part in five repetitive experiments. Tables 2–5 list the results of validations on benchmark data sets.

It is obvious that the model worked well on four golden standard data sets from Tables 2–5. In terms of the results yielded by the *enzyme* data set, the average accuracy, precision, sensitivity, specificity, and MCC are 88.96%, 89.76%, 87.92%, 90.01%, and 77.93% with standard deviations of 1.13%, 1.82%, 1.25%, 1.50%, and 2.29%, respectively. As for

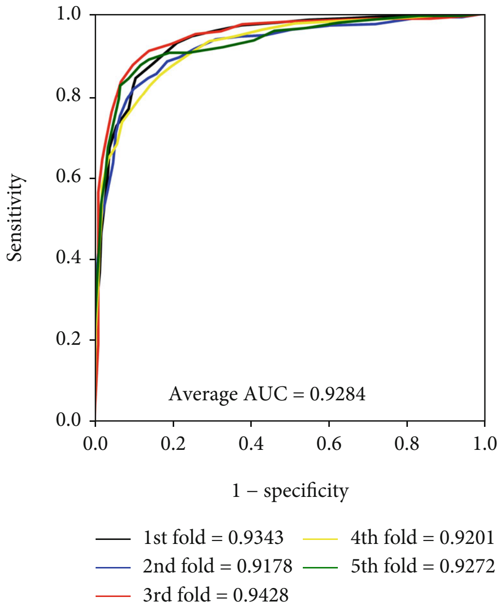


Figure 5: ROC curves performed by our approach on the *ion channel* data set.

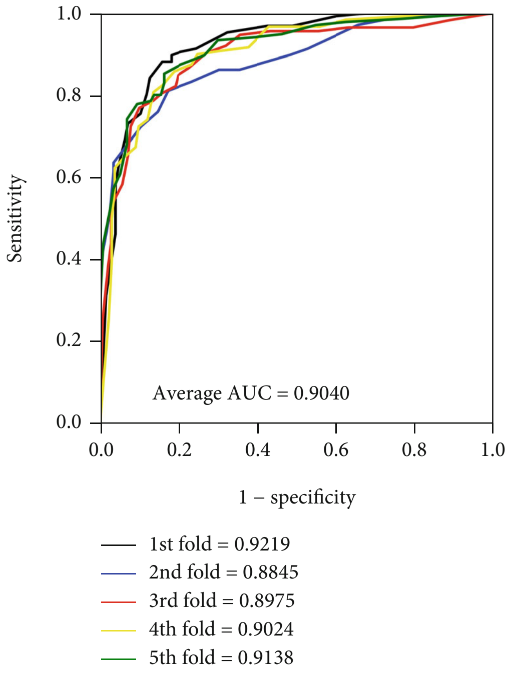


Figure 6: ROC curves performed by our approach on the *GPCRs* data set.

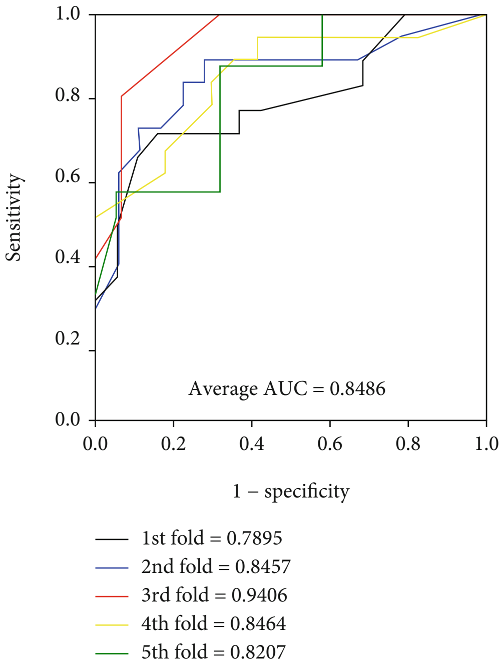


Figure 7: ROC curves performed by our approach on the *nuclear receptor* data set.

the results yielded on the *ion channel* data set, the accuracy, precision, sensitivity, specificity, and MCC come to be 86.37%, 86.24%, 86.45%, 86.24%, and 72.72% with standard deviations of 1.90%, 2.02%, 2.74%, 1.75%, and 3.84%, respectively. When performing the model on the *GPCR* data set, we obtained the average accuracy, precision, sensitivity, specificity, and MCC of 82.88%, 83.10%, 82.53%, 83.32%, and 65.78% with standard deviations of 1.49%, 3.30%, 3.26%, 2.24%, and 3.03%, respectively. When verifying the proposed model on the *nuclear receptor* data set, the model generates average accuracy, precision, sensitivity, specificity, and MCC of 76.92%, 74.45%, 82.97%, 71.04%, and 54.94% with standard deviations of 6.30%, 9.01%, 8.43%, 12.14%, and 10.91%, respectively. The difference between the sample quantities caused the gap of the evaluating criteria and standard deviations between four benchmark data sets. The average AUC of the proposed model were 0.9509, 0.9284, 0.9040, and 0.8486, respectively. Figures 4–7 give the ROC curves for the four benchmark data sets.

3.4. Comparison between the Models with PHOG Descriptor and LPQ Descriptor. For fairly evaluating the performance of the PHOG descriptor, we also conducted the experiments with local phase quantization (LPQ) which has a wide application prospect in spatial fuzzy image texture description processing for blurred-invariant property [38, 39]. Table 6 displays the comparison between PHOG and LPQ with rotation forest. The summarized table clearly indicates that the model with PHOG descriptors has a performance promotion than the LPQ descriptors on four golden standard data sets. In particular, the precision, sensitivity, specificity, and MCC all improved in the *ion channel* and *GPCR* data sets. Figure 8 plots the ROC curves of the PHOG and LPQ models on four data sets with mean AUC values. As can be noted, the AUC values of the PHOG model are higher than the LPQ model. Especially in the *GPCR* and *nuclear receptor* data sets, AUC gaps attend to 1.81% and 2.75%, respectively. Hence, our model can effectively describe PSSM to identify potential interacting drug-target pairs.

3.5. Comparison with Other Classifiers. At present, the classifiers which were used in predicting DTIs are mainly based on

Table 6: The comparison between LPQ and PHOG with rotation forest in terms of accuracy (Acc.), precision (Pre.), sensitivity (Sen.), specificity (Spec.), and Matthews correlation coefficient (MCC) on four types of benchmark data sets.

Data set	Method	Acc. (%)	Prec. (%)	Sen. (%)	Spec. (%)	MCC (%)
Enzyme	LPQ+RF	88.48 ± 0.87	90.30 ± 1.92	87.10 ± 1.52	90.65 ± 1.87	77.79 ± 1.76
	PHOG+RF	88.96 ± 1.13	89.76 ± 1.82	87.92 ± 1.25	90.01 ± 1.50	77.93 ± 2.29
Ion Channel	LPQ+RF	85.36 ± 1.00	85.22 ± 1.84	85.53 ± 1.37	85.13 ± 2.15	70.70 ± 2.00
	PHOG+RF	86.37 ± 1.90	86.24 ± 2.02	86.45 ± 2.74	86.24 ± 1.75	72.72 ± 3.84
GPCRs	LPQ+RF	82.02 ± 2.82	81.59 ± 5.65	82.34 ± 3.43	81.62 ± 4.39	63.92 ± 5.56
	PHOG+RF	82.88 ± 1.49	83.10 ± 3.30	82.53 ± 3.26	83.32 ± 2.24	65.78 ± 3.03
Nuclear receptor	LPQ+RF	75.78 ± 6.78	75.66 ± 6.80	77.62 ± 6.17	72.93 ± 14.28	50.89 ± 15.46
	PHOG+RF	76.92 ± 6.30	74.45 ± 9.01	82.97 ± 8.43	71.04 ± 12.14	54.94 ± 10.91

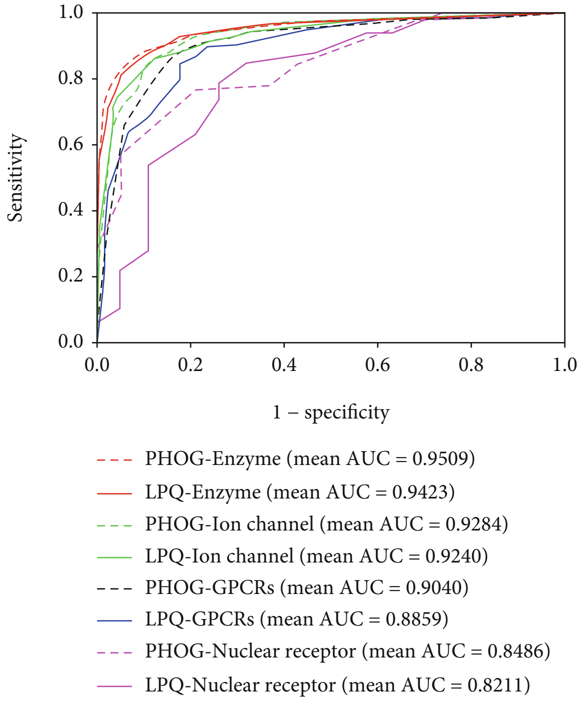


Figure 8: Performance comparison between LPQ and PHOG on four golden standard data sets.

traditional machine learning methods. In this section, we adopted advanced SVM and LGBM to combine PHOG descriptors. In the rotation forest, we set parameters $K = 28$ and $L = 26$ by utilizing the grid search method. After parameter optimization, SVM embedded Gaussian kernel function with parameters $C = 0.2$ and Gamma = 40. Parameter C prevents SVM from over fitting, and Gamma determines the number of support vectors. LGBM is based on a gradient boosting framework, and it is widely used in classification in industrial practice for it is time-saving and memory-conserving. After conducting grid method searching, the best results can be obtained by setting the number of leaves to 60, the learning rate to 0.05, and the number of training rounds to 40.

Figure 9 gives the results of RF, SVM, and LGBM on the *enzyme*, *ion channel*, *GPCR*, and *nuclear receptor* data sets, and it clearly reports that RF has a better performance with the PHOG descriptor than the other classifiers on verifying interacting drug-target pairs. The mean accuracy of RF is 8.30%, 8.07%, 11.43%, and 9.56% higher than SVM on the four golden standard data sets. Compared with the LGBM algorithm, the accuracy of RF improved 3.90%, 4.34%, 1.65%, and 8.33%, respectively. Figures 10 and 11 depict the ROC curves of the benchmark data sets generated by the rates of true positive (TP) and false positive (FP). In addition, mean AUC values are also attached to each graph for more intuitively describing the effect of different classifiers. The reliability of predicting DTIs of the model is proportional to the value of AUC. It can be observed that RF has performance promotions of 9.27%, 9.63%, 12.52%, and 11.49% against SVM on the four benchmark data sets. The value gaps of AUC between RF and LGBM are 8.97%, 7.59%, 9.38%, and 11.48%, respectively, on the four data sets. Accordingly, RF is more competitive than the other models in predicting DTIs.

3.6. Comparison with Other Methods. To date, many researchers have innovatively provided effective solutions for the prediction of DTIs. In order to further validate the efficiency of our model, we selected such previous models as MLCLE [40], NetCBP [41], SIMCOMP [42], WNN-GIP [43], AM-PSSM [44], NetLapRLS [45], MSCMF [20], and Bigram-PSSM [46] to analyze the performance of the proposed model. Meanwhile, all of these models are under the 5-fold CV framework on benchmark data sets. The average AUC values of these methods obviously indicate that the effect of our model has a significant enhancement in prediction in Table 7. In terms of the *enzyme*, *ion channel*, and *GPCR* data sets, the growths of the AUC reached 0.0029, 0.0119, and 0.0320, respectively. With regard to the *nuclear receptor* data set, Bigram-PSSM has the best performance with an AUC improvement of 0.0204 than our model. The results illustrate that the model which embeds PHOG descriptors and rotation forest is competent to effectively identify DTIs.

4. Conclusion

In this paper, we fused the pyramid histogram of oriented gradients (PHOG), Position-Specific Scoring Matrix (PSSM),

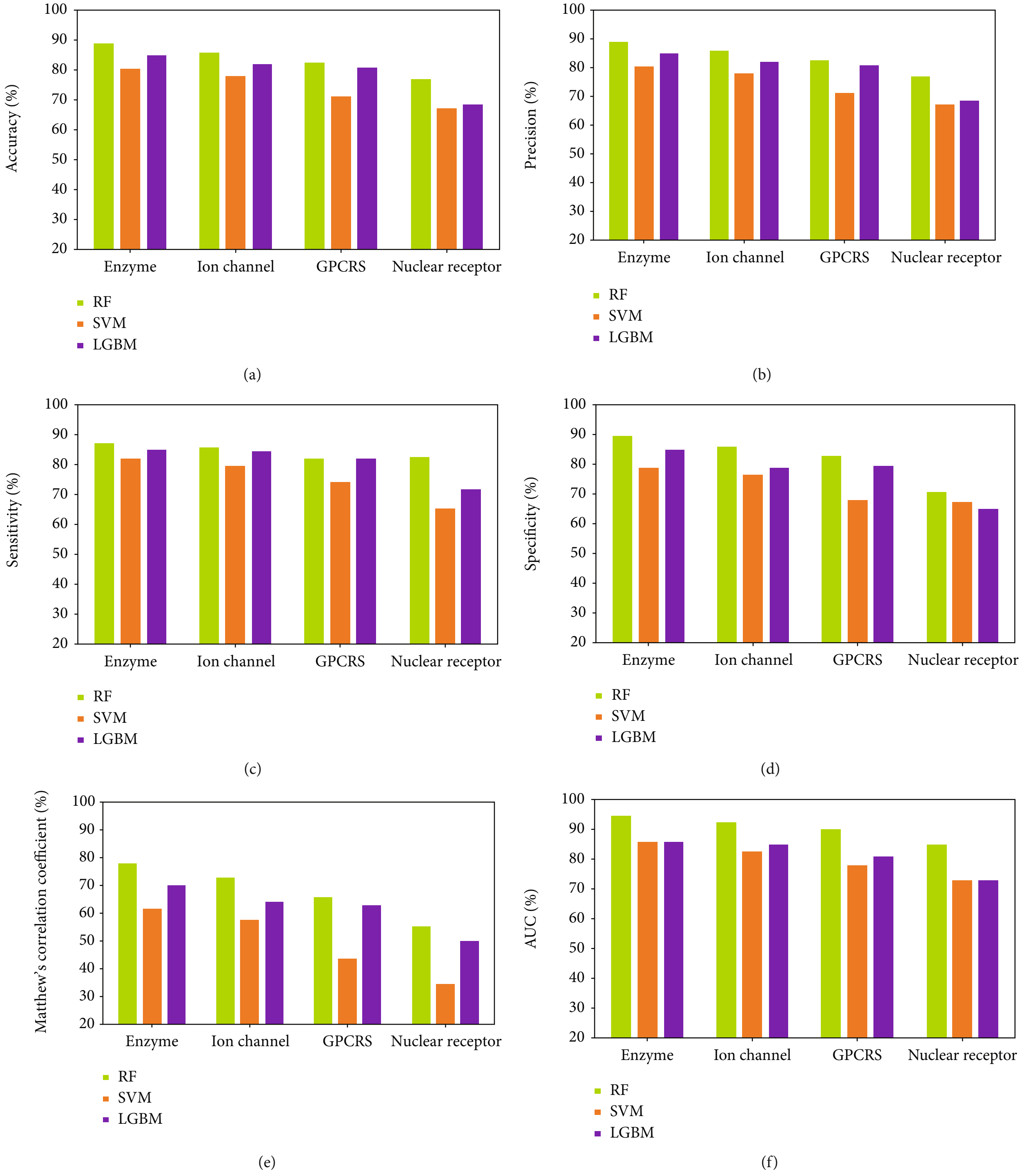


Figure 9: Comparison of experimental outcomes of RF, SVM, and LGBM on four benchmark data sets with six evaluating indicators: (a) accuracy; (b) precision; (c) sensitivity; (d) specificity; (e) MCC; (f) AUC.

and rotation forest (RF) into a novel computational model to predict the interactions between drugs and targets. To prove the reliability of the proposed model, a series of experiments have been conducted. Specifically, we first altered the feature extraction method of the proposed model with LPQ to assess the feature description ability of PHOG. This paper also experimented on state-of-the-art LGBM and SVM with the same features to validate the performance of RF. Among them, the proposed model achieves mean accuracies of 88.96%, 86.37%, 82.88%, and 76.92% on the *enzyme*, *ion*

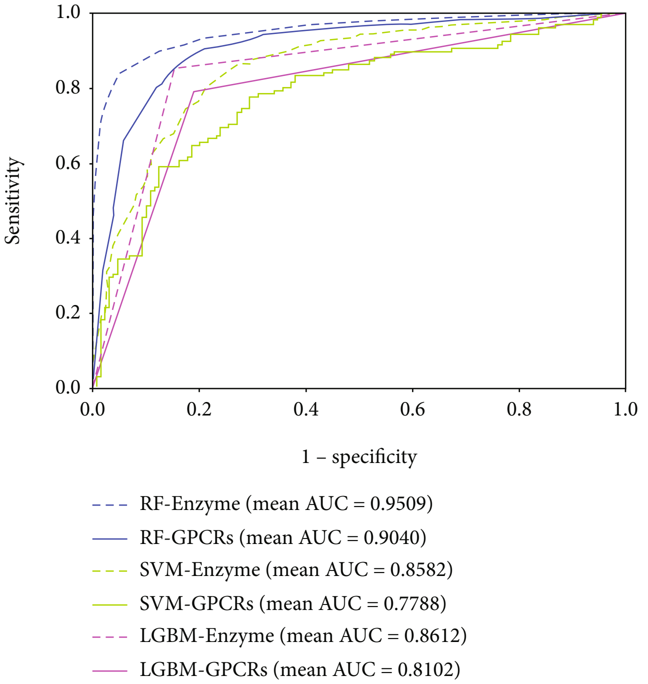


Figure 10: ROC curves performed by RF, SVM, and LGBM on the *enzyme* and *GPCR* data sets.

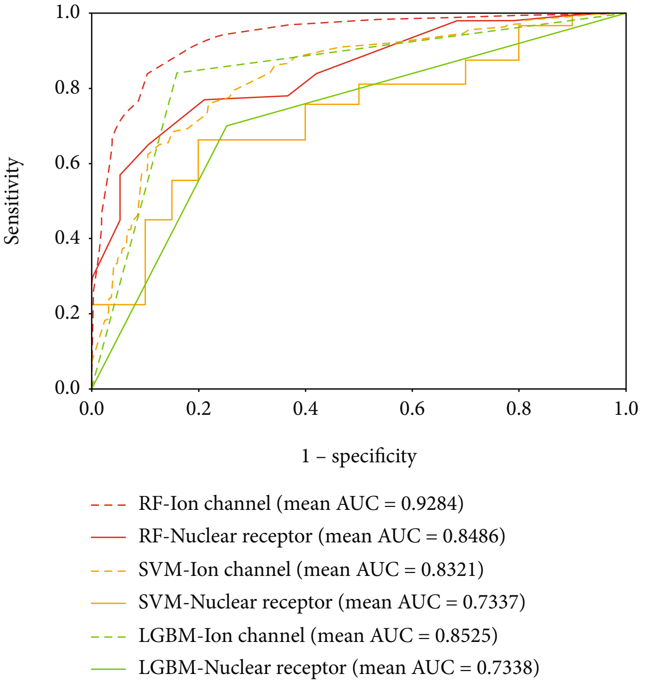


Figure 11: ROC curves performed by RF, SVM, and LGBM on the *ion channel* and *nuclear receptor* data sets.

Table 7: The comparison of AUC values obtained by the proposed model and other advanced model on four benchmark data sets.

Model	Enzyme	Ion channel	GPCRs	Nuclear receptor
MLCLE	0.842	0.795	0.850	0.790
NetCBP	0.8251	0.8034	0.8235	0.8394
SIMCOMP	0.863	0.776	0.867	0.856
WNN-GIP	0.861	0.775	0.872	0.839
AM-PSSM	0.843	0.722	0.839	0.767
NetLapRLS	0.9013	0.9165	0.7711	0.6772
MSCMF	0.9142	0.776	0.867	0.856
Bigram-PSSM	0.948	0.889	0.872	0.869
Our method	0.9509	0.9284	0.9040	0.8486

channel, *G protein-coupled receptor* (*GPCR*), and *nuclear receptor* data sets. The results obviously illustrate that the PHOG features can trace the local characteristics and assist the model to improve the accuracy even compared with LPQ. Meanwhile, the model is considered to be an extraordinarily suitable tool for providing candidates of drug discovery. In the subsequent work, we will experiment with more methods to further raise the feasibility of the prediction model.

5. Limitation and Future Work

Although the model shows an improved prediction ability than other models, we still noticed the singleness of the local feature, and the noise existing in features also has an adverse effect on forming describers. The main limitations of the model can be explained from two aspects. On the one side, the utilized feature extraction method is sensitive to local feature information. However, it is hard to excavate the global feature information of samples. On another side, the same number of unlabelled samples is randomly selected to be negative samples as the known interacting drug-target pairs; hence, the model wastes a large number of unselected negative samples. The feature studies will mainly focus on the processes of feature extraction and classification. The external edge features which have an excellent application prospect in the field of image tamper prevention will integrate the internal features to comprehensively describe bioinformation with less noise. Meanwhile, unsupervised learning models will be adopted to confront the waste of data sets, and it will make full use of high-throughput unbalanced data. These improvements will bring new challenges and opportunities to develop robust prediction tools for enhancing the model prediction accuracy.

Data Availability

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

Conflicts of Interest

The author declares that there is no conflict of interest.

Authors’ Contributions

Conceptualization was handled by Z.-Y.Z.; methodology, software, and validation was handled by Z.-YZ.; formal analysis was handled by J.P. and Y.-A.H.; data curation was handled by Y.-A.H. and X.-K.Z.; project administration was handled by W.-Z.H., S.-W.Z., and C.-Q.Y.; and funding acquisition was handled by W.-Z.H.

Acknowledgments

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

References

[1] X. Y. Yan and S. W. Zhang, "Identifying drug-target interactions with decision templates," *Current Protein & Peptide Science*, vol. 19, no. 5, pp. 498–506, 2018.

[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, 2016.

[3] F. Zhu, B. C. Han, P. Kumar et al., "Update of TTD: therapeutic target database," *Nucleic Acids Research*, vol. 38, suppl_1, pp. D787–D791, 2010.

[4] D. S. Wishart, C. Knox, A. C. Guo et al., "DrugBank: a knowledgebase for drugs, drug actions and drug targets," *Nucleic Acids Research*, vol. 36, suppl_1, pp. D901–D906, 2008.

[5] E. L. Willighagen, A. Waagmeester, O. Spjuth et al., "The ChEMBL database as linked open data," *Journal of Cheminformatics*, vol. 5, no. 1, p. 23, 2013.

[6] M. Kanehisa and S. Goto, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, 2000.

[7] S. D'Souza, K. V. Prema, and B. Seetharaman, "Machine learning models for drug–target interactions: current knowledge and future directions," *Drug Discovery Today*, vol. 25, no. 4, pp. 748–756, 2020.

[8] C. Vénien-Bryan, Z. Li, L. Vuillard, and J. A. Boutin, "Cryo-electron microscopy and X-ray crystallography: complementary approaches to structural biology and drug discovery," *Acta Crystallogr F Struct Biol Commun*, vol. 73, no. 4, pp. 174–183, 2017.

[9] Y. Kurogi and O. F. Güner, "Pharmacophore modeling and three-dimensional database searching for drug design using catalyst," *Current Medicinal Chemistry*, vol. 8, no. 9, pp. 1035–1055, 2001.

[10] O. F. Güner, "History and evolution of the pharmacophore concept in computer-aided drug design," *Current Topics in Medicinal Chemistry*, vol. 2, no. 12, pp. 1321–1332, 2002.

[11] Q. Zou, J. Zeng, L. Cao, and R. Ji, "A novel features ranking metric with application to scalable visual and bioinformatics data classification," *Neurocomputing*, vol. 173, pp. 346–354, 2016.

[12] Y. Ding, J. Tang, and F. Guo, "Identification of drug–target interactions via fuzzy bipartite local model," *Neural Computing and Applications*, vol. 32, no. 14, pp. 10303–10319, 2020.

[13] B. Krisztian, L. Peška, and J. Koller, "Modified linear regression predicts drug-target interactions accurately," *PLoS One*, vol. 15, no. 4, article e0230726, 2020.

[14] Y. Ye, Y. Wen, Z. Zhang, S. He, and X. Bo, "Drug-target interaction prediction based on adversarial Bayesian personalized ranking," *BioMed Research International*, vol. 2021, 16 pages, 2021.

[15] M. Bagherian, R. B. Kim, C. Jiang, M. A. Sartor, H. Derksen, and K. Najarian, "Coupled matrix–matrix and coupled tensor–matrix completion methods for predicting drug–target interactions," *Briefings in Bioinformatics*, vol. 22, no. 2, pp. 2161–2171, 2021.

[16] K. Pliakos and C. Vens, "Drug-target interaction prediction with tree-ensemble learning and output space reconstruction," *BMC Bioinformatics*, vol. 21, no. 1, p. 49, 2020.

[17] J. Y. An, F. R. Meng, and Z. J. Yan, "An efficient computational method for predicting drug-target interactions using weighted extreme learning machine and speed up robot features," *Bio Data Mining*, vol. 14, no. 1, p. 3, 2021.

[18] T. V. Laarhoven, S. B. Nabuurs, and E. Marchiori, "Gaussian interaction profile kernels for predicting drug-target interaction," *Bioinformatics*, vol. 27, no. 21, pp. 3036–3043, 2011.

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

[20] X. Zheng, H. Ding, H. Mamitsuka, and S. Zhu, "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, Illinois, USA, 2013.

[21] N. Hecker, J. Ahmed, J. von Eichborn et al., "SuperTarget goes quantitative: update on drug-target interactions," *Nucleic Acids Research*, vol. 40, pp. D1113–D1117, 2012.

[22] I. Schomburg, A. Chang, C. Ebeling et al., "BRENDA, the enzyme database: updates and major new developments," *Nucleic Acids Research*, vol. 32, no. 90001, pp. 431D–4433, 2004.

[23] 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.

[24] Z. Wu, F. Cheng, L. Li, W. Li, G. Liu, and Y. Tang, "SDTNBI: an integrated network and chemoinformatics tool for systematic prediction of drug–target interactions and drug repositioning," *Briefings in Bioinformatics*, vol. 18, no. 2, pp. bbw012–bbw347, 2016.

[25] X. Liang, W. Zhu, Z. Lv, and Q. Zou, "Molecular computing and bioinformatics," *Multidisciplinary Digital Publishing Institute*, vol. 24, no. 13, article 2358, 2019.

[26] W. Zhang, Y. Chen, and D. Li, "Drug-Target interaction prediction through label propagation with linear neighborhood information," *Molecules*, vol. 22, no. 12, p. 2056, 2017.

[27] Y. Wang, J. Xiao, T. O. Suzek, J. Zhang, J. Wang, and S. H. Bryant, "PubChem: a public information system for analyzing bioactivities of small molecules," *Nucleic Acids Research*, vol. 37, no. Web Server, pp. W623–W633, 2009.

[28] Q. Zou, G. Lin, X. Jiang, X. Liu, and X. Zeng, "Sequence clustering in bioinformatics: an empirical study," *Briefings in Bioinformatics*, vol. 21, no. 1, pp. 1–10, 2018.

[29] L. Nanni and S. Brahnam, "Set of approaches based on position specific scoring matrix and amino acid sequence for primary category enzyme classification," *Journal of Artificial Intelligence and Systems*, vol. 2, no. 1, pp. 38–52, 2020.

[30] J. Liang, Y. Han, and Q. Hu, "Semi-supervised image clustering with multi-modal information," *Multimedia Systems*, vol. 22, no. 2, pp. 149–160, 2016.

[31] N. Gour and P. Khanna, "Automated glaucoma detection using GIST and pyramid histogram of oriented gradients (PHOG) descriptors," *Pattern Recognition Letters*, vol. 137, pp. 3–11, 2020.

[32] D. A. Abdullah, M. H. Akpınar, and A. Şengür, "Local feature descriptors-based ECG beat classification," *Health Information Science Systems*, vol. 8, no. 1, p. 20, 2020.

[33] 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.

[34] Z. H. Guo, Z. H. You, Y. B. Wang, H. C. Yi, and Z. H. Chen, "A learning-based method for LncRNA-disease association identification combing similarity information and rotation forest," *iScience*, vol. 19, pp. 786–795, 2019.

[35] F. Yuan, L. Lu, and Q. Zou, "Analysis of gene expression profiles of lung cancer subtypes with machine learning algorithms," *Biochimica et Biophysica Acta-Molecular Basis of Disease*, vol. 1866, no. 8, article 165822, 2020.

[36] K. Qu, F. Guo, X. Liu, Y. Lin, and Q. Zou, "Application of machine learning in microbiology," *Frontiers in Microbiology*, vol. 10, p. 827, 2019.

[37] Q. Zou, W. Chen, Y. Huang, X. Liu, and Y. Jiang, "Identifying multi-functional enzyme by hierarchical multi-label classifier," *Journal of Computational and Theoretical Nanoscience*, vol. 10, no. 4, pp. 1038–1043, 2013.

[38] N. Loris, B. Sheryl, and L. Alessandra, "Local phase quantization descriptor for improving shape retrieval/classification," *Pattern Recognition Letters*, vol. 33, no. 16, pp. 2254–2260, 2012.

[39] P. Michelangelo, N. Loris, L. Anna, A. S. Katriina, H. Jari, and S. Stefano, "Non-binary coding for texture descriptors in subcellular and stem cell image classification," *Current Bioinformatics*, vol. 8, no. 2, pp. 208–219, 2013.

[40] K. Pliakos, C. Vens, and G. Tsoumakas, "Predicting drug-target interactions with multi-label classification and label partitioning," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pp. 1–1, 2019.

[41] H. Chen and Z. Zhang, "A semi-supervised method for drug-target interaction prediction with consistency in networks," *PLoS One*, vol. 8, no. 5, article e62975, 2013.

[42] H. Öztürk, E. Ozkirimli, and A. Özgür, "A comparative study of SMILES-based compound similarity functions for drug-target interaction prediction," *BMC Bioinformatics*, vol. 17, no. 1, p. 128, 2016.

[43] T. V. Laarhoven and E. Marchiori, "Predicting drug-target interactions for new drug compounds using a weighted nearest neighbor profile," *PLoS One*, vol. 8, no. 6, article e66952, 2013.

[44] L. Nanni, A. Lumini, and S. Brahnam, "A set of descriptors for identifying the protein–drug interaction in cellular networking," *Journal of Theoretical Biology*, vol. 359, pp. 120–128, 2016.

[45] 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.

[46] Z. Mousavian, S. Khakabimamaghani, K. Kavousi, and A. Masoudi-Nejad, "Drug–target interaction prediction from PSSM based evolutionary information," *Journal of Pharmacological and Toxicological Methods*, vol. 78, pp. 42–51, 2016.

Hindawi
BioMed Research International
Volume 2021, Article ID 5556784, 10 pages
https://doi.org/10.1155/2021/5556784

Research Article

Identifying Infliximab- (IFX-) Responsive Blood Signatures for the Treatment of Rheumatoid Arthritis

ShiJian Ding,[1] **ZhanDong Li**,[2] **Tao Zeng**,[3] **Yu-Hang Zhang**,[4] **Tao Huang**,[3,5] **and Yu-Dong Cai**[1]

[1]*School of Life Sciences, Shanghai University, Shanghai, China*
[2]*College of Food Engineering, Jilin Engineering Normal University, Changchun, China*
[3]*Bio-Med Big Data Center, CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China*
[4]*Channing Division of Network Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA*
[5]*CAS Key Laboratory of Tissue Microenvironment and Tumor, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China*

Correspondence should be addressed to Tao Huang; tohuangtao@126.com and Yu-Dong Cai; cai_yud@126.com

Received 26 February 2021; Revised 18 March 2021; Accepted 24 March 2021; Published 7 April 2021

Academic Editor: Quan Zou

Rheumatoid arthritis (RA) is a severe chronic pathogenic inflammatory abnormality that damages small joints. Comprehensive diagnosis and treatment procedures for RA have been established because of its severe symptoms and relatively high morbidity. Medication and surgery are the two major therapeutic approaches. Infliximab (IFX) is a novel biological agent applied for the treatment of RA. IFX improves physical functions and benefits the achievement of clinical remission even under discontinuous medication. However, not all patients react to IFX, and distinguishing IFX-sensitive and IFX-resistant patients is quite difficult. Thus, how to predict the therapeutic effects of IFX on patients with RA is one of the urgent translational medicine problems in the clinical treatment of RA. In this study, we present a novel computational method for the identification of the applicable and substantial blood gene signatures of IFX sensitivity by liquid biopsy, which may assist in the establishment of a clinical drug sensitivity test standard for RA and contribute to the revelation of unique IFX-associated pharmacological mechanisms.

1. Introduction

Rheumatoid arthritis (RA) is a severe chronic pathogenic inflammatory abnormality that damages small joints [1, 2]. Some patients with severe and progressive RA may also suffer from pathogenic lesions in various body systems, including the skin, eyes, lungs, and blood vessels [1]. According to the statistics provided by the American College of Rheumatology in 2008, more than 1.3 million people in the USA suffer from pathogenic rheumatoid arthritis with typical symptoms [3, 4]; thus, RA is the leading cause of arthritis among multiple pathogeneses.

Comprehensive diagnosis and treatment procedures for RA have been established because of its severe symptoms and relatively high morbidity [1, 5–8]. The diagnosis of RA can be divided into two major procedures: symptom-dependent diagnosis [9] and clinical laboratory examinations [10]. A group of typical symptoms can be used to preliminary screen for RA. The typical symptoms of RA include the swelling of small joints (which may further extend to larger joints, such as the hips and shoulders), fatigue, fever, and anemia, which are quite easy to identify and recognize [11, 12]. However, the types and severity of these symptoms vary among patients. Thus, RA is difficult to diagnose based only on clinical symptoms. Apart from clinical symptoms, blood tests and imaging tests have also been applied for the accurate diagnosis of RA. Erythrocyte sedimentation rate [13] and C-reactive protein [13] are nonspecific biomarkers for the diagnosis and progression monitoring of RA that reflect the degree of inflammatory responses in the whole body. The gold standard for RA diagnosis is clinical liquid biopsy, and rheumatoid factor [14] and anticyclic citrullinated peptide [15] antibodies are the specific biomarkers for the blood test screening of RA. Pathogenic lesions in the joints can be regarded as a diagnosis measurement for RA and can be easily identified by X-ray [16], ultrasound [17], and magnetic resonance imaging [18].

Medication and surgery are the two major therapeutic approaches for RA [19, 20]. Different from surgery-based approaches, which focus on the direct relief of joint damage, most medications focus on the prevention of abnormal inflammatory immune responses and therefore indirectly relieve systemic symptoms. Four subgroups of medications are clinically applied, namely, nonsteroidal anti-inflammatory drugs, steroids, disease-modifying antirheumatic drugs, and biological agents [21]. Among these drugs, infliximab (IFX) is a novel biological agent applied for the treatment of RA. As a chimeric monoclonal antibody, IFX has been approved by the Food and Drug Administration (FDA) for the treatment of multiple immune-associated diseases, including RA, early in 2007 [22]. IFX targets one of the pathogenic proinflammatory cytokines, tumor necrosis factor-alpha (TNF-α), and thus has been applied in classical TNF antagonist therapy against multiple chronic autoimmune inflammatory diseases, including RA [23]. The clinical application of IFX has been studied for nearly 30 years; IFX greatly improves physical functions and is beneficial for the achievement of clinical remission even under discontinuous medication [24, 25]. However, not all patients react to IFX, and its anti-inflammatory effects vary among patients with RA. IFX-sensitive and IFX-resistant patients are difficult to distinguish based on traditional clinical examination; thus, how to predict the therapeutic effects of IFX on patients with RA is one of the urgent translational medicine problems in the clinical treatment of RA.

With the development of liquid biopsy and high-throughput sequencing technologies, a recent study [26] revealed that patients with different drug susceptibility against IFX have different pretherapeutic blood expression patterns; thus, liquid biopsy and the transcriptomic profiling of patients' blood may help predict the therapeutic effects of IFX in RA and therefore assist in clinical medication. However, the application of whole transcriptome sequencing on every patient with RA is not feasible; therefore, the identification of accurate and efficient transcriptomic biomarkers and their respective pathogenic expression pattern would be the priority. In this study, we presented a novel computational method to identify the applicable and substantial blood gene biomarkers of IFX sensitivity by liquid biopsy. This study may assist in the establishment of a test standard for clinical drug sensitivity for RA treatment and contribute to the revelation of unique IFX-associated pharmacological mechanisms.

2. Materials and Methods

2.1. Data. The blood gene expression profiles of 140 patients with RA before IFX treatment were downloaded from the Gene Expression Omnibus under accession number GSE78068 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE78068) [26]. Among the patients, 42 showed remission and 98 showed nonremission. 19595 genes were involved to constitute the blood gene expression profiles. Predicting the response of patients with RA to IFX before therapy using their blood will help in deciding previse treatments, which is the ultimate goal of precision medicine. We would like to build such IFX response prediction models for patients with RA based on their blood gene expression profiles.

2.2. Monte Carlo Feature Selection (MCFS). In this work, the MCFS method, which is a decision tree- (DT-) based feature selection method [27–29], was used to select important gene candidates. MCFS can randomly select a few feature subsets from the original features. Each feature subset consists of a small number of features from the original ones. One bootstrap sample dataset can be induced from this feature subset, and then, multiple DTs can be learned and tested on this bootstrap sample datasets. The process is repeated several times to produce many feature subsets. Each feature subset can help learn the same number of DTs.

The contribution of each feature in these DTs can be evaluated by the relative importance (RI) score, which is calculated as follows:

$$\mathrm{RI}_f = \sum_{\tau=1}^{pt} (\mathrm{wAcc})^u \sum_{n_f(\tau)} \mathrm{IG}\left(n_f(\tau)\right) \left(\frac{\text{no. in } n_f(\tau)}{\text{no. in } \tau}\right)^v, \quad (1)$$

where wAcc indicates the weighted accuracy and $n_f(\tau)$ represents a node of feature f in DT τ. The information gain of $n_f(\tau)$ in the DT is measured by $\mathrm{IG}[n_f(\tau)]$, and no.in $n_f(\tau)$ points the number of training samples in node $n_f(\tau)$. In addition, the weighting factors u and v have a value of 1 by default.

After each feature was assigned a RI value, we ranked all features in a list with the decreasing order of their RI values. In addition to the feature list, the MCFS method also outputs some most important features, called informative features, which are some top features in the list. These features are accessed by determining a threshold of RI values via a permutation test on class labels and one-sided Student's t-test.

This study used the MCFS program retrieved from http://www.ipipan.eu/staff/m.draminski/mcfs.html. Its default parameters were used.

2.3. Incremental Feature Selection (IFS). IFS is widely used to determine the optimal number of features for constructing a classification model with an integrated supervised classifier [30]. On the basis of a ranked feature list obtained from MCFS, a series of feature subsets are produced with a step interval of 10. For example, the first feature subset includes the top 10 features and the second feature subset includes the top 20 features. For each feature subset, a classifier (e.g., support vector machine (SVM)) is trained on a training data induced from this feature subset. An optimal feature subset is selected when it has the highest performance among the candidate feature subsets, where performance is evaluated by the Matthews correlation coefficients (MCC) [31] under 10-fold cross-validation [32]. The classifier with such optimal feature subset can be built and was called the optimum classifier in this study.

2.4. Support Vector Machine (SVM). SVM is a classification algorithm suitable for linear and nonlinear data [33–39]. For a dataset with two classes, SVM tries to find out an optimum hyperplane, which can divide samples in two classes with a maximum margin. However, in many cases, such hyperplane is not easy or impossible to be discovered. SVM employs a kernel trick to convert the original sample in a low-dimensional space to a new sample in some high-dimensional space, in which the optimum hyperplane can be easily constructed. For a new sample, it is also mapped into the high-dimensional space and its class is determined according to the side of the hyperplane it lies. In this study, we used the tool "SMO" in Weka [40, 41] to quickly implement SVM. Such SVM is optimized by the sequential minimum optimization algorithm [42]. For convenience, default parameters were adopted, where the kernel was a polynomial function.

2.5. Random Forest (RF). RF [43] is a metaclassifier that is widely applied in biological and biomedical researches [44–52]. RF includes many DTs as members. To construct a DT, a dataset, in which samples are randomly selected, with replacement, from the original dataset, is constructed. Such dataset has the same size of the original dataset. The DT is grown at each node by determining an optimal splitting way on some features, which were randomly selected from all features. Given a new sample, each DT gives its prediction. RF integrates these predictions with majority voting. Similar to SVM, we also employed a tool "RandomForest" in Weka [40, 41], which implements RF. Likewise, the default parameters were used, where the number of DTs was set to ten.

2.6. Rule Learning. In addition to "black-box" classification algorithms, the interpretable rules for a classification model can also be extracted to explain the feature differences between groups of patients with particular response to drug treatment. To accelerate this procedure, we directly picked up the informative features extracted by the MCFS method. These features were further filtered by the Johnson Reducer algorithm [53]. The remaining features were fed into the repeated incremental pruning to produce error reduction (RIPPER) algorithm [54] to extract classification rules. Obtained rules were represented by IF-ELSE rules. The above rule learning procedures were also integrated in the MCFS program, which was directly adopted in this study.

2.7. Measurement. The MCC [31] within 10-fold cross-validation was applied in this work to evaluate the classification performance of different classification models. The MCC for the binary problem is calculated as follows:

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TN} + \mathrm{FN}) \times (\mathrm{TN} \times \mathrm{FP}) \times (\mathrm{TP} + \mathrm{FN}) \times (\mathrm{TP} + \mathrm{FP})}}, \tag{2}$$

where TP, TN, FP, and FN represent the number of true-positive, true-negative, false-positive, and false-negative samples, respectively. As a measurement for the classification model, the MCC value ranges from −1 to +1, where a value of +1 indicates that the classification model has the best performance. To date, it has wide application in bioinformatics for evaluating the performance of different classification models [55–61].

Besides, we also employed other five measurements to give a full evaluation on different classification models. They were sensitivity (SN), specificity (SP), accuracy (ACC), precision, and $F1$-measure, which can be calculated by the following equations:

$$\begin{cases} \mathrm{SN} = \dfrac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \\ \mathrm{SP} = \dfrac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}, \\ \mathrm{ACC} = \dfrac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{TN} + \mathrm{FP}}, \\ \mathrm{Precision} = \dfrac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \\ F1\text{-measure} = \dfrac{2 \cdot \mathrm{Recall} \cdot \mathrm{Precision}}{\mathrm{Recall} + \mathrm{Precision}}. \end{cases} \tag{3}$$

3. Results

In this study, we gave a computational investigation on the blood gene expression profiles of patients with RA before IFX treatment. The entire procedures are illustrated in Figure 1. This section gave the results of all procedures.

3.1. Results of MCFS. The blood gene expression profiles were first analyzed by the MCFS method. As a result, each feature was assigned a RI value, which indicated its importance. The RI values of all features are listed in Table S1. Accordingly, all features were ranked in a list by the decreasing order of features' RI values. Such list is also provided in Table S1.

3.2. Results of IFS. Based on the feature list obtained in the above section, the IFS method is followed. It first constructed several feature subsets. Then, on each feature subset, a classifier was built using SVM or RF as the classification algorithm. Each classifier was evaluated by 10-fold cross-validation. Six measurements (see equations (2) and (3)) were obtained for each classifier, which are listed in Tables S2 and S3. For an easy observation, a curve was plotted for each classification algorithm, as shown in Figure 2, in which MCC was set as the y-axis and the number of features was set as the x-axis. It can be observed that for SVM, the highest MCC was 0.760 when top 1260 features were used. Thus, these 1260 features constituted the optimum feature subset of SVM and the SVM classifier with these features was the optimum SVM classifier. Other measurements of such classifier are listed in Table 1. As for RF, the highest MCC was 0.611 when only top ten features were adopted. An optimum RF classifier was built based on these top ten features. The MCC of the optimum RF classifier was much lower than that of the optimum SVM classifier. Other five measurements of such RF classifier are provided in Table 1. Evidently, each measurement was inferior to that of the optimum SVM

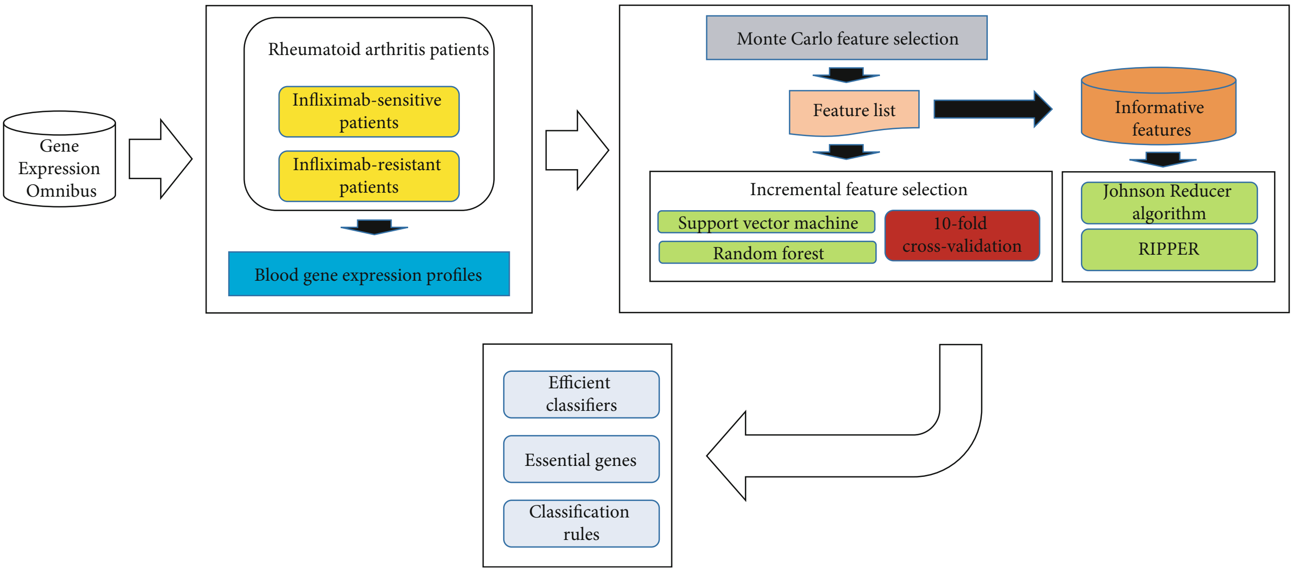

Gene Expression Omnibus
Rheumatoid arthritis patients
Infliximab-sensitive patients
Infliximab-resistant patients
Blood gene expression profiles
Monte Carlo feature selection
Feature list
Informative features
Incremental feature selection
Support vector machine
Random forest
10-fold cross-validation
Johnson Reducer algorithm
RIPPER
Efficient classifiers
Essential genes
Classification rules

Figure 1: Entire procedures to investigate the blood gene expression profiles of rheumatoid arthritis patients. Profiles are retrieved from Gene Expression Omnibus, which are analyzed by the Monte Carlo feature selection method. A feature list is obtained, which is fed into the incremental feature selection method to construct efficient classifiers and extract essential genes. On the other hand, informative features, which are some top features in the list, are used to construct classification rules via Johnson Reducer and RIPPER algorithms.

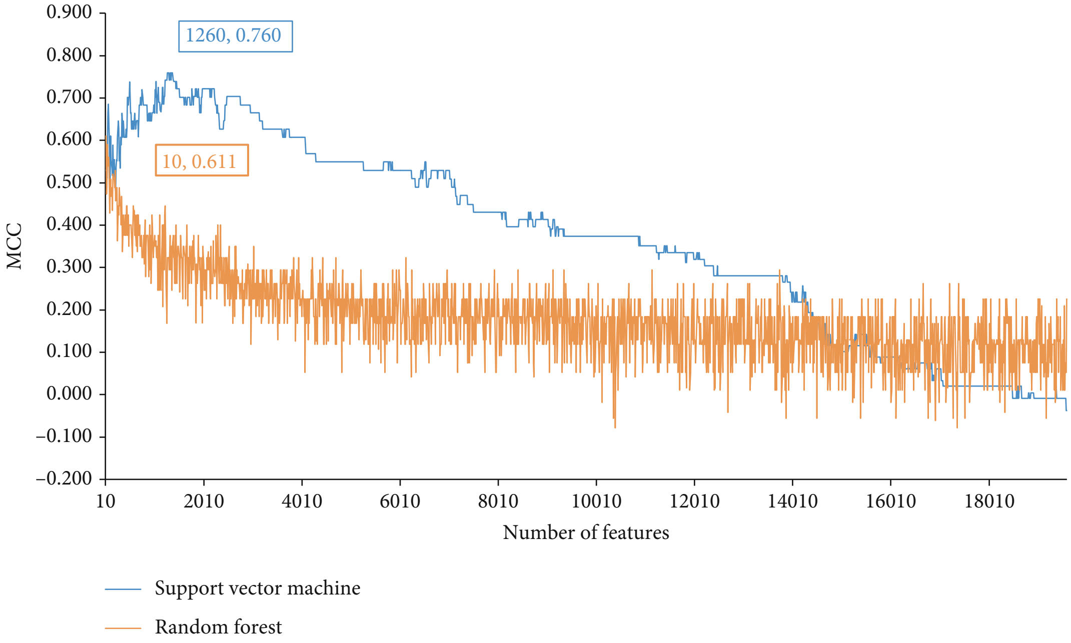


Figure 2: IFS curves with different classification algorithms on different numbers of features (genes). The support vector machine yields the highest MCC of 0.760 when top 1260 features are used, whereas the random forest generates the highest MCC of 0.611 when top 10 features are adopted.

Table 1: Performance of some key support vector machine (SVM) and random forest (RF) classifiers.

Classification algorithm	Number of features	Sensitivity	Specificity	Accuracy	Precision	$F1$-measure
SVM	1260	0.690	0.990	0.900	0.967	0.806
SVM	60	0.619	0.980	0.871	0.929	0.743
RF	10	0.643	0.929	0.843	0.794	0.711

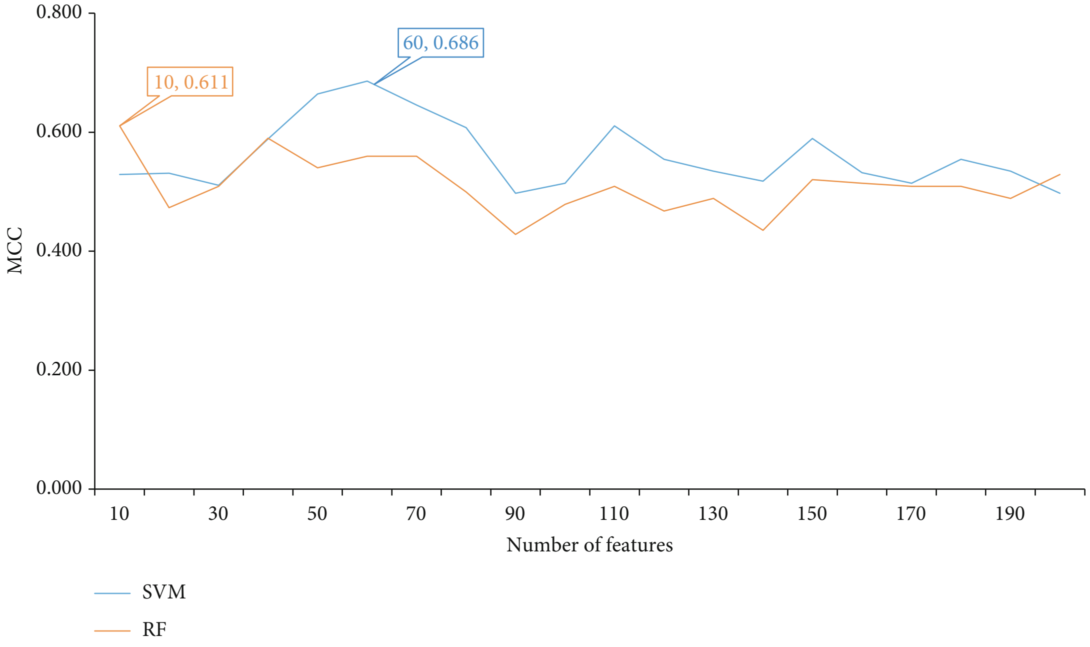


Figure 3: IFS curves with different classification algorithms on top 10-200 features. The support vector machine (SVM) yields the MCC of 0.686 when only top 60 features are used. It is still higher than the highest MCC yielded by the optimum random forest (RF) classifier.

classifier. Therefore, it can be concluded that the optimum SVM classifier is better than the optimum RF classifier.

As mentioned above, the optimum SVM classifier needed much more features than the optimum RF classifier. In fact, the SVM classifier can yield good performance when much less features were used. As shown in Figure 3, when top 60 features were used, the SVM classifier can generate the MCC of 0.686, which was still higher than that of the optimum RF classifier. The detailed performance of such SVM classifier is listed in Table 1. It can be seen that all measurements, except SN, of this SVM classifier were higher than those of the optimum RF classifier. Thus, we picked up such SVM classifier as the proposed classifier because it can provide good performance and was much more efficient than the optimum SVM classifier.

3.3. Rule Learning Results. The optimum SVM and RF classifiers gave good performance. However, they were black-box algorithms. Few medical insights can be captured from these classifiers. In view of this, we further employed a rule learning procedure. The informative features yielded by MCFS were processed by the Johnson Reducer and RIPPER algorithms one by one. As a result, three rules were constructed, as shown in Table 2, where two rules were for prediction of IFX-sensitive patients and the last one was for identification of IFX-resistant patients. We counted two measurements: support and accuracy for each rule, as listed in Table 2. Each rule covered some samples, and all accuracies were quite high, implying the utility of the rules.

Furthermore, to test the effectiveness of the above rule learning procedures, we did the 10-fold cross-validation three times. Six measurements calculated by equations (2) and (3) were counted. The MCC was 0.439, and other five measurements were 0.421 (SN), 0.939 (SP), 0.783 (ACC), 0.746 (precision), and 0.538 ($F1$-measure), respectively. Clearly, the rule classifier was inferior to the optimum SVM and RF classifiers. However, it can explain the detailed gene expression pattern for distinguishing patients with RA who respond to IFX treatment or not.

4. Discussion

IFX is one of the major clinically applied drugs for RA. However, the sensitivity and effectiveness of this drug vary among patients. Recent publications confirmed that the sensitivity of this drug against RA can be predicted by obtaining the expression profiling pattern of patients' pretherapeutic blood. However, the core signatures/biomarkers for the prediction and understanding of IFX sensitivity are difficult to identify. We identified gene signatures for drug therapeutic effect evaluation and established a series of quantitative rules that explain the detailed accurate recognition of patients with different IFX sensitivity using a novel computational approach on the expression profiling of pretherapeutic blood. All the identified signatures have been confirmed by recent publications, and the detailed analysis of the representative genes and rules is discussed below.

4.1. Gene Signatures Associated with IFX Response. In this study, with some computational methods, several genes associated with IFX response were identified. Here, we selected some of them for detailed analysis, which are listed in Table 3.

DISC1, which encodes a scaffold protein, participates in the synthesis of hemoglobin in peripheral blood [62, 63]. Although no direct evidence confirmed that the blood expression of *DISC1* may directly contribute to the pathogenesis of RA, recent publications validated that *DISC1* may

Table 2: Classification rules yielded by RIPPER.

Index	Condition	Result	Support#	Accuracy$
1	DISC1 ≤ 5.2301	IFX-sensitive patient	7.86%	90.91%
2	SAMD11 ≥ 4.2938	IFX-sensitive patient	7.14%	90.00%
3	Others	IFX-resistant patient	85.00%	80.67%

#Support is defined as the proportion of samples satisfying the rule to all samples. $Accuracy is defined as the proportion of correctly predicted samples to the samples satisfying the rule.

Table 3: Some top genes associated with IFX response.

Gene symbol	Description	RI score
DISC1	DISC1 scaffold protein	0.0506
SAMD11	Sterile alpha motif domain containing 11	0.0410
EID2B	EP300 interacting inhibitor of differentiation 2B	0.0345
NTS	Neurotensin	0.0257
STAT2	Signal transducer and activator of transcription 2	0.0233
HELZ	Helicase with zinc finger	0.0198
SUMO2	Small ubiquitin-like modifier 2	0.0190

participate in TNF-α-associated biological processes [64, 65]. A further study on the biological processes of TNF-α receptors reported that our predicted gene, *DISC1*, may be functionally related to MIPT3, a microtubule-interacting protein associated with TNF receptors; thus, *DISC1* has potential regulatory effects on TNF-associated biological processes [66]. The therapeutic effects of IFX rely on the pharmacological regulation on TNF-α-associated biological processes [67, 68]; therefore, as a regulator for the TNF receptor, the different expression patterns of *DISC1* can indicate different TNF-related biological processes and affect the pharmacological effects of IFX.

SAMD11, as a transcription coactivator, contributes to the development of a photoreceptor [69]. In 2009, *SAMD11* was identified as a susceptibility gene for RA by high-throughput genotyping techniques [70]. Multiple publications have also confirmed its specific role in inflammatory regulation [71, 72] and even validated its functional relationship with IL17 [73, 74]. The pharmacological effects of IFX are mediated by TNF-α inhibition; thus, IFX is functionally related to the regulation of immune responses via the interactions between IL17 and TNF-α [75, 76]. As a specific regulator of IL17, our candidate gene, *SMAD11*, may also participate in the regulation of IFX-mediated pharmacological responses in patients with RA. Therefore, the expression level of *SMAD11* in the blood may indirectly reflect the reactiveness of IFX in patients with RA.

EID2B, as an RA-associated gene, participates in MyoD-dependent transcription, glucocorticoid receptor-dependent transcription, and muscle differentiation as a functional repressor [77, 78]. MyoD-associated biological functions induce muscle lesions by interacting with abnormal regulation on TNF-α pathways, targeted by IFX during the initiation and progression of RA [79]. Further studies on the pharmacological effects of IFX confirmed that MyoD may directly participate in IFX-associated therapeutic metabolism in Crohn's disease, which is another autoimmune disease [80]. Our predicted gene *EID2B* may also participate in the regulation of the therapeutic effects of IFX by interacting with the core regulator, MyoD. Therefore, the expression level of *EID2B* in the peripheral blood may reflect *EID2B* expression in mesenchymal cells and immune cells in the blood and may also be a novel parameter for the prediction of the prognosis of IFX-dependent therapeutics.

NTS, which encodes a common precursor for peptides neuromedin N and neurotensin, participates in the regulation of fat metabolism in muscles [81, 82]. A recent research confirmed its specific pathogenic role in the pathophysiology of RA [83]. As for its distinctive role for the sensitivity of IFX, NTS participates in TNF-α-associated biological processes [84, 85]. Considering that IFX acts as the inhibitor of TNF-α-associated biological processes [23], the therapeutic effects of IFX are functionally associated with the abnormal activation status of TNF-α-associated biological processes. Therefore, as an effective participator of TNF-α-associated biological processes, the expression pattern of NTS may reflect the activation status of TNF-α-associated biological processes and therefore indicates the therapeutic effects of IFX.

STAT2, as an effective member of the STAT protein family, participates in the transcriptional regulation mediated by type I interferons (IFNs) [86] and the Jak kinase signaling cascade [87, 88]. The treatment effects of IFX are functionally connected to the IFN signaling cascade [89–91]; hence, IFN-associated biological processes may be crucial for the pharmacological effects of IFX. Therefore, as a transcriptional regulator at the downstream of IFN-associated biological processes, the expression level of our predicted gene *STAT2* may also be an alternative following the expression alteration of genes in the type I IFN-associated pathways. This finding indicates the potential identification ability of *STAT2* on the therapeutic effects of IFX.

The *HELZ* gene can encode a member of the RNA helicase superfamily I class. *HELZ* participates in RNA hydrolysis in multiple tissues [92]. In fact, TNF-α and its related immune regulatory functions are regulated by multiple RNA helicases, including helicase with zinc finger (HELZ) [93–95]. Therefore, the expression pattern of *HELZ* may also affect the therapeutic effects of IFX by interfering with TNF-α-associated biological processes. Similarly, *SUMO2* also affects IFX sensitivity by interfering with TNF-α-associated biological processes [96, 97].

Apart from unannotated RNA transcripts with no validated protein products, all the predicted genes have been confirmed to be functionally related to TNF-α-associated biological processes in either physical or pathological conditions. Therefore, these genes may further affect the therapeutic effects and sensitivity of IFX in patients with RA. This finding validates the efficacy and accuracy of our prediction method and analysis.

4.2. Signature Rules Associated with IFX Response. Apart from the qualitative analysis of each top-ranked gene signatures in our prediction list, we also set up a series of quantitative recognition rules for the detailed and accurate recognition of IFX-sensitive and IFX-resistant patients. The first rule involves only the *DISC1* gene. A low *DISC1* expression (<5 FPKM) indicates that the patient may be sensitive to IFX. Based on the analysis, a low *DISC1* expression may indicate abnormal TNF-α-associated immune activation status in RA [66]. Therefore, patients with a low *DISC1* expression pattern may have an activated TNF-associated signaling pathway and are definitely sensitive to therapeutic effects. The expression level of *DISC1* is quite high in normal conditions (>10 FPKM). Therefore, our threshold may indicate low *DISC1* expression level and corresponds to our analysis above.

The second rule involves the *SAMD11* gene. Different from *DISC1*, a high *SAMD11* expression level may indicate an activated inflammatory status in the whole body and a high TNF-α expression level [75, 76]. Therefore, the therapeutic effects of IFX may also be more effective in conditions with more potential pharmacological targets. This finding corresponds to our prediction rules. According to recent publications, the expression level of *SAMD11* in normal whole blood is <1 FPKM. Therefore, under the predicted conditions, the expression level of *SAMD11* may be upregulated and result in the stronger activation of the TNF-α signaling pathway. Patients with specific *SAMD11* expression levels that follow our rules would be sensitive to IFX. By contrast, patients with expression profiling that does not follow these quantitative rules may be resistant to IFX-mediated RA therapeutics.

5. Conclusions

The identified blood gene signatures participate in IFX-sensitive pharmacological processes in patients with RA. Thus, these genes may be potential biomarkers for the distinction of IFX-sensitive and IFX-resistant patients at the transcriptomic level. Several quantitative signature rules for the distinction of patients have also been verified by other recent publications. Therefore, our newly presented method provides comprehensive qualitative and quantitative prediction standards for prognosis guidance on the clinical application of IFX on patients with RA.

Data Availability

The data used to support the findings of this study have been deposited in the Gene Expression Omnibus repository (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE78068).

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Authors' Contributions

ShiJian Ding and ZhanDong Li contributed equally to this work.

Acknowledgments

This research was funded by the Strategic Priority Research Program of Chinese Academy of Sciences (XDB38050200), the National Key R&D Program of China (2017YFC1201200, 2018YFC0910403), the Shanghai Municipal Science and Technology Major Project (2017SHZDZX01), the National Natural Science Foundation of China (31701151), the Shanghai Sailing Program (16YF1413800), the Youth Innovation Promotion Association of Chinese Academy of Sciences (CAS) (2016245), and the Fund of the Key Laboratory of Tissue Microenvironment and Tumor of Chinese Academy of Sciences (202002).

Supplementary Materials

Table S1: list of genes ranked based on the RI score from MCFS. Table S2: performance of IFS with SVM. Table S3: performance of IFS with RF. *(Supplementary materials)*

References

[1] G. R. Burmester and J. E. Pope, "Novel treatment strategies in rheumatoid arthritis," *Lancet*, vol. 389, no. 10086, pp. 2338–2348, 2017.

[2] J. S. Smolen, D. Aletaha, and I. B. McInnes, "Rheumatoid arthritis," *Lancet*, vol. 388, no. 10055, pp. 2023–2038, 2016.

[3] R. C. Lawrence, D. T. Felson, C. G. Helmick et al., "Estimates of the prevalence of arthritis and other rheumatic conditions in the United States: part II," *Arthritis Rheum*, vol. 58, no. 1, pp. 26–35, 2008.

[4] C. G. Helmick, D. T. Felson, R. C. Lawrence et al., "Estimates of the prevalence of arthritis and other rheumatic conditions in the United States: part I," *Arthritis Rheum*, vol. 58, no. 1, pp. 15–25, 2008.

[5] R. N. Maini and M. Feldmann, "Cytokine therapy in rheumatoid arthritis," *Lancet*, vol. 348, no. 9030, pp. 824-825, 1996.

[6] A. J. Hayes, "Angioneogenesis in rheumatoid arthritis," *Lancet*, vol. 354, no. 9176, pp. 423-424, 1999.

[7] I. B. McInnes and G. Schett, "Pathogenetic insights from the treatment of rheumatoid arthritis," *Lancet*, vol. 389, no. 10086, pp. 2328–2337, 2017.

[8] G. R. Burmester, J. M. Kremer, F. van den Bosch et al., "Safety and efficacy of upadacitinib in patients with rheumatoid arthritis and inadequate response to conventional synthetic disease-modifying anti-rheumatic drugs (SELECT-NEXT): a randomised, double-blind, placebo-controlled phase 3 trial," *Lancet*, vol. 391, no. 10139, pp. 2503–2512, 2018.

[9] J. M. Torpy, G. D. Perazza, and R. M. Golub, "JAMA patient page. Rheumatoid arthritis," *JAMA*, vol. 305, no. 17, p. 1824, 2011.

[10] R. Ferrari, “An examination of expectations for rheumatoid arthritis disability in Germany: comparison with Canadian data,” *Seminars in Arthritis and Rheumatism*, vol. 44, no. 2, p. e4, 2014.

[11] T. Mine, K. Ihara, H. Kawamura, R. Kuriyama, and R. Date, “Knee arthritis without other joint symptoms in the elderly with seronegative elderly onset rheumatoid arthritis,” *The Open Orthopaedics Journal*, vol. 10, no. 1, pp. 793–796, 2016.

[12] J. Isacson, E. Allander, and L. A. Brostrom, “17-year follow-up of symptoms and signs in the knee joint in rheumatoid arthritis,” *Scandinavian Journal of Rheumatology*, vol. 17, no. 5, pp. 325–331, 1988.

[13] F. Davatchi, I. S. Abari, S. Soroosh, M. Soroosh, and B. S. Abdollahi, “Performance of the 2010 rheumatoid arthritis classification criteria,” *International Journal of Rheumatic Diseases*, vol. 15, no. 5, pp. 455–461, 2012.

[14] M. Infantino, M. Manfredi, F. Meacci et al., “Anti-citrullinated peptide antibodies and rheumatoid factor isotypes in the diagnosis of rheumatoid arthritis: an assessment of combined tests,” *Clinica Chimica Acta*, vol. 436, pp. 237–242, 2014.

[15] R. H. Mackey, L. H. Kuller, K. D. Deane et al., “Rheumatoid arthritis, anti-cyclic citrullinated peptide positivity, and cardiovascular disease risk in the Women's Health Initiative,” *Arthritis & Rhematology*, vol. 67, no. 9, pp. 2311–2322, 2015.

[16] E. de Miguel, T. E. C. O.-D. A. I. Group, A. Pecondón-Español et al., “A reduced 12-joint ultrasound examination predicts lack of X-ray progression better than clinical remission criteria in patients with rheumatoid arthritis,” *Rheumatology International*, vol. 37, no. 8, pp. 1347–1356, 2017.

[17] P. Sivakumaran, S. Hussain, L. Attipoe, and C. Ciurtin, “Diagnostic accuracy of simplified ultrasound hand examination protocols for detection of inflammation and disease burden in patients with rheumatoid arthritis,” *Acta Radiologica*, vol. 60, no. 1, pp. 92–99, 2019.

[18] S. Witulski, T. J. Vogl, S. Rehart, and P. Ottl, “Evaluation of the TMJ by means of clinical TMD examination and MRI diagnostics in patients with rheumatoid arthritis,” *BioMed Research International*, vol. 2014, Article ID 328560, 2014.

[19] D. Vanags, B. Williams, B. Johnson et al., “Therapeutic efficacy and safety of chaperonin 10 in patients with rheumatoid arthritis: a double-blind randomised trial,” *Lancet*, vol. 368, no. 9538, pp. 855–863, 2006.

[20] L. Klareskog, D. van der Heijde, J. P. de Jager et al., “Therapeutic effect of the combination of etanercept and methotrexate compared with each treatment alone in patients with rheumatoid arthritis: double-blind randomised controlled trial,” *Lancet*, vol. 363, no. 9410, pp. 675–681, 2004.

[21] M. Fechtenbaum, J. L. Nam, and P. Emery, “Biologics in rheumatoid arthritis: where are we going?,” *British Journal of Hospital Medicine (London, England)*, vol. 75, no. 8, pp. 448–456, 2014, 451-6.

[22] K. E. Donahue, G. Gartlehner, D. E. Jonas et al., “Systematic review: comparative effectiveness and harms of disease-modifying medications for rheumatoid arthritis,” *Annals of Internal Medicine*, vol. 148, no. 2, pp. 124–134, 2008.

[23] G. Martius, S. Cameron, M. Rave-Fränk, C. Hess, H. Wolff, and I. Malik, “The anti-TNF-α antibody infliximab inhibits the expression of FAT-transporter-protein FAT/CD36 in a selective hepatic-radiation mouse model,” *International Journal of Molecular Sciences*, vol. 16, no. 3, pp. 4682–4697, 2015.

[24] J. K. Eriksson, J. K. Wallman, H. Miller et al., “Infliximab versus conventional combination treatment and seven-year work loss in early rheumatoid arthritis: results of a randomized Swedish trial,” *Arthritis Care Res (Hoboken)*, vol. 68, no. 12, pp. 1758–1766, 2016.

[25] D. Dimopoulou, T. Dimitroulas, E. Akriviadis, and A. Garyfallos, “Infliximab as a treatment option for patients with rheumatoid arthritis and primary biliary cirrhosis,” *Rheumatology International*, vol. 35, no. 11, pp. 1913–1916, 2015.

[26] S. Nakamura, K. Suzuki, H. Iijima et al., “Identification of baseline gene expression signatures predicting therapeutic responses to three biologic agents in rheumatoid arthritis: a retrospective observational study,” *Arthritis Research & Therapy*, vol. 18, no. 1, p. 159, 2016.

[27] M. Draminski, A. Rada-Iglesias, S. Enroth, C. Wadelius, J. Koronacki, and J. Komorowski, “Monte Carlo feature selection for supervised classification,” *Bioinformatics*, vol. 24, no. 1, pp. 110–117, 2008.

[28] X. Pan, T. Zeng, Y.-H. Zhang et al., “Investigation and prediction of human interactome based on quantitative features,” *Frontiers in Bioengineering and Biotechnology*, vol. 8, p. 730, 2020.

[29] Y.-H. Zhang, Z. Li, T. Zeng, W. C. Lu, T. Huang, and Y.-D. Cai, “Identifying the immunological gene signatures of immune cell subtypes,” *BioMed Research International*, vol. 2021, Article ID 6639698, 2021.

[30] H. A. Liu and R. Setiono, “Incremental feature selection,” *Applied Intelligence*, vol. 9, no. 3, pp. 217–230, 1998.

[31] B. Matthews, “Comparison of the predicted and observed secondary structure of T4 phage lysozyme,” *Biochimica et Biophysica Acta (BBA)-Protein Structure*, vol. 405, no. 2, pp. 442–451, 1975.

[32] R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” in *In International joint Conference on artificial intelligence*, Lawrence Erlbaum Associates Ltd., 1995.

[33] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[34] J.-P. Zhou, L. Chen, and Z.-H. Guo, “iATC-NRAKEL: an efficient multi-label classifier for recognizing anatomical therapeutic chemical classes of drugs,” *Bioinformatics*, vol. 36, no. 5, pp. 1391–1396, 2020.

[35] J.-P. Zhou, L. Chen, T. Wang, and M. Liu, “iATC-FRAKEL: a simple multi-label web server for recognizing anatomical therapeutic chemical classes of drugs with their fingerprints only,” *Bioinformatics*, vol. 36, no. 11, pp. 3568-3569, 2020.

[36] S. P. Wang, Q. Zhang, J. Lu, and Y.-D. Cai, “Analysis and prediction of nitrated tyrosine sites with the mRMR method and support vector machine algorithm,” *Current Bioinformatics*, vol. 13, no. 1, pp. 3–13, 2018.

[37] X. G. Chen, W. W. Shi, and L. Deng, “Prediction of disease comorbidity using HeteSim scores based on multiple heterogeneous networks,” *Current Gene Therapy*, vol. 19, no. 4, pp. 232–241, 2019.

[38] Y. Zhu, B. Hu, L. Chen, and Q. Dai, “iMPTCE-Hnetwork: a multi-label classifier for identifying metabolic pathway types of chemicals and enzymes with a heterogeneous network,” *Computational and Mathematical Methods in Medicine*, vol. 2021, Article ID 6683051, 2021.

[39] F. Yuan, Z. Li, L. Chen et al., “Identifying the signatures and rules of circulating extracellular microRNA for distinguishing

cancer subtypes," *Frontiers in Genetics*, vol. 12, p. 651610, 2021.

[40] E. Frank, M. Hall, L. Trigg, G. Holmes, and I. H. Witten, "Data mining in bioinformatics using Weka," *Bioinformatics*, vol. 20, no. 15, pp. 2479–2481, 2004.

[41] I. H. Witten and E. Frank, *Data Mining:Practical Machine Learning Tools and Techniques*, San Francisco, Morgan, Kaufmann, 2005.

[42] J. Platt, *Sequential minimal optimizaton: a fast algorithm for training support vector machines*, Technical Report MSR-TR-98-14, 1998.

[43] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[44] X. Y. Pan, Y. N. Zhang, and H. B. Shen, "Large-scale prediction of human protein-protein interactions from amino acid sequence based on latent topic features," *Journal of Proteome Research*, vol. 9, no. 10, pp. 4992–5001, 2010.

[45] Y. Jia, R. Zhao, and L. Chen, "Similarity-based machine learning model for predicting the metabolic pathways of compounds," *IEEE Access*, vol. 8, pp. 130687–130696, 2020.

[46] H. Liang, L. Chen, X. Zhao, and X. Zhang, "Prediction of drug side effects with a refined negative sample selection strategy," *Computational and Mathematical Methods in Medicine*, vol. 2020, Article ID 1573543, 2020.

[47] X. Zhao, L. Chen, and J. Lu, "A similarity-based method for prediction of drug side effects with heterogeneous information," *Mathematical Biosciences*, vol. 306, pp. 136–144, 2018.

[48] X. Zhao, L. Chen, Z. H. Guo, and T. Liu, "Predicting drug side effects with compact integration of heterogeneous networks," *Current Bioinformatics*, vol. 14, no. 8, pp. 709–720, 2019.

[49] J. R. Li, L. Lu, Y.-. H. Zhang et al., "Identification of synthetic lethality based on a functional network by using machine learning algorithms," *Journal of Cellular Biochemistry*, vol. 120, no. 1, pp. 405–416, 2018.

[50] E. Kwon, M. Cho, H. Kim, and H. S. Son, "A study on host tropism determinants of influenza virus using machine learning," *Current Bioinformatics*, vol. 15, no. 2, pp. 121–134, 2020.

[51] P. Chen, T. Shen, Y. Zhang, and B. Wang, "A sequence-segment neighbor encoding schema for protein hotspot residue prediction," *Current Bioinformatics*, vol. 15, no. 5, pp. 445–454, 2020.

[52] Y.-H. Zhang, H. Li, T. Zeng et al., "Identifying transcriptomic signatures and rules for SARS-CoV-2 infection," *Frontiers in Cell and Developmental Biology*, vol. 8, p. 627302, 2021.

[53] D. S. Johnson, "Approximation algorithms for combinatorial problems," *Journal of Computer and System Sciences*, vol. 9, no. 3, pp. 256–278, 1974.

[54] W. W. Cohen, "Fast effective rule induction," in *Twelfth International Conference on Machine Learning*, pp. 115–123, Tahoe City, California, 1995.

[55] L. Chen, S. Wang, Y. H. Zhang et al., "Identify key sequence features to improve CRISPR sgRNA efficacy," *IEEE Access*, vol. 5, pp. 26582–26590, 2017.

[56] L. Chen, C. Chu, Y.-H. Zhang et al., "Identification of drug-drug interactions using chemical interactions," *Current Bioinformatics*, vol. 12, no. 6, pp. 526–534, 2017.

[57] L. Liu, L. Chen, Y. H. Zhang et al., "Analysis and prediction of drug-drug interaction by minimum redundancy maximum relevance and incremental feature selection," *Journal of Biomolecular Structure & Dynamics*, vol. 35, no. 2, pp. 312–329, 2017.

[58] H. Liu, B. Hu, L. Chen, and L. Lu, "Identifying protein subcellular location with embedding features learned from networks," *Current Proteomics*, vol. 17, 2020.

[59] X. Pan, H. Li, T. Zeng et al., "Identification of protein subcellular localization with network and functional embeddings," *Frontiers in Genetics*, vol. 11, p. 626500, 2021.

[60] Y.-H. Zhang, Z. Li, T. Zeng et al., "Detecting the multiomics signatures of factor-specific inflammatory effects on airway smooth muscles," *Frontiers in Genetics*, vol. 11, p. 599970, 2021.

[61] Y.-H. Zhang, T. Zeng, L. Chen, T. Huang, and Y. D. Cai, "Determining protein-protein functional associations by functional rules based on gene ontology and KEGG pathway," *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, vol. 1869, no. 6, p. 140621, 2021.

[62] J. R. Liu, Q. Liu, J. Khoury et al., "Hypoxic preconditioning decreases nuclear factor *κ*B activity via disrupted in schizophrenia-1," *The International Journal of Biochemistry & Cell Biology*, vol. 70, pp. 140–148, 2016.

[63] A. Mohammadi, E. Rashidi, and V. G. Amooeian, "Brain, blood, cerebrospinal fluid, and serum biomarkers in schizophrenia," *Psychiatry Research*, vol. 265, pp. 25–38, 2018.

[64] C. Noto, V. K. Ota, M. L. Santoro et al., "Depression, cytokine, and cytokine by treatment interactions modulate gene expression in antipsychotic naïve first episode psychosis," *Molecular Neurobiology*, vol. 53, no. 8, pp. 5701–5709, 2016.

[65] N. F. Berbari, N. W. Kin, N. Sharma, E. J. Michaud, R. A. Kesterson, and B. K. Yoder, "Mutations in Traf3ip1 reveal defects in ciliogenesis, embryonic development, and altered cell size regulation," *Developmental Biology*, vol. 360, no. 1, pp. 66–76, 2011.

[66] J. A. Morris, G. Kandpal, L. Ma, and C. P. Austin, "DISC1 (disrupted-in-schizophrenia 1) is a centrosome-associated protein that interacts with MAP1A, MIPT3, ATF4/5 and NUDEL: regulation and loss of interaction with mutation," *Human Molecular Genetics*, vol. 12, no. 13, pp. 1591–1608, 2003.

[67] I. Voloshyna, S. Seshadri, K. Anwar et al., "Infliximab Reverses Suppression of Cholesterol Efflux Proteins by TNF-*α*: A Possible Mechanism for Modulation of Atherogenesis," *BioMed Research International*, vol. 2014, Article ID 312647, 8 pages, 2014.

[68] S. Yoshida, T. Takeuchi, T. Kotani et al., "Infliximab, a TNF-*α* inhibitor, reduces 24-h ambulatory blood pressure in rheumatoid arthritis patients," *Journal of Human Hypertension*, vol. 28, no. 3, pp. 165–169, 2014.

[69] M. Corton, A. Avila-Fernández, L. Campello et al., "Identification of the photoreceptor transcriptional co-repressor SAMD11 as novel cause of autosomal recessive retinitis pigmentosa," *Scientific Reports*, vol. 6, no. 1, p. 35370, 2016.

[70] Q. Sha, R. Tang, and S. Zhang, "Detecting susceptibility genes for rheumatoid arthritis based on a novel sliding-window approach," *BMC Proc*, vol. 3, Supplement 7, p. S14, 2009.

[71] W. Liu, Q. Ma, K. Wong et al., "Brd4 and JMJD6-associated anti-pause enhancers in regulation of transcriptional pause release," *Cell*, vol. 155, no. 7, pp. 1581–1595, 2013.

[72] A. M. Zawada, J. S. Schneider, A. I. Michel et al., "DNA methylation profiling reveals differences in the 3 human monocyte subsets and identifies uremia to induce DNA methylation changes during differentiation," *Epigenetics*, vol. 11, no. 4, pp. 259–272, 2016.

[73] J. Diegelmann, T. Olszak, B. Göke, R. S. Blumberg, and S. Brand, "A novel role for interleukin-27 (IL-27) as mediator

of intestinal epithelial barrier protection mediated via differential signal transducer and ctivator of transcription (STAT) protein signaling and induction of antibacterial and anti-inflammatory proteins," *The Journal of Biological Chemistry*, vol. 287, no. 1, pp. 286–298, 2012.

[74] Y. Tian, H. Shen, L. Xia, and J. Lu, "Elevated serum and synovial fluid levels of interleukin-34 in rheumatoid arthritis: possible association with disease progression via interleukin-17 production," *Journal of Interferon & Cytokine Research*, vol. 33, no. 7, pp. 398–401, 2013.

[75] T. Hennerici, R. Pollmann, T. Schmidt et al., "Increased frequency of T follicular helper cells and elevated interleukin-27 plasma levels in patients with pemphigus," *PLoS One*, vol. 11, no. 2, article e0148919, 2016.

[76] D. A. Alves da Silva, M. V. da Silva, C. C. Oliveira Barros et al., "TNF-α blockade impairs in vitro tuberculous granuloma formation and down modulate Th1, Th17 and Treg cytokines," *PLoS One*, vol. 13, no. 3, article e0194430, 2018.

[77] E. P. D. Reis, D. M. Paixão, O. J. B. Brustolini et al., "Expression of myogenes in longissimus dorsi muscle during prenatal development in commercial and local Piau pigs," *Genetics and Molecular Biology*, vol. 39, no. 4, pp. 589–599, 2016.

[78] E. L. Huttlin, R. J. Bruckner, J. A. Paulo et al., "Architecture of the human interactome defines protein communities and disease networks," *Nature*, vol. 545, no. 7655, pp. 505–509, 2017.

[79] R. J. Boutrup, J. Farup, K. Vissing, M. Kjaer, and U. R. Mikkelsen, "Skeletal muscle stem cell characteristics and myonuclei content in patients with rheumatoid arthritis: a cross-sectional study," *Rheumatology International*, vol. 38, no. 6, pp. 1031–1041, 2018.

[80] K. Subramaniam, K. Fallon, T. Ruut et al., "Infliximab reverses inflammatory muscle wasting (sarcopenia) in Crohn's disease," *Alimentary Pharmacology & Therapeutics*, vol. 41, no. 5, pp. 419–428, 2015.

[81] J. Li, J. Song, Y. Y. Zaytseva et al., "An obligatory role for neurotensin in high-fat-diet-induced obesity," *Nature*, vol. 533, no. 7603, pp. 411–415, 2016.

[82] J. Piątek, H. Witmanowski, J. Paluszak, H. Krauss, and J. Krawczyk, "The effects of neurotensin on selected parameters of lipid metabolism in rats," *Peptides*, vol. 26, no. 5, pp. 837–843, 2005.

[83] J. P. Famaey, T. Appelboom, and J. Wybran, "The nervous system and the pathophysiology of rheumatoid arthritis: the role of neurotensin and opioid peptides," *The Journal of Rheumatology*, vol. 13, no. 3, pp. 651-652, 1986.

[84] D. Zhao, Y. Zhan, H. W. Koon et al., "Metalloproteinase-dependent transforming growth factor-α release mediates neurotensin-stimulated MAP kinase activation in hhuman colonic epithelial cells," *The Journal of Biological Chemistry*, vol. 279, no. 42, pp. 43547–43554, 2004.

[85] L. I. F. Moura, A. M. A. Dias, E. Suesca et al., "Neurotensin-loaded collagen dressings reduce inflammation and improve wound healing in diabetic mice," *Biochimica et Biophysica Acta*, vol. 1842, no. 1, pp. 32–43, 2014.

[86] C. Yue, J. Xu, M. D. Tan Estioko et al., "Host STAT2/type I interferon axis controls tumor growth," *International Journal of Cancer*, vol. 136, no. 1, pp. 117–126, 2015.

[87] M. L. Slattery, A. Lundgreen, S. A. Kadlubar, K. L. Bondurant, and R. K. Wolff, "JAK/STAT/SOCS-signaling pathway and colon and rectal cancer," *Molecular Carcinogenesis*, vol. 52, no. 2, pp. 155–166, 2013.

[88] H. Wu, Y. Wu, Z. Ai et al., "Vitamin C enhances Nanog expression via activation of the JAK/STAT signaling pathway," *Stem Cells*, vol. 32, no. 1, pp. 166–176, 2014.

[89] L. G. van Baarsen, C. A. Wijbrandts, F. Rustenburg et al., "Regulation of IFN response gene activity during infliximab treatment in rheumatoid arthritis is associated with clinical response to treatment," *Arthritis Research & Therapy*, vol. 12, no. 1, p. R11, 2010.

[90] M. Takeshita, K. Suzuki, J. Kikuchi et al., "Infliximab and etanercept have distinct actions but similar effects on cytokine profiles in rheumatoid arthritis," *Cytokine*, vol. 75, no. 2, pp. 222–227, 2015.

[91] C. P. Mavragani, D. T. la, W. Stohl, and M. K. Crow, "Association of the response to tumor necrosis factor antagonists with plasma type I interferon activity and interferon-beta/alpha ratios in rheumatoid arthritis patients: a post hoc analysis of a predominantly Hispanic cohort," *Arthritis and Rheumatism*, vol. 62, no. 2, pp. 392–401, 2010.

[92] P. A. Hasgall, D. Hoogewijs, M. B. Faza, V. G. Panse, R. H. Wenger, and G. Camenisch, "The putative RNA helicase HELZ promotes cell proliferation, translation initiation and ribosomal protein S6 phosphorylation," *PLoS One*, vol. 6, no. 7, article e22107, 2011.

[93] K. Mosallanejad, Y. Sekine, S. Ishikura-Kinoshita et al., "The DEAH-box RNA helicase DHX15 activates NF-B and MAPK signaling downstream of MAVS during antiviral responses," *Science Signaling*, vol. 7, no. 323, p. ra40, 2014.

[94] V. R. R. Mendonça, L. C. L. Souza, G. C. Garcia et al., "DDX39B (BAT1), TNF and IL6 gene polymorphisms and association with clinical outcomes of patients with Plasmodium vivax malaria," *Malaria Journal*, vol. 13, no. 1, p. 278, 2014.

[95] J. Pircher, T. Czermak, M. Merkle et al., "Hepatitis C virus induced endothelial inflammatory response depends on the functional expression of TNFα receptor subtype 2," *PLoS One*, vol. 9, no. 11, article e113351, 2014.

[96] A. M. Mabb, S. M. Wuerzberger-Davis, and S. Miyamoto, "PIASy mediates NEMO sumoylation and NF-κB activation in response to genotoxic stress," *Nature Cell Biology*, vol. 8, no. 9, pp. 986–993, 2006.

[97] J. Liu, M. Sha, Q. Wang et al., "Small ubiquitin-related modifier 2/3 interacts with p65 and stabilizes it in the cytoplasm in HBV-associated hepatocellular carcinoma," *BMC Cancer*, vol. 15, no. 1, 2015.

Hindawi
BioMed Research International
Volume 2021, Article ID 9987819, 15 pages
https://doi.org/10.1155/2021/9987819

Research Article

Improving the Prognosis of Colon Cancer through Knowledge-Based Clinical-Molecular Integrated Analysis

Danyang Tong,[1] Yu Tian,[1] Qiancheng Ye,[1] Jun Li,[2] Kefeng Ding,[2] and Jingsong Li[1,3]

[1] *Engineering Research Center of EMR and Intelligent Expert System, Ministry of Education, College of Biomedical Engineering and Instrument Science, Zhejiang University, No. 38 Zheda Road, Hangzhou, 310027 Zhejiang Province, China*
[2] *Department of Surgical Oncology, Second Affiliated Hospital, Zhejiang University School of Medicine, No. 88 Jiefang Road, Hangzhou, 31009 Zhejiang Province, China*
[3] *Research Center for Healthcare Data Science, Zhejiang Lab, Hangzhou, China*

Correspondence should be addressed to Jingsong Li; ljs@zju.edu.cn

Received 3 March 2021; Revised 19 March 2021; Accepted 23 March 2021; Published 7 April 2021

Academic Editor: Quan Zou

Background. Colon cancer has high morbidity and mortality rates among cancers. Existing clinical staging systems cannot accurately assess the prognostic risk of colon cancer patients. This study was aimed at improving the prognostic performance of the colon cancer clinical staging system through knowledge-based clinical-molecular integrated analysis. *Methods*. 374 samples from The Cancer Genome Atlas Colon Adenocarcinoma (TCGA-COAD) dataset were used as the discovery set. 98 samples from the Clinical Proteomic Tumor Analysis Consortium (CPTAC) dataset were used as the validation set. After converting gene expression data into pathway dysregulation scores (PDSs), the random survival forest and Cox model were used to identify the best prognostic supplementary factors. The corresponding clinical-molecular integrated prognostic model was built, and the improvement of prognostic performance was assessed by comparing with the clinical prognostic model. *Results*. The PDS of 14 pathways played important roles in prognostic prediction together with clinical prognostic factors through the random survival forest. Further screening with the Cox model revealed that the PDS of the pathway hsa00532 was the best clinical prognostic supplementary factor. The integrated prognostic model constructed with clinical factors and the identified molecular factor was superior to the clinical prognostic model in discriminative performance. Kaplan-Meier (KM) curves of patients grouped by PDS suggested that patients with a higher PDS had a poorer prognosis, and stage II patients could be distinctly distinguished. *Conclusions*. Based on the knowledge-based clinical-molecular integrated analysis, a clinical-molecular integrated prognostic model and corresponding nomogram for colon cancer overall survival prognosis was built, which showed better prognostic performance than the clinical prognostic model. The PDS of the pathway hsa00532 is a considerable clinical prognostic supplementary factor for colon cancer and may represent a potential prognostic marker for stage II colon cancer. The PDS calculation involves only 16 genes, which supports its potential for clinical application.

1. Introduction

Colon cancer is one of the top cancers in terms of incidence and mortality in both China and America [1, 2]. Recent global surveillance of cancer trends revealed that further research on colon cancer is needed, as the age-standardized 5-year net survival of colon cancer ranges from approximately 15% to 75% in different countries [3].

Currently, the tumor, node, and metastasis (TNM) stage system proposed by the American Joint Committee on Cancer (AJCC) is the most commonly used clinical staging tool for colon cancer. However, the accuracy of the 7th TNM staging system for assessing the prognostic risk of colorectal cancer patients still needs to be improved, especially for stage II and stage IIIA patients [4]. The 8th edition of the TNM staging system was aimed at building an important bridge from a “population-based” to a more “personalized” approach to cancer stage [5]. The 8th edition of the TNM staging system for breast cancer did so by including the HER2 and ER statuses in its prognostic staging [6]. Several studies claimed that modifications of the TNM staging system for colorectal cancer showed improved prognostic performance [4, 7, 8]. However, no structural changes were made in the 8th edition of the TNM staging system for colon cancer [9]. Therefore, to achieve a more personalized prognosis for colon cancer patients, incorporating more prognostic factors in addition to current clinical prognostic factors would be a considerable choice.

Incorporating molecular factors, such as gene expression data, would be a considerable option for improving the performance of colon cancer prognosis. However, in gene expression-based analyses of heterogeneous diseases, a single gene often provides weak information [10]. However, the gene set obtained directly by analyzing a large number of genes is not stable and will change with changes in the training samples [11]. Several studies about the prognosis of colon cancer tried to select hypoxia-related genes or tumor microenvironment-related genes through literature reviews, but further screenings of these selected genes were still required in subsequent prognostic analyses [12, 13]. Therefore, the introduction of representative functional units, such as gene sets or pathways, may yield a more stable performance and may simultaneously provide certain biological annotations to improve the interpretability of the results [14–17]. In addition, converting the gene expression profile into personalized pathway activities showed a better prediction performance than using the origin gene expression profile in previous studies [18, 19].

In recent years, machine learning methods have been widely used for cancer prognostic analysis. When performing prognostic analyses through machine learning methods, the introduction of prior knowledge, such as pathway information, can further improve the performance of the model [20]. In most associated studies, molecular prognostic features were obtained by considering only the molecular features; therefore, new molecular features obtained through analysis may not be effectively combined with clinical features [21].

In this study, we conducted a knowledge-based clinical-molecular integrated analysis through a machine learning method, identified new pathway-based molecular prognostic factors to supplement the clinical TNM staging system for colon cancer overall survival prognosis prediction, and verified the improved performance of the clinical-molecular integrated prognostic models compared to the clinical prognostic model.

2. Materials and Methods

2.1. Data Acquisition and Processing. Gene mRNA expression data from primary tumors and related clinical data of 452 patients in The Cancer Genome Atlas Colon Adenocarcinoma (TCGA-COAD) project were obtained from cBioPortal as the discovery set, and gene expression data from normal adjacent tissues of 41 patients in the TCGA-COAD were obtained from the UCSC Xena as the reference set [22]. The mRNA sequence data of the discovery set and reference set used in this study were generated with the Illumina HiSeq 2000 platform and processed by the RNAseqV2 pipeline, which uses RNA-Seq by expectation maximization upper quartile (RSEM-UQ) for quantification. To validate the prognostic performance of the identified pathway-based factors, one independent dataset that offered identical clinical data and gene mRNA expression from primary tumors generated with a similar pipeline of 106 colon cancer patients was obtained from the Clinical Proteomic Tumor Analysis Consortium (CPTAC) from the LinkedOmics as the validation set [23]. The mRNA sequence data of the validation set used in this study were generated with the Illumina HiSeq 4000 platform and processed by the RNAseqV2 pipeline with RSEM-UQ for quantification. Both datasets can be used for an integrated analysis of clinical data and omics data.

Patients with primary tumors with both clinical data and gene expression data in the discovery set and validation set were included in this study. All data were cleaned and checked after data acquisition. The clinical data included T, N, and M stages and overall survival information. Other clinical prognostic factors, such as age and location, were not included because this study is focused on supplementing the clinical TNM staging system. The T stage was categorized into T1, T2, T3, and T4 stages (1 = T1, 2 = T2, 3 = T3, and 4 = T4 in subsequent analyses); the N stage was categorized into N0, N1, and N2 stages (0 = N0, 1 = N1, and 2 = N2 in subsequent analyses); and the M stage was categorized into M0 and M1 stages (0 = M0 and 1 = M1 in subsequent analyses). All gene expression data values were further log-transformed (Log2 (value + 1)) for subsequent analysis.

The following exclusion criteria were applied to the samples: containing Tis, N1c, or MX; lack of clear T, N, and M stages; and invalid survival information. In gene expression data, genes that could not be targeted with accurate HUGO Gene Nomenclature Committee (HGNC) symbols in the discovery set, validation set, and reference set were removed. Besides, genes with missing expression values or zero values were removed as well.

2.2. Study Design. First, we converted the gene expression data into pathway dysregulation scores (PDSs) based on prior knowledge from the Kyoto Encyclopedia of Genes and Genomes (KEGG) human pathway database. Then, we conducted a clinical-molecular integrated analysis by combining machine learning methods and survival analysis to identify the best molecular prognostic factors among the converted pathway-based factors. Finally, a clinical-molecular integrated prognostic model was constructed using clinical factors and the identified molecular factors for overall survival prediction and compared with the corresponding clinical prognostic model. The overall pipeline of this study is shown in Figure 1.

2.3. PDS Calculation. Among the pathway-based approaches, two methods, PARADIGM and Pathifier, are widely used to estimate the pathway dysregulation information in a particular sample [24, 25]. However, PARADIGM requires pathway

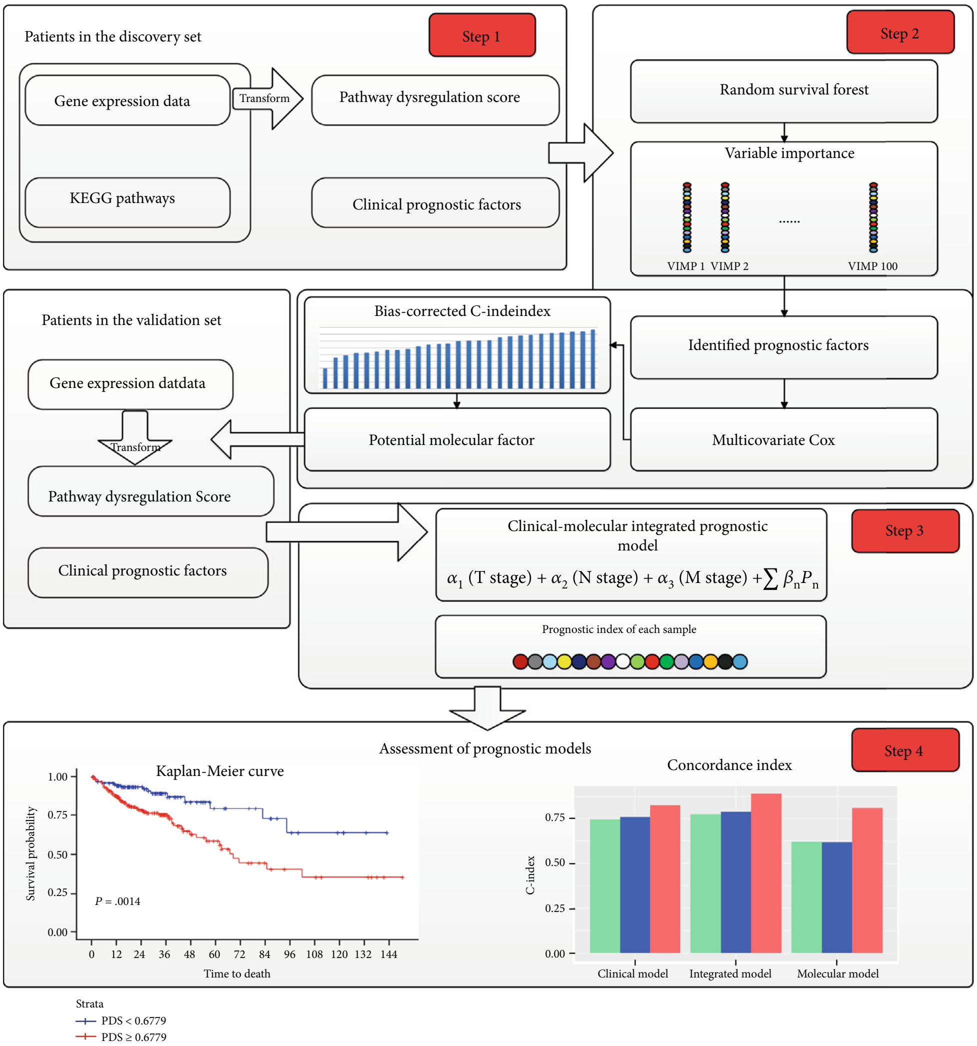

Patients in the discovery set
Step 1
Gene expression data
Transform
Pathway dysregulation score
KEGG pathways
Clinical prognostic factors
Step 2
Random survival forest
Variable importance
......
VIMP 1
VIMP 2
VIMP 100
Patients in the validation set
Gene expression datdata
Transform
Pathway dysregulation Score
Clinical prognostic factors
Bias-corrected C-indeindex
Identified prognostic factors
Potential molecular factor
Multicovariate Cox
Clinical-molecular integrated prognostic model
α_1 (T stage) + α_2 (N stage) + α_3 (M stage) +$\sum \beta_n P_n$
Step 3
Prognostic index of each sample
Assessment of prognostic models
Step 4
Kaplan-Meier curve
Survival probability
1.00
0.75
0.50
0.25
0.00
P = .0014
0
12
24
36
48
60
72
84
96
108
120
132
144
Time to death
Strata
PDS < 0.6779
PDS ≥ 0.6779
Concordance index
C-index
0.75
0.50
0.25
0.00
Clinical model
Integrated model
Molecular model

Figure 1: The overall pipeline of the study. Step 1: calculation of PDS; step 2: identification of molecular prognostic factors; step 3: construction of clinical-molecular integrative prognostic model; step 4: assessment of the integrative prognostic model.

mechanisms and is inappropriate for complex or incomplete pathways. Pathifier requires only the expression data of genes involved in each pathway and is more suitable for this study. In addition, previous studies confirmed that the PDS calculated by this method can effectively characterize pathway abnormalities [17, 19, 24, 26]. The PDS quantifies the biological difference of a specific pathway between a diseased sample and normal samples with a numeric value range from 0 to 1, and it is transformed from the gene expression data by the R package Pathifier [24]. The PDS in each sample indicates the distance of deviation between the projection of a specific pathway and the projection of normal samples on the principle component curve. The pathway information was obtained from KEGG with the R package KEGGREST (version 1.26.1).

In this study, the PDSs of 327 human pathways obtained from KEGG were calculated based on this method.

2.4. Identification of Molecular Prognostic Factors. In this study, the random survival forest was used to screen prognostic factors that could supplement clinical prognosis, and then the multicovariate Cox model was used to identify prognostic factors that could be the best supplementary factors for clinical prognostic factors.

2.4.1. Identification with Random Survival Forest. The random survival forest is an ensemble tree-based method used to analyze right-censored survival data [27]. The nonparametric random survival forest model can assess the nonlinear effects of variables and explore the complex interactions between variables. In addition, variables in the random survival forest model that do not have prognostic ability can be filtered by variable importance. The variable selection procedure through the random survival forest in this study consists of the following three steps:

(A) Construct a random survival forest model with candidate variables. The numbers of trees that offer the lowest error rate were chosen

(B) In the constructed random survival forest model, variables with importance greater than 0 are selected and recorded

Table 1: Detailed information of the data used for analysis.

Characteristic	Discovery set TCGA-COAD	Validation set CPTAC	Reference set Normal samples
Patients, *n*	374	98	41
Survival status, *n* (%)			
Alive	293 (78.3)	90 (91.8)	29 (70.7)
Dead	81 (21.7)	8 (8.2)	12 (29.3)
Age[a] in years, mean (SD, range)	66.75 (12.73, 31-90)	65.43 (11.56, 35-93)	70.34 (13.23, 40-90)
Gender, *n* (%)			
Male	199 (53.2)	41 (41.8)	20 (48.8)
Female	175 (46.8)	57 (58.2)	21 (51.2)
Overall survival time in months, mean (median, range)	30.24 (24.27, 0.47-150.07)	27.96 (30, 1-44)	27.66 (24.37, 0-101.40)
T stage, *n* (%)			
T1	9 (2.4)	0 (0)	Not available
T2	65 (17.4)	12 (12.2)	
T3	258 (69.0)	73 (74.5)	
T4	42 (11.2)	13 (13.3)	
N stage, *n* (%)			
N0	226 (60.4)	52 (53.1)	Not available
N1	84 (22.5)	31 (31.6)	
N2	64 (17.1)	15 (15.3)	
M stage, n (%)			
M0	315 (84.2)	91 (92.9)	Not available
M1	59 (15.8)	7 (7.1)	
Number of genes, *n*	10877	10877	10877

[a] The characteristic "Age" refers to the age at initial diagnosis in the discovery set and reference set but refers to the age at procurement in the validation set. SD: standard deviation.

Table 2: Description of the pathways identified by the random survival forest.

KEGG pathway ID	Pathway name	Number of genes involved in the pathway in this study
hsa00450	Selenocompound metabolism—Homo sapiens (human)	13
hsa00532	Glycosaminoglycan biosynthesis—chondroitin sulfate/dermatan sulfate—Homo sapiens (human)	16
hsa02010	ABC transporters—Homo sapiens (human)	24
hsa04380	Osteoclast differentiation—Homo sapiens (human)	105
hsa04614	Renin-angiotensin system—Homo sapiens (human)	14
hsa04750	Inflammatory mediator regulation of TRP channels—Homo sapiens (human)	65
hsa04911	Insulin secretion—Homo sapiens (human)	41
hsa04971	Gastric acid secretion—Homo sapiens (human)	40
hsa04975	Fat digestion and absorption—Homo sapiens (human)	13
hsa05032	Morphine addiction—Homo sapiens (human)	38
hsa05133	Pertussis—Homo sapiens (human)	56
hsa05152	Tuberculosis—Homo sapiens (human)	128
hsa05167	Kaposi sarcoma-associated herpesvirus infection—Homo sapiens (human)	148
hsa05321	Inflammatory bowel disease (IBD)—Homo sapiens (human)	34

(C) Considering the existence of random processes, steps A and B would be repeated 100 times to generate a matrix of variables with a variable importance value greater than 0. The prognostic factors that were recorded as important prognostic factors multiple times were regarded as important prognostic factors

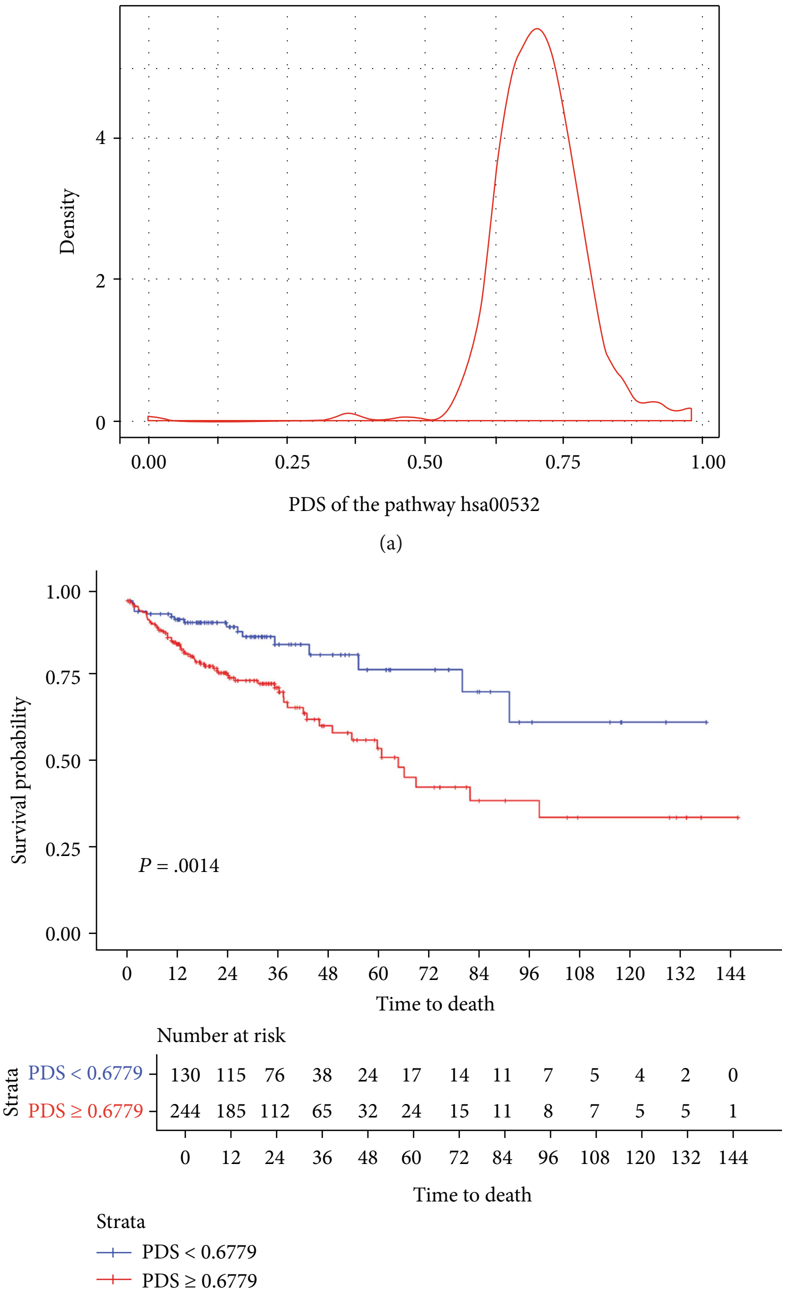
Table 3: Bias-corrected C-indexes of 27 different clinical-molecular integrated models.

Covariates used in the model	Bias-corrected Harrell's C-index (±95% CI)
T, N, M, hsa00532, hsa04911, hsa05133, hsa05152	0.775 ± 0.0038
T, N, M, hsa00532	0.773 ± 0.0038
T, N, M, hsa00532, hsa04911, hsa05133	0.773 ± 0.0038
T, N, M, hsa02010, hsa05152, hsa05321	0.773 ± 0.0038
T, N, M, hsa02010, hsa05167, hsa05321	0.772 ± 0.0037
T, N, M, hsa00532, hsa04380, hsa04911, hsa05133	0.772 ± 0.0039
T, N, M, hsa00532, hsa05133, hsa05152	0.772 ± 0.0040
T, N, M, hsa00532, hsa04380, hsa04971, hsa05133	0.771 ± 0.0039
T, N, M, hsa00532, hsa05133, hsa05167	0.771 ± 0.0040
T, N, M, hsa00532, hsa04975, hsa05133	0.770 ± 0.0041
T, N, M, hsa02010, hsa04911, hsa05133, hsa05152	0.768 ± 0.0040
T, N, M, hsa02010, hsa05133, hsa05152	0.768 ± 0.0040
T, N, M, hsa00532, hsa04971, hsa05133	0.767 ± 0.0038
T, N, M, hsa00532, hsa04380, hsa05133	0.766 ± 0.0041
T, N, M, hsa02010, hsa04911, hsa05133	0.764 ± 0.0038
T, N, M, hsa02010, hsa05133, hsa05167	0.764 ± 0.0040
T, N, M, hsa00532, hsa04750, hsa05133	0.764 ± 0.0041
T, N, M, hsa02010, hsa04911	0.763 ± 0.0037
T, N, M, hsa02010, hsa04380, hsa04911, hsa05133	0.763 ± 0.0040
T, N, M, hsa05133, hsa05152	0.761 ± 0.0042
T, N, M, hsa02010, hsa04750, hsa05133	0.759 ± 0.0040
T, N, M, hsa00450, hsa04911	0.758 ± 0.0036
T, N, M, hsa04380, hsa05133	0.758 ± 0.0043
T, N, M, hsa04911, hsa05133	0.756 ± 0.0039
T, N, M, hsa00450, hsa04911, hsa05133, hsa05152	0.756 ± 0.0040
T, N, M, hsa04380, hsa04911, hsa05133	0.754 ± 0.0041
T, N, M, hsa05133, hsa05167	0.754 ± 0.0042
T, N, M	0.746 ± 0.0040

CI: confidence interval.

In this study, identification of molecular prognostic factors through the random survival forest was implemented with the following procedures. First, we performed a rough screening on all molecular factors. The clinical prognostic factors and all molecular factors were used as variables in the random survival forest. Variables that showed positive prognostic power more than 90 times according to the variable selection procedure were identified as the potential important prognostic factors. Then, we tried to identify robust molecular prognostic factors that could supplement the clinical prognostic factors. The potential important prognostic factors identified by the rough screening were screened again. Here, the identified potential important molecular factors and the clinical prognostic factors were used as variables of the random survival forest. The variable selection procedure was repeated 10 times to ensure the robustness. In each repetition, variables that showed positive prognostic power over 95 times were recorded as important prognostic factors. Finally, molecular factors that were recorded as important prognostic factors in all 10 repetitions were regarded as the final important prognostic factors identified by the random survival forest.

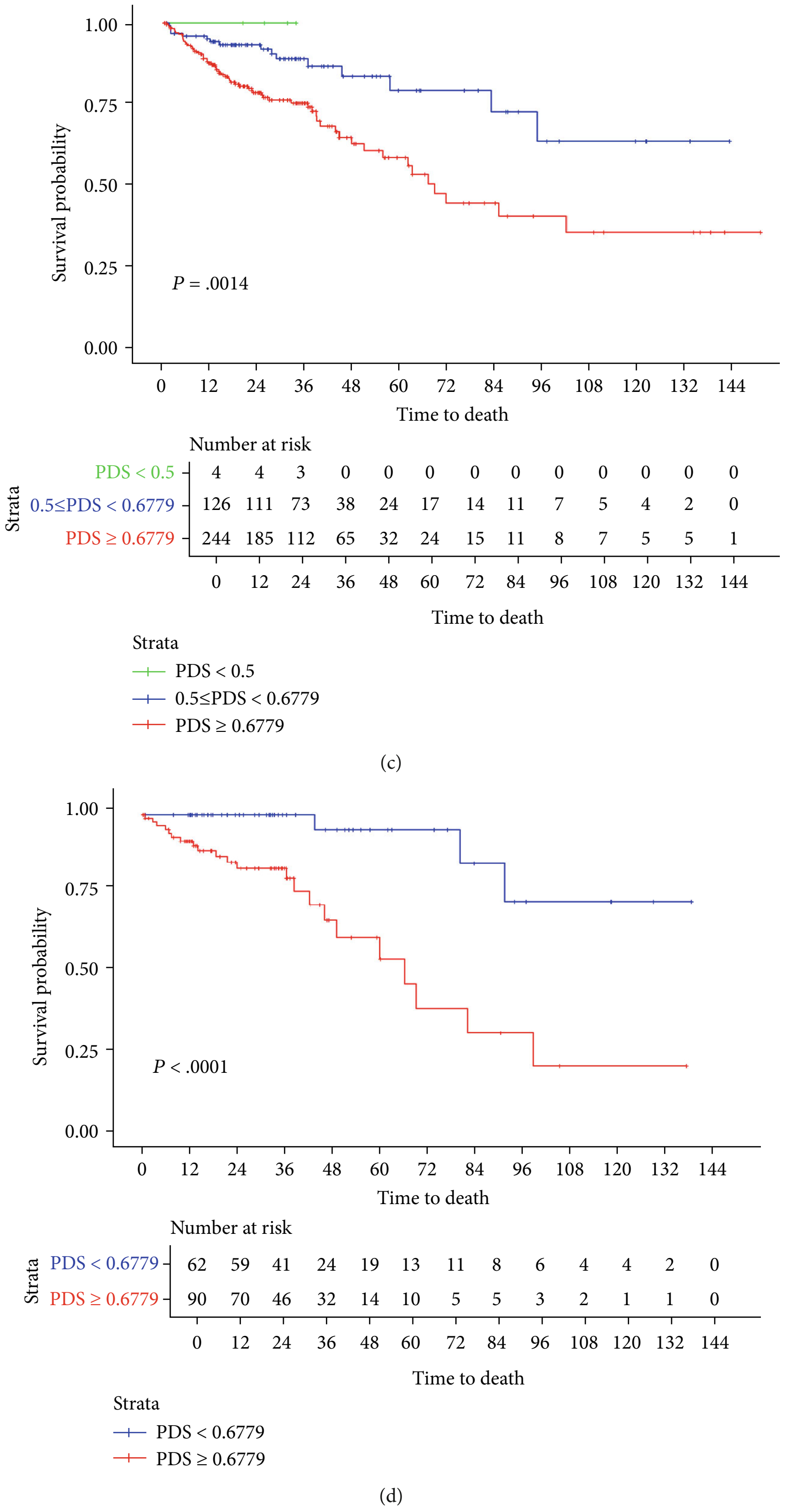
2.4.2. Identification with Multicovariate Cox Model. Multicovariate Cox models were constructed to identify the best molecular factors for clinical prognostic supplementation. These models were constructed with clinical prognostic factors and different combinations of molecular prognostic factors identified in Section 2.4.1. The models in which molecular factors showed no statistical significance of prognostic importance (with a P value of the covariate larger than 0.05) were excluded. The discrimination performance of the remaining models was measured by the bias-corrected concordance index (C-index). Molecular prognostic factors in the model with the best discrimination performance were regarded as the best molecular prognostic factors. If multiple



(a)

(b)

Figure 2: Continued.



(c)

(d)

Figure 2: Observation of the PDS of the pathway hsa00532 in the discovery set. (a) Density distribution of the PDS of the pathway hsa00532 in the discovery set. (b) KM curve plotted based on two groups of patients in the discovery set divided by the PDS with a threshold of 0.6779. (c) KM curve plotted based on three groups of patients in the discovery set divided by the PDS with thresholds of 0.5 and 0.6779. (d) KM curve plotted based on two groups of stage II patients in the discovery set divided by the PDS with a threshold of 0.6779.

models showed similar discrimination performance, the molecular factors that used the least number of genes were regarded as the best molecular prognostic factors.

2.5. Construction of the Clinical-Molecular Integrated Prognostic Model. The clinical prognostic factor T, N, and M stages and the identified best molecular prognostic factors were used to construct the clinical-molecular integrated prognostic model. Therefore, a multicovariate Cox model was built, with the formula as follows:

$$h(t) = h(0)\ \exp\left(\alpha_1(\text{T stage}) + \alpha_2(\text{N stage}) + \alpha_3(\text{M stage}) + \sum \beta_n M_n\right), \tag{1}$$

where $h(t)$ is the risk of death at time t, $h(0)$ is the baseline risk, α is the regression coefficient of clinical prognostic factors, β is the regression coefficient of molecular prognostic factors, and M is the identified molecular prognostic factor. In addition, identical clinical prognostic factors and molecular prognostic factors were used to construct the corresponding clinical prognostic model and molecular prognostic model. Comparisons of these models were performed to evaluate the improvement of prognostic performance between the clinical-molecular integrated prognostic model and the clinical prognostic model. Finally, a nomogram was constructed based on the clinical-molecular integrated prognostic model to predict the 3-year colon cancer overall survival.

2.6. Assessment of the Clinical-Molecular Integrated Prognostic Model. First, according to the distribution of the PDS of the corresponding molecular factors identified, patients in the discovery set were divided into different groups. The grouping was based on the highest degree of differentiation of survival curves. These findings could provide a direct observation of the relevance of the identified molecular prognostic factors and survival.

Second, based on the clinical prognostic factors and identified molecular prognostic factors, one clinical prognostic model, one molecular prognostic model, and one clinical-molecular integrated prognostic model were constructed on the discovery set. Internal validation through bootstrapping with 200 iterations was used to assess the discrimination performance of these models on the discovery set. External validation through stratified bootstrapping with 200 iterations was used to assess the discrimination performance of these models on the validation set. As the mean survival time of patients with metastasis was shorter than that of patients without metastasis, the performance of the prognostic model might have been affected. Therefore, models for nonmetastatic patients were built with the same prognostic factors and compared with the same assessment.

Finally, to compare the prognostic performance of directly using gene expression data and using converted PDS in this study, genes involved in the pathways were combined with clinical prognostic factors in the clinical-molecular integrated prognostic model. Comparisons between the gene-based integrated prognostic model and pathway-based integrated prognostic model were conducted.

Table 4: Regression coefficients of the knowledge-based clinical-molecular integrated prognostic model.

Covariate	Coefficient ± SE	HR	95% CI	*P* value
T stage				
T2	−1.62 ± 1.42	0.20	0.012-3.18	.25
T3	0.38 ± 1.02	1.47	0.20-10.79	.71
T4	1.22 ± 1.05	3.38	0.43-26.37	.25
N stage				
N1	−0.01 ± 0.31	0.99	0.54-1.80	.96
N2	0.67 ± 0.30	1.95	1.07-3.54	.03
M stage				
M1	1.02 ± 0.28	2.78	1.60-4.82	<.001
hsa00532*	2.86 ± 1.42	17.53	1.08-283.24	.04

SE: standard error; HR: hazard ratio; CI: confidence interval. *Covariate hsa00532 used in the model is the PDS of pathway has00532.

The constructed integrated prognostic model had a potential problem of overfitting as it contains multiple covariates. The bias-corrected Harrell's C-index which overcomes the problem of overfitting was chosen to evaluate the overall discriminative performance of the models in internal validation [28]. The origin Harrell's C-index was used in external validation of the overall discriminative performance. Uno's C-index, which is free of censoring, was chosen to evaluate the discriminative performance of the models at the 3-year time point [29]. A two-sided Wilcoxon signed-rank test was used to compare the 200 C-indexes generated from the 200 iterations of the bootstrapping procedure to quantify the discriminative difference of the C-index between different models.

2.7. Statistical Analysis. All statistical analyses were performed using R statistical software (version 3.5.3). Construction of Cox models and the nomogram and internal validation of Harrell's C-index and calibration plot were performed with the rms R package. External validation of Harrell's C-index was performed with the Hmisc R package. Uno's C-index was calculated with the survC1 R package. The Wilcoxon signed-rank test was performed with the stats R package. The random survival forest was performed with the randomForestSRC R package.

3. Results

3.1. Results of Data Processing. After data acquisition and processing, this study included 374 cases in the TCGA-COAD data as the discovery set, 98 colon cancer cases in the CPTAC as the validation set, and 41 colon cancer normal adjacent tissue data from the TCGA as the reference set. Both the discovery set and the validation set included the T, N, and M stages with identical categories and overall survival information including overall survival time and overall survival status. The detailed information of the final dataset used for analysis is shown in Table 1.

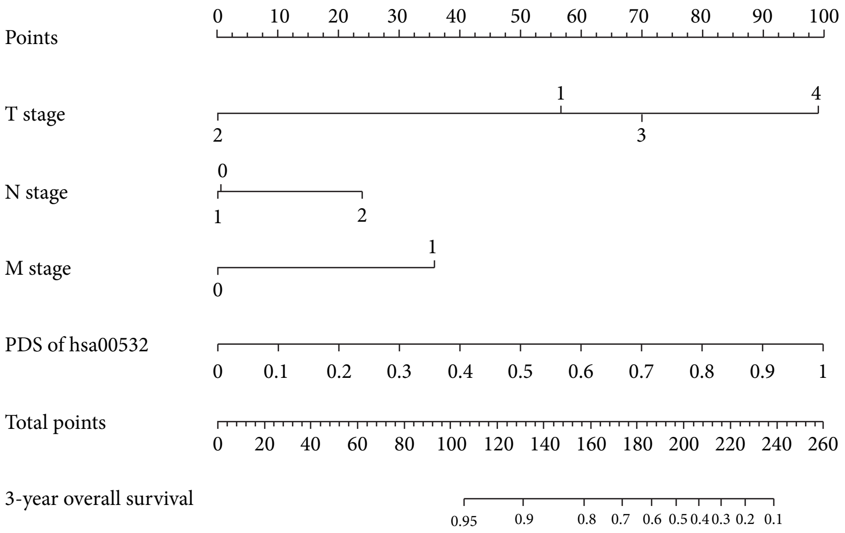


Figure 3: Nomogram for predicting the 3-year overall survival of colon cancer patients. To use the nomogram, first, the position of each variable of an individual patient on the corresponding axis should be found. Next, a line to the point axis for the number of points should be drawn upwards to determine the number of points of each variable. Then, the points from all the variables should be added. Finally, a line from the total point axis should be drawn downward to determine the likelihood of 3-year survival probabilities at the lower line of the nomogram.

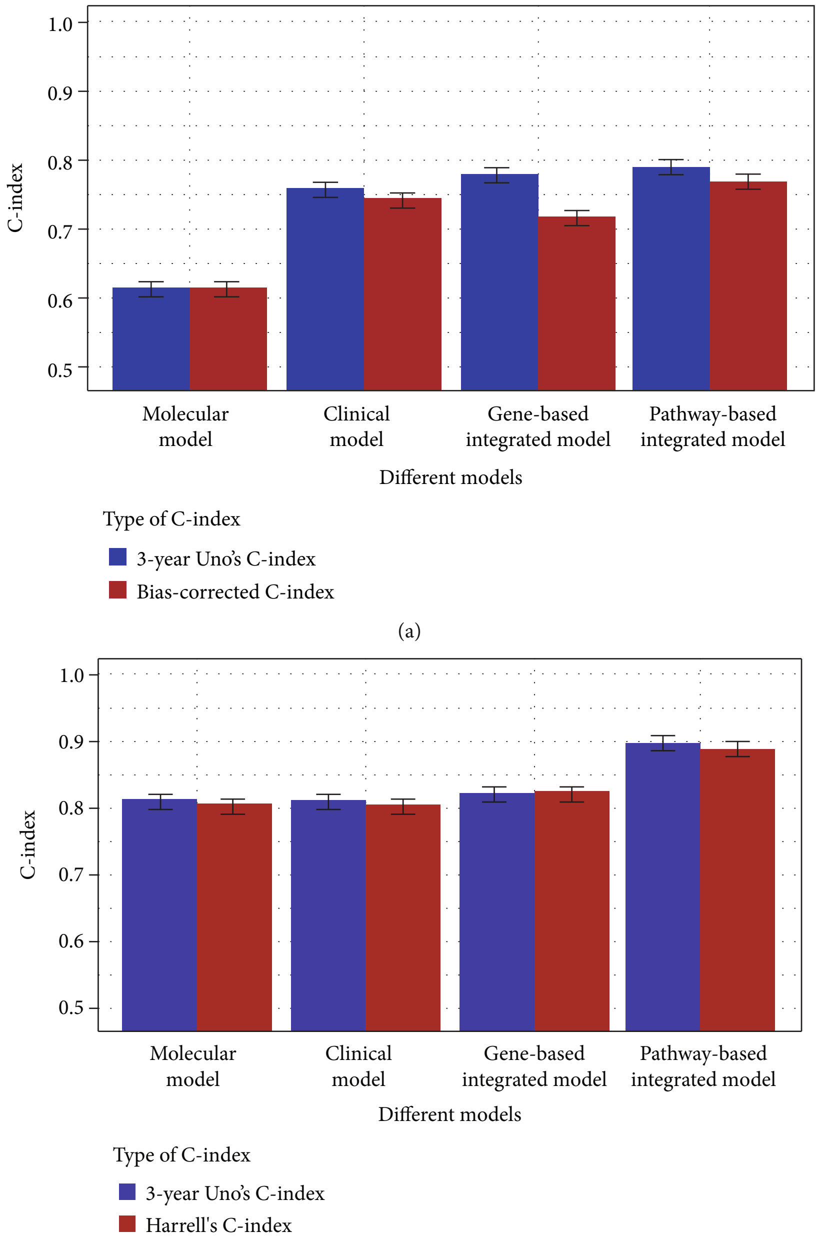
3.2. Identification of Molecular Prognostic Factors. Through the random survival forest, a total of 14 pathways were screened as potential molecular prognostic factors as shown in Table 2. After further screening through the multicovariate Cox model, 27 combinations of different pathways were found to have significant prognostic effects in the integrated models. Based on the bias-corrected C-indexes of these 27 different clinical-molecular integrated models shown in Table 3, and the numbers of genes used in the analysis of each pathway shown in Table 2, we concluded that the PDS of the pathway has00532 should be the best molecular prognostic factor to supplement clinical prognosis among these 14 pathways. In this study, 16 genes were included in the analysis: XYLT1, XYLT2, B4GALT7, B3GALT6, B3GAT3, CSGALNACT1, CSGALNACT2, CHSY1, CHPF, CHPF2, DSE, CHST11, CHST12, CHST3, CHST15, and CHST14. The other 4 genes, CHSY3, CHST13, CHST7, and UST, were removed during data processing as these four genes were not matched in the validation set. These 16 genes were used for gene-based model construction in subsequent analyses.

Observation of the distribution of the PDS of the pathway hsa00532 in the discovery set suggested that it approximately obeyed a normal distribution as shown in Figure 2(a). Therefore, patients in the discovery set were divided into a high-PDS group and a low-PDS group. Based on the difference in Kaplan-Meier (KM) curves between different patient groups, a threshold of 0.6779 was considered to most clearly separate these two groups, with the corresponding KM curves shown in Figure 2(b). In addition, several peaks at approximately less than 0.5 of the density distribution led us to separate the patients into three groups according to thresholds of 0.5 and 0.6779, with the corresponding KM curves shown in Figure 2(c). The high-PDS and low-PDS groups divided by 0.6779 showed significant survival differences in stage II colon cancer patients, as shown in Figure 2(d).

3.3. Constructed Knowledge-Based Clinical-Molecular Integrated Prognostic Model. With the identified knowledge-based prognostic factor, the PDS of the pathway hsa00532, and clinical prognostic factor T, N, and M stages, our knowledge-based clinical-molecular integrated prognostic model was built. To assess the improvement of our model compared with the clinical prognostic model, the corresponding clinical prognostic model based on T, N and M stages and the molecular prognostic model based on the PDS of pathway has00532 were constructed. The multicovariate Cox model was used to determine the regression coefficients of the models, with the coefficients of the knowledge-based clinical-molecular integrated prognostic model summarized in Table 4 and regression coefficients of the other models summarized in Table S1, Table S2, and Table S3. A corresponding nomogram that predicts the 3-year overall survival was constructed and is shown in Figure 3.

3.4. Assessment of the Prognostic Models for all Colon Cancer Patients. The discriminative performance of different models was measured with both Harrell's C-index for overall performance and 3-year Uno's C-index for performance at specific time points and is shown in Figure 4. In the internal validation, our model outperformed in terms of overall prognostic performance compared to the clinical prognostic model (0.773 vs 0.746, $P < .001$) and the molecular prognostic model (0.773 vs 0.619, $P < .001$) as shown in Figure 4(a). In the external validation, our model again outperformed in terms of overall prognostic performance compared to the clinical prognostic model (0.893 vs 0.808, $P < .001$) and the molecular prognostic model (0.893 vs 0.810, $P < .001$) as shown in Figure 4(b).

The prognostic performance of our model at the 3-year time point was assessed by Uno's C-index and calibration plot. The 3-year Uno's C-index in the discovery set suggested



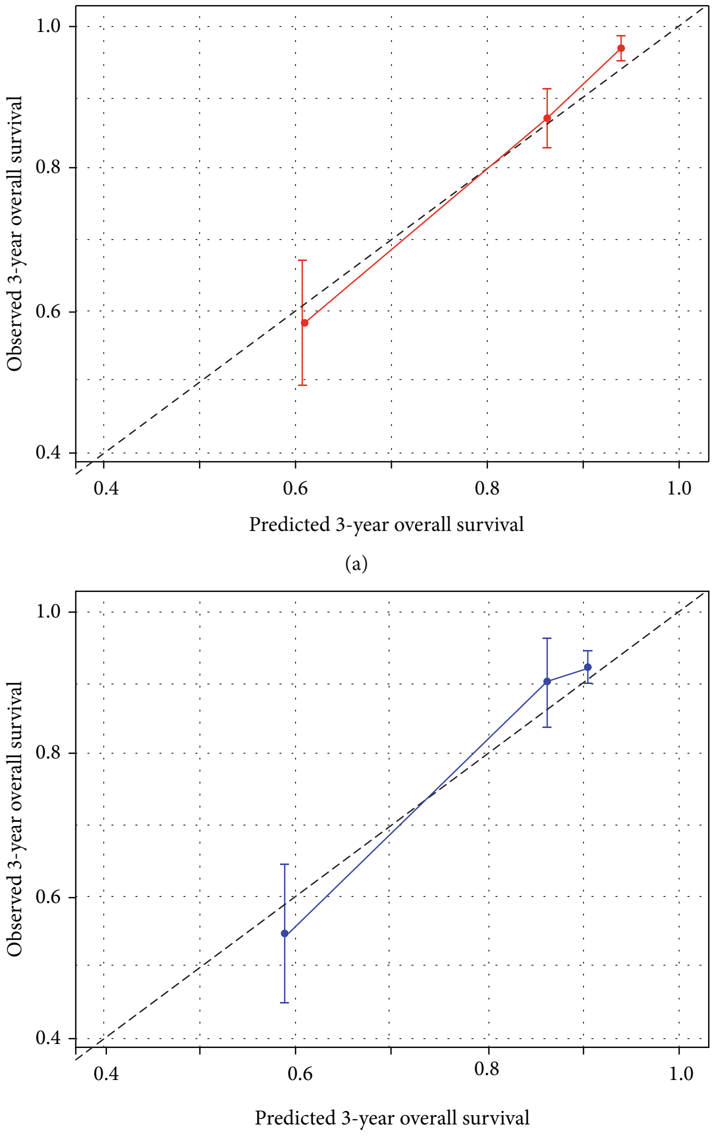
(a)

(b)

FIGURE 4: C-index of our pathway-based integrated model and other models for patients in the discovery set (a) and validation set (b).

that our model has the best discriminative performance compared to the clinical model (0.793 vs. 0.762, $P < .001$) and the molecular model (0.793 vs. 0.619, $P < .001$) as shown in Figure 4(a), whereas in the validation set, the comparison results were 0.899 vs. 0.816 ($P < .001$) compared to the clinical model and 0.899 vs. 0.816 ($P < .001$) compared to the molecular model as shown in Figure 4(b). The calibration plot of these models also showed that our model has a superior calibration performance compared with the clinical model at a 3-year time point as shown in Figure 5.

3.5. Assessment of the Prognostic Models for Nonmetastatic Colon Cancer Patients. Because the mean survival time of metastatic patients is shorter than that of nonmetastatic patients, the prognostic performance of the prognostic models may be affected. Therefore, the clinical prognostic model, molecular prognostic model, and clinical-molecular integrated prognostic model were constructed with the same prognostic factors used for nonmetastatic patients. The same assessments were performed on these models for nonmetastatic patients, with nonmetastatic patients in the discovery set and the validation set. In the internal validation, the integrated model outperformed in terms of overall prognostic performance compared to the clinical prognostic model (0.712 vs. 0.665, $P < .001$ for bias-corrected Harrell's C-index, 0.763 vs. 0.709, $P < .001$ for the 3-year Uno's C-index) and the molecular prognostic model (0.712 vs. 0.655, $P < .001$ for bias-corrected Harrell's C-index, 0.763 vs. 0.659, $P < .001$ for the 3-



(a)

(b)

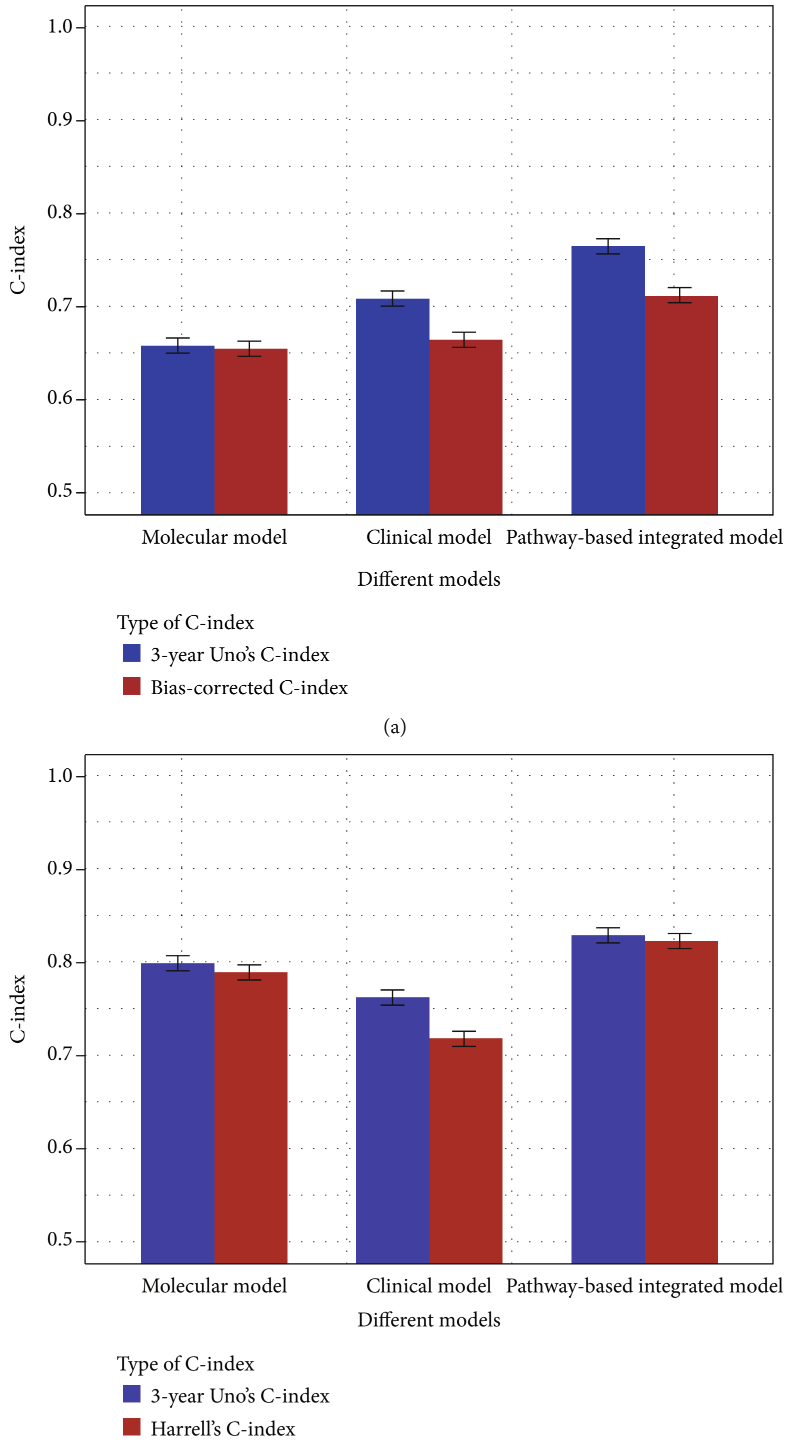
Figure 5: Calibration plot of our pathway-based integrated model (a) and clinical model (b) at the 3-year time point.

year Uno's C-index) as shown in Figure 6(a). In the external validation, the integrated model again outperformed in terms of overall prognostic performance compared to the clinical prognostic model (0.824 vs. 0.720, $P < .001$ for Harrell's C-index, 0.829 vs. 0.743, $P < .001$ for the 3-year Uno's C-index) and the molecular prognostic model (0.824 vs. 0.791, $P < .001$ for Harrell's C-index, 0.829 vs. 0.799, $P < .001$ for the 3-year Uno's C-index) as shown in Figure 6(b).

3.6. Pathway-Based Model Is Superior to the Gene-Based Model. Previous studies have claimed that the introduction of representative functional units should improve gene expression-based studies [10, 14–16, 30]. Therefore, genes involved in the pathway hsa00532 were used to construct a gene-based clinical-molecular integrated prognostic model. The regression coefficients of the gene-based model suggested that only two genes (CSGALNACT1 and DSE) were prognostically related when combined with clinical factors, and they are summarized in Table S3. Compared with our knowledge-based integrated model, the C-indexes of the gene-based integrated model in the discovery set were lower, with 0.721 vs. 0.773 ($P < .001$) for the bias-corrected C-index, 0.783 vs. 0.793 ($P < .001$) for the 3-year Uno's C-index in the discovery set, 0.825 vs. 0.893 ($P < .001$) for Harrell's C-index, and 0.826 vs. 0.899 ($P < .001$) for the 3-year Uno's C-index in the validation set as shown in Figure 4. These results suggest that the pathway-based integrated model is superior to the gene-based integrated model in discriminative performance because the gene-based integrated model might include too many redundant prognostic factors.

4. Discussion

4.1. Principal Results. Through knowledge-based clinical-molecular integrated analysis by the random survival forest



(a)

(b)

Figure 6: C-index of our pathway-based integrated model and other models for nonmetastatic patients in the discovery set (a) and the validation set (b).

and multicovariate Cox model, this study successfully identified the PDS of the pathway hsa00532 as the best molecular prognostic factor for supplementing the prognostic performance of the T, N, and M stages in overall survival prediction. The results of internal validation and external validation suggested that the knowledge-based clinical-molecular integrated prognostic model had the best discriminative performance and improved calibration performance than the clinical prognostic model. The regression coefficients of the covariate in different models in Table 4, Table S1, and Table S2 indicated that tiny changes were observed for the clinical prognostic factors, while the molecular prognostic factor keeps as an independent prognostic factor in the final clinical-molecular integrated prognostic model. These further indicated that our final clinical-molecular integrated prognostic model does satisfy the aim of our study. In addition, the pathway-based models were superior to the gene-based model, which indicates that the incorporation of pathway information can make more use of the expression information of genes involved in a pathway rather than directly using the expression information of genes.

The observation of the KM curves based on patient groups divided by thresholds of 0.5 and 0.6779 suggested that patients with higher PDSs had worse survival. In addition, the KM curves of stage II patients divided into high-PDS and low-PDS groups could be distinctly distinguished, indicating that the PDS of the pathway hsa00532 might be a potential biomarker for separating high-risk stage II colon cancer patients.

The pathway hsa00532 is named glycosaminoglycan biosynthesis-chondroitin sulfate/dermatan sulfate on the KEGG website, and it is related to the biosynthesis of chondroitin sulfate and dermatan sulfate. Previous studies have indicated that the dermatan sulfate chain is different between colon cancer and normal colonic mucosa, and chondroitin sulfate is associated with tumor metastasis [31–34]. However, the PDS of the pathway hsa00532 showed no relevance to the metastasis status, with a Pearson correlation coefficient of 0.04 for the discovery dataset. In addition, the pathway hsa00532 showed considerable supplementary power in the models for both all-stage and nonmetastatic colon cancer patients, while the metastasis status and the PDS of the pathway hsa00532 were regarded as independent significant prognostic factors in the constructed model. One possible explanation for this finding is that although the PDS in this study is generated from gene expression data, a series of regulatory and expression biology processes from the transcriptome is still needed to generate the actual pathway products. Two genes, CSGALNACT1 and DSE, involved in the pathway hsa00532 might be potential markers for colon cancer prognosis because only these two genes showed a potential prognostic effect in the gene-based integrated model based on the regression coefficients of the gene-based model summarized in Table S3. Further validation is required to validate the prognostic effect of these two genes as currently published papers have not mentioned them in conjunction with colon cancer prognosis.

Recent studies on colon cancer prognosis mainly focused on finding better molecular prognostic features, while our study was aimed at supplementing the current clinical staging system with molecular features [12, 13, 35]. Compared to a recent colon cancer prognosis study which incorporated both clinical prognostic features and gene expression profiles, our study integrated the clinical prognostic features and gene expression profile in a conditional way rather than joining the two types of features independently [12]. The conditional modelling strategy is more suitable for our study as this study was aimed at supplementing the prognosis performance of the current TNM staging system.

4.2. Limitations. There are still limitations in this study. The clinical prognostic factors in this study involved only the T, N, and M stages, while in the actual clinical treatment of colon cancer, there are many other factors that need to be considered, such as the patient's physical condition and the chemotherapy or radiotherapy regimen. In addition, due to the short follow-up time of the validation set, it was not possible to further validate the performance of our model on long-term prognosis. The conditional modelling strategy for clinical-molecular integration could satisfy the demand of the current study, although it could not fully utilize the correlation structure between clinical and molecular factors [36]. Further studies about how to make better use of molecular features should be considered. The current study used only gene expression data from the transcriptome, and the addition of other types of omics data, such as genome or epigenome data, may further improve the accuracy of molecular features and better supplement the clinical prognosis. However, with the current technology, how to balance the improvement in discriminative performance and the cost of sequencing remains to be considered.

Through cooperation with local hospitals, we can collect more real-world follow-up patients and sequence their tumor samples to generate more molecular data. Therefore, further validation of our model could be conducted. The involvement of more clinical prognostic factors in clinical-molecular analysis could make a more detailed and specific supplement to clinical prognosis. New integrative models based on a conditional strategy or even a joint modelling strategy would be required to deal with new data. In addition, considering that the PDS of the pathway hsa00532 can effectively distinguish the risk of stage II patients, further research and validation should be performed with more data. After further validation with real data, further research related to the PDS of the pathway has00532, such as immunohistochemistry and other methods appropriate for clinical use, should be conducted, with corresponding web-based tools being developed.

5. Conclusions

In conclusion, this study identified that the PDS of the pathway hsa00532 can be used as a supplementary prognostic factor for the three clinical prognostic factor T, N, and M stages. The clinical-molecular integrated prognostic model constructed with these three clinical prognostic factors and the identified molecular prognostic factor is superior to the clinical prognostic model, molecular prognostic model, or gene-based integrated prognostic model in prognostic performance. A corresponding nomogram including the three clinical prognostic factors and the identified molecular prognostic factor was constructed for possible clinical use. In addition, the PDS of the pathway hsa00532 showed a significant ability to distinguish high risk stage II colon cancer patients and is a potential prognostic marker. The PDS calculation of the pathway hsa00532 involves only 16 genes; therefore, it has good prospects for clinical use after further validation with real data.

Data Availability

The dataset analyzed during the current study is available in the cBioPortal, http://www.cbioportal.org; generated by the National Cancer Institute CPTAC and available in the LinkedOmics, http://linkedomics.org/cptac-colon/ and available in the UCSC Xena, http://xena.ucsc.edu/.

Disclosure

Thanks are due to the Research Square for accepting this article as a preprint; the preprint can be found in https://www.researchsquare.com/article/rs-36892/v1, and the preprint DOI is 10.21203/rs.3.rs-36892/v1.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Authors' Contributions

DT, YT, and JL contributed to the conception of the study. DT, YT, and QY performed the data preparation. DT and YT performed the data analyses and wrote the manuscript. JSL and KD provided critical revisions. All authors read and approved the final manuscript. Danyang Tong and Yu Tian contributed equally to this work.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 81771936 and No. 81801796), the Major Scientific Project of Zhejiang Lab (No. 2020ND8AD01), and the National Key Research and Development Program of China (No. 2018YFC0116901). In addition, we acknowledge the TCGA Research Network and all staff members involved in the TCGA Project, the CPTAC Project, the cBioPortal Project, and the UCSC Xena Project for providing the data.

Supplementary Materials

Supplementary 1. Table S1: regression coefficients of the clinical prognostic model.

Supplementary 2. Table S2: regression coefficients of the molecular prognostic model.

Supplementary 3. Table S3: regression coefficients of the gene-based clinical-molecular integrated prognostic model.

References

[1] W. Chen, R. Zheng, P. D. Baade et al., "Cancer statistics in China, 2015," *CA: A Cancer Journal for Clinicians.*, vol. 66, no. 2, pp. 115–132, 2016.

[2] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2019," *CA: A Cancer Journal for Clinicians*, vol. 69, no. 1, pp. 7–34, 2018.

[3] C. Allemani, T. Matsuda, V. di Carlo et al., "Global surveillance of trends in cancer survival 2000-14 (CONCORD-3): analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries," *The Lancet*, vol. 391, no. 10125, pp. 1023–1075, 2018.

[4] J. Li, B. C. Guo, L. R. Sun et al., "TNM staging of colorectal cancer should be reconsidered by T stage weighting," *World Journal of Gastroenterology*, vol. 20, no. 17, pp. 5104–5112, 2014.

[5] M. B. Amin, F. L. Greene, S. B. Edge et al., "The eighth edition AJCC cancer staging manual: continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging," *CA: A Cancer Journal for Clinicians*, vol. 67, no. 2, pp. 93–99, 2017.

[6] A. E. Giuliano, S. B. Edge, and G. N. Hortobagyi, "Eighth edition of the AJCC cancer staging manual: breast cancer," *Annals of Surgical Oncology*, vol. 25, no. 7, pp. 1783–1785, 2018.

[7] J. Li, C. H. Yi, Y. T. Hu et al., "TNM staging of colorectal cancer should be reconsidered according to weighting of the T stage verification based on a 25-year follow-up," *Medicine*, vol. 95, no. 6, p. e2711, 2016.

[8] X. X. Kong, J. Li, Y. B. Cai et al., "A modified TNM staging system for non-metastatic colorectal cancer based on nomogram analysis of SEER database," *Bmc Cancer*, vol. 18, no. 1, p. 50, 2018.

[9] M. R. Weiser, "AJCC 8th edition: colorectal cancer," *Annals of Surgical Oncology*, vol. 25, no. 6, pp. 1454-1455, 2018.

[10] A. Subramanian, P. Tamayo, V. K. Mootha et al., "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 43, pp. 15545–15550, 2005.

[11] L. Ein-Dor, I. Kela, G. Getz, D. Givol, and E. Domany, "Outcome signature genes in breast cancer: is there a unique set?," *Bioinformatics*, vol. 21, no. 2, pp. 171–178, 2005.

[12] J. H. Lee, S. Jung, W. S. Park et al., "Prognostic nomogram of hypoxia-related genes predicting overall survival of colorectal cancer-analysis of TCGA database," *Scientific Reports*, vol. 9, no. 1, p. 1803, 2019.

[13] R. Zhou, D. Zeng, J. Zhang et al., "A robust panel based on tumour microenvironment genes for prognostic prediction and tailoring therapies in stage I-III colon cancer," *eBioMedicine*, vol. 42, pp. 420–430, 2019.

[14] A. H. Bild, G. Yao, J. T. Chang et al., "Oncogenic pathway signatures in human cancers as a guide to targeted therapies," *Nature*, vol. 439, no. 7074, pp. 353–357, 2006.

[15] E. B. van den Akker, W. M. Passtoors, R. Jansen et al., "Meta-analysis on blood transcriptomic studies identifies consistently coexpressed protein-protein interaction modules as robust markers of human aging," *Aging Cell*, vol. 13, no. 2, pp. 216–225, 2014.

[16] S. G. Ma, M. R. Kosorok, J. A. Huang, and Y. Dai, "Incorporating higher-order representative features improves prediction in network-based cancer prognosis analysis," *BMC Medical Genomics*, vol. 4, no. 1, p. ???, 2011.

[17] B. T. Fa, C. W. Luo, Z. Tang, Y. Yan, Y. Zhang, and Z. Yu, "Pathway-based biomarker identification with crosstalk analysis for robust prognosis prediction in hepatocellular carcinoma," *eBioMedicine*, vol. 44, pp. 250–260, 2019.

[18] D. Kim, R. Li, A. Lucas, S. S. Verma, S. M. Dudek, and M. D. Ritchie, "Using knowledge-driven genomic interactions for multi-omics data analysis: metadimensional models for

predicting clinical outcomes in ovarian carcinoma," *Journal of the American Medical Informatics Association*, vol. 24, no. 3, pp. 577–587, 2017.

[19] S. J. Huang, C. Yee, T. Ching, H. Yu, and L. X. Garmire, "A novel model to combine clinical and pathway-based transcriptomic information for the prognosis prediction of breast cancer," *Plos Computational Biology*, vol. 10, no. 9, article e1003851, 2014.

[20] J. M. Bae, J. H. Kim, Y. Kwak et al., "Distinct clinical outcomes of two CIMP-positive colorectal cancer subtypes based on a revised CIMP classification system," *British Journal of Cancer*, vol. 116, no. 8, pp. 1012–1020, 2017.

[21] K. Chaudhary, O. B. Poirion, L. Q. Lu, and L. X. Garmire, "Deep learning-based multi-omics integration robustly predicts survival in liver cancer," *Clinical Cancer Research*, vol. 24, no. 6, pp. 1248–1259, 2018.

[22] K. A. Hoadley, C. Yau, T. Hinoue et al., "Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer," *Cell*, vol. 173, no. 2, pp. 291–304.e6, 2018.

[23] S. Vasaikar, C. Huang, X. Wang et al., "Proteogenomic analysis of human colon cancer reveals new therapeutic opportunities," *Cell*, vol. 177, no. 4, pp. 1035–1049.e19, 2019.

[24] Y. Drier, M. Sheffer, and E. Domany, "Pathway-based personalized analysis of cancer," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, no. 16, pp. 6388–6393, 2013.

[25] C. J. Vaske, S. C. Benz, J. Z. Sanborn et al., "Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM," *Bioinformatics*, vol. 26, no. 12, pp. i237–i245, 2010.

[26] C. Liu, S. Srihari, S. Lal et al., "Personalised pathway analysis reveals association between DNA repair pathway dysregulation and chromosomal instability in sporadic breast cancer," *Molecular Oncology*, vol. 10, no. 1, pp. 179–193, 2016.

[27] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, "Random survival forests," *Annals of Applied Statistics*, vol. 2, no. 3, pp. 841–860, 2008.

[28] F. E. Harrell, R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati, "Evaluating the yield of medical tests," *Journal of the American Medical Association*, vol. 247, no. 18, pp. 2543–2546, 1982.

[29] H. Uno, T. Cai, M. J. Pencina, R. B. D'Agostino, and L. J. Wei, "On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data," *Statistics in Medicine*, vol. 30, no. 10, pp. 1105–1117, 2011.

[30] V. K. Mootha, C. M. Lindgren, K. F. Eriksson et al., "PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes," *Nature Genetics*, vol. 34, no. 3, pp. 267–273, 2003.

[31] J. Iida, A. M. Meijne, J. R. Knutson, L. T. Furcht, and J. B. McCarthy, "Cell surface chondroitin sulfate proteoglycans in tumor cell adhesion, motility and invasion," *Seminars in Cancer Biology*, vol. 7, no. 3, pp. 155–162, 1996.

[32] C. M. Lee, T. Tanaka, T. Murai et al., "Novel chondroitin sulfate-binding cationic liposomes loaded with cisplatin efficiently suppress the local growth and liver metastasis of tumor cells in vivo," *Cancer Research*, vol. 62, no. 15, pp. 4282–4288, 2002.

[33] M. M. Fuster and J. D. Esko, "The sweet and sour of cancer: glycans as novel therapeutic targets," *Nature Reviews Cancer*, vol. 5, no. 7, pp. 526–542, 2005.

[34] K. Daidouji, K. Takagaki, S. Yoshihara, H. Matsuya, M. Sasaki, and M. Endo, "Neoplastic changes in saccharide sequence of dermatan sulfate chains derived from human colon cancer," *Digestive Diseases and Sciences*, vol. 47, no. 2, pp. 331–337, 2002.

[35] H. Jiang, J. du, J. M. Gu, L. Jin, Y. Pu, and B. Fei, "A 65-gene signature for prognostic prediction in colon adenocarcinoma," *International Journal of Molecular Medicine*, vol. 41, no. 4, pp. 2021–2027, 2018.

[36] E. López de Maturana, L. Alonso, P. Alarcón et al., "Challenges in the integration of omics and non-omics data," *Genes*, vol. 10, no. 3, p. 238, 2019.

Hindawi
BioMed Research International
Volume 2021, Article ID 5561569, 13 pages
https://doi.org/10.1155/2021/5561569

Research Article

LC-MS/MS-Based Quantitative Proteomics Analysis of Different Stages of Non-Small-Cell Lung Cancer

Murong Zhou,[1] **Yi Kong**,[2] **Xiaobin Wang**,[3] **Wen Li**,[3] **Si Chen**,[3,4] **Li Wang**,[5] **Chengbin Wang**,[2] **and Qian Zhang**[6]

[1]*College of Physics and Optoelectronic Engineering, Shenzhen University, Shenzhen, China*
[2]*Medical School of Chinese PLA & Medical Laboratory Center, First Medical Center of Chinese PLA General Hospital, Beijing, China*
[3]*Shenzhen University Health Science Center, Shenzhen, China*
[4]*Department of Immunology, Shenzhen University Health Science Center, Shenzhen, China*
[5]*Department of Dermatology, Shenzhen University General Hospital, Shenzhen University, Shenzhen, China*
[6]*Department of Dermatology, Huazhong University of Science and Technology Union Shenzhen Hospital, Shenzhen, China*

Correspondence should be addressed to Li Wang; pfkwangli@126.com, Chengbin Wang; wangcbin301@163.com, and Qian Zhang; zhangqianbisheng@163.com

Received 20 January 2021; Revised 25 January 2021; Accepted 7 February 2021; Published 27 February 2021

Academic Editor: Quan Zou

Lung cancer has a higher incidence rate and mortality rate than all other cancers. Early diagnosis and treatment of lung cancer remain a major challenge, and the 5-year survival rate of its patients is only 15%. Basic and clinical research, especially the discovery of biomarkers, is crucial for improving the diagnosis and treatment of lung cancer patients. To identify novel biomarkers for lung cancer, we used the iTRAQ8-plex labeling technology combined with liquid chromatography-tandem mass spectrometry (LC-MS/MS) to analyze the serum and urine of patients with different stages of lung adenocarcinoma and healthy individuals. A total of 441 proteins were identified in the serum, and 1,161 proteins were identified in the urine. The levels of elongation factor 1-alpha 2, proteasome subunit alpha type, and spermatogenesis-associated protein increased significantly in the serum of patients with lung cancer compared with those in healthy controls. The levels of transmembrane protein 143, cadherin 5, fibronectin 1, and collectin-11 decreased significantly in the serum of patients with metastases compared with those of nonmetastatic lung cancer patients. In the urine of stage III and IV lung cancer patients, the prostate-specific antigen and prostatic acid phosphatase decreased significantly, whereas neutrophil defensin 1 increased significantly. The results of LC-MS/MS were confirmed by enzyme-linked immunosorbent assay (ELISA) for transmembrane protein 143, cadherin 5, fibronectin 1, and collectin-11 in the serum. These proteins may be a potential early diagnosis and metastasis biomarkers for lung adenocarcinoma. Furthermore, the relative content of these markers in the serum and urine could be used to determine the progression of lung adenocarcinoma and achieve accurate staging and diagnosis.

1. Introduction

The incidence and mortality rates of lung cancer are higher than those of other cancers [1]. The overall 5-year survival rate for lung cancer is only 15% [2]. Lung cancer is divided into small cell carcinoma (SCLC) and non-small-cell carcinoma (NSCLC). SCLC accounts for 10%-15% of lung cancers and is sensitive to radiotherapy and chemotherapy [3, 4]. However, approximately 85% of lung cancers are NSCLC [5]. The median survival time of patients with advanced lung cancer is only 10 months [6]. Although the diagnosis and treatment of lung cancer have significantly improved, the current treatment methods are still not satisfactory. Improved early diagnosis and targeted treatment of lung cancer are required in clinical practices to improve patient outcomes.

Proteomics is the science of studying protein composition and alterations in cells, tissues, and organisms. Proteomics is widely used in basic and clinical medical research [7] for the identification of biomarkers [8, 9], posttranslational protein modifications [10], and the regulation of signaling pathways [11]. Proteomics research of lung cancer has focused on the classification of lung cancer [12], the correlation between protein and gene expressions [13, 14], identification of new molecular targets [15, 16], and the development of new drugs [17, 18]. Early diagnostic markers for lung cancer [19, 20] have been identified, but the markers still lack accuracy and sensitivity. Therefore, exploration and discovery of reliable and sensitive markers for the early diagnosis of lung cancer are a research priority.

Quantitative proteomics can determine relative changes in protein content. Based on this technology, differences in protein abundance between healthy individuals and patients with cancer can be defined to identify disease markers. Proteomics has been widely used in human disease research [21]. The postlabeling analysis of proteins using the iTRAQ8-plex technology, combined with Data Dependent Acquisition (DADA), is currently the standard labeling method used in quantitative proteomics [18, 22, 23]. The iTRAQ8-plex technology is also a commonly used quantitative method, which can be applied to quantitative analysis in proteomics research [24, 25].

For the first time, we combined iTRAQ8-plex labeling with liquid chromatography-tandem mass spectrometry (LC-MS/MS) as a quantitative proteomics analysis approach to compare protein abundance in the serum and urine samples of healthy controls to that in the serum and urine samples of stage I, II, III, and IV lung adenocarcinoma patients. The purpose of this study was to identify a set of biomarkers for the early diagnosis and metastasis prediction in patients with lung adenocarcinoma using serum and urine.

2. Materials and Methods

2.1. Chemicals and Reagents. Pierce™ Top12 Abundant Protein Depletion Spin Columns, dithiothreitol (DTT), indole-3-acetic acid (IAA), a bicinchoninic acid (BCA) kit, iTRAQ reagents, and Ziptip solid-phase microextraction reagents were purchased from Thermo Fisher Scientific, Inc. (Waltham, MA, USA). Trypsin was obtained from Promega. All organic reagents were high-performance liquid chromatography (HPLC) grade, and Milli-Q ultrapure water was used in all experiments (Millipore, Bradford, USA). Other reagents were analytical grade reagents, unless otherwise indicated.

2.2. Collection and Storage of Serum and Urine Samples. Serum and urine samples were collected from patients that were diagnosed with lung adenocarcinoma at the General Hospital of the Chinese People's Liberation Army from June 2017 to June 2018. The discovery set consisted of 30 healthy individuals and 70 lung adenocarcinoma patients at early stages (stage Ia1, $n = 10$; stage Ia2, $n = 10$; stage Ia3, $n = 10$; stage Ib, $n = 10$; and stage II, $n = 10$) and late stages (stage III, $n = 10$; stage IV, $n = 10$). Healthy controls were age- and gender-matched to the lung adenocarcinoma patients (Table 1). Of note, serum and urine samples were collected before the 70 lung adenocarcinoma patients had undergone chemical or medical treatment. All samples were collected from patients with an empty stomach in the morning. All participants provided signed informed consent, and samples were collected following the protocol approval. All methods were carried out in accordance with the approved guidelines, and all experimental protocols were approved by the ethics committee of the Chinese PLA General Hospital (Number S2018-007-001). Each serum sample was allowed to clot for 45 min and then centrifuged at 2,000 rpm for 10 min. For urine samples, a morning midstream urine specimen was collected and centrifuged at 1,500 rpm for 5 min. All samples were aliquoted and stored at −80°C until use.

2.3. Sample Preparation and Enzyme Digestion. For the LC-MS analysis, frozen serum and urine samples were thawed on ice. Serum samples with each group (100 μL of serum) were merged; the urine samples within each group (2 mL of urine) were merged. Pierce™ Top12 Abundant Protein Depletion Spin Columns were used to remove high-abundance proteins from the serum samples of each group. The filtrate was collected and denatured with 8 M urea. DTT and IAA were added to reduce alkylation, and 1 μg of Trypsin was then added at a ratio of 1 : 30 (enzyme : protein). The sample was hydrolyzed overnight at 37°C. Urine samples were precipitated with precooled acetone, and urea was used to redissolve the precipitate, followed by protein quantification using a BCA kit. Lastly, 3.3 μg Trypsin was added to 100 μg protein from each group, and the samples were enzymatically hydrolyzed (1 : 30, enzyme : protein).

2.4. Use of the iTRAQ-8plex for Labeling and Separation. After digestion, iTRAQ113-119 was used to separately label the serum and urine of stage Ia1-IV, and iTRAQ121 was used to label the serum and urine of the normal control group. For iTRAQ-8plex labeling, 150 μL isopropanol was added to each labeling reagent, and the mixture was added to each polypeptide sample (100 μg) after shaking and mixing. The reaction was carried out at room temperature for 2 hours. Then, water (100 μL) was added to terminate the reaction, and samples were freeze-dried after mixing.

The labeled peptide fragments were mixed, and 15 components were separated using the Agilent 1200 HPLC separation system. The chromatographic column was a high-pH RP C18 (4.6 mm × 250 mm, 5 μm, 300 A). Mobile phase A was the aqueous phase containing 20 mM ammonium acetate (pH 10). Mobile phase B was ACN/water containing 20 mM ammonium acetate (ACN/Water, 9/1, *v*/*v*, pH 10). The mobile phase gradient was 5% B to 35% B and 35%–40% B to 90% B; 90% B was then maintained for 10 minutes. The fractionated samples were desalinated using Ziptip solid-phase microextraction and analyzed.

2.5. SDS-PAGE. Serum samples were loaded onto a 10% sodium dodecyl sulfate- (SDS-) polyacrylamide gel electrophoresis (SDS-PAGE gel) (Invitrogen™, Thermo Fisher Scientific, Inc., New York, USA) and run at 100 V for 100 min in a running buffer. A prestained protein standard (Solarbio Science & Technology Co., Ltd., Beijing, China) was used to track protein migration. The resulting gels were stained with a fast silver stain kit (Beyotime, Beijing, China). The protocol was previously described [26].

2.6. LC-MS/MS Analysis. The prepared samples were separated and identified by liquid chromatography- (Ultimate3000,

Table 1: Clinical profiles and demographics of healthy controls and lung adenocarcinoma patients with different stages.

Demographics	Control	Stage Ia1	Stage Ia2	Stage Ia3	Stage Ib	Stage II	Stage III	Stage IV
n	30	10	10	10	10	10	10	10
Age	55 ± 8.32	53 ± 5.91	55.9 ± 6.79	54.3 ± 4.06	57.4 ± 8.37	57.9 ± 6.35	57.2 ± 5.39	57.8 ± 9.96
Male	21	7	7	7	7	7	7	7
Female	9	3	3	3	3	3	3	3
Smoking history	12	4	4	4	4	4	4	4
Ethnicity	Han (30)	Han (10)	Han (10)	Han (10)	Han (10)	Han (10)	Han (10)	Han (10)

121: healthy control; 113: stage Ia 1; 114: stage Ia2; 115: stage Ia3; 116: stage Ib; 117: stage II; 118: stage III; 119: stage IV.

Thermo Fisher Scientific, Inc.) tandem mass spectrometry (Q-Exactive, Thermo Fisher Scientific, Inc.). Samples were first loaded separately on a trap column (150 μm × 20 mm) packed with SP-300-ODS-AP (3 μm particle diameter; 100 nm pore size in house). Each sample was eluted into an analytical column (75 μm × 15 mm), packed with SP-300-ODS-AP, and separated at a flow rate of 500 nL•min^{-1} with an elution gradient consisting of mobile phase B (80% acetonitrile, 20% H_2O, 0.1% formic acid) and mobile phase A (99.9% H_2O, 0.1% formic acid). Elution gradient solutions were added as follows: B was increased from 4% to 10% in 5 minutes, from 10% to 12.5% in 10 minutes, from 12.5% to 27.5% in 75 minutes, from 27.5% to 50% in 110 minutes, and from 50% to 95% in 10 minutes. The LC system was automatically equilibrated with mobile phase A for approximately 10 minutes before the next analysis. Fractions were continuously detected in the Q-Exactive hybrid quadrupole-orbitrap mass spectrometer with a nanoelectrospray ionization source at a capillary temperature of 250°C and a spray voltage of 2,500 V.

2.7. Enzyme-Linked Immunosorbent Assay (ELISA) Analysis. The 100 serum samples in the experiment were analyzed using an ELISA for transmembrane protein 143, cadherin 5, fibronectin 1, and collectin-11 according to the manufacturer's instructions.

2.8. Data Analysis. The MS/MS spectra were searched using the Proteome Discoverer 2.2 against a nonredundant Sequest database (released in January 2010; Homo sapiens, 20,367 entries). For protein identification, we set thresholds of 10 ppm for intact peptide tolerance masses and 0.02 Da for fragment ions. The analysis allowed for two missed cleavages from the trypsin digest, and iTRAQ (N-terminal, 144 Da), iTRAQ (Lys, 144 Da), oxidized methionine (16 Da), and carbamidomethyl (C, 57 Da) were set as potential variable modifications. Each peptide integrated intensity was normalized to the sum of its channel intensities. The normalized channels were averaged over all peptides of a protein, and the standard deviation of the mean was determined for each normalized channel of a peptide. The results of the SEQUEST database search for each reversed-phase elution were further analyzed (Table S1, serum; Table S2, urine). A difference multiple greater than 1.3 or less than 0.77 in serum and that greater than 1.5 or less than 0.67 in urine were considered statistically significant. The clinical characteristics were compared using Student's t test, Fisher's exact test, or Wilcoxon rank test, whichever was appropriate. The concentrations of transmembrane protein 143, cadherin 5, fibronectin 1, and collectin-11 in the different stages of lung cancer were compared to those of the control group using Student's t test. Data are expressed as the mean ± standard deviation (SD). Results were analyzed using SPSS 8.0 software and Origin 8.5 statistics. The analysis of variance was performed to determine any significant differences ($P \leq 0.05$).

3. Results

3.1. Identification of Removed High-Abundance Proteins. We used Pierce™ Top 12 Abundant Protein Depletion Spin Columns to remove high-abundance proteins from each serum group, as indicated by SDS-PAGE electrophoresis (Figure 1). After the removal of high-abundance proteins, the distribution of protein bands was significantly better than that of samples without the removal of high-abundance proteins. The eight samples separated well after removal of high-abundance proteins; the bands were similar, and the color depth was consistent among the samples.

3.2. Identification of Polypeptides in Serum and Urine Samples. We examined the iTRAQ labeling efficiency. A total of 12,155 peptides were identified in the urine sample, and 12,153 peptides were labeled using the iTRAQ reagent. Thus, the labeling efficiency of iTRAQ was 99.9%. The labeling efficiency of the serum sample reached 99.86%.

According to the LC-MS analysis (Table S1, serum; Table S2, urine), the relative molecular mass distribution of proteins was mainly concentrated below 200 kDa in the serum and urine samples, because the experiment used the classic bottom-up technique (Figures 2(a) and 2(b)). Proteins with molecular weights of up to 540 kDa were identified. Peptide sequence lengths ranged from 10 to 13 peaks; polypeptide lengths were concentrated in the 6–25 range, and 90% of peptide lengths were within 24 kDa, as expected (Figures 2(c) and 2(d)). The theoretical distribution was fitted to a sixth degree polynomial, and R^2 was greater than 0.95. On the polypeptide m/z distribution map, the abscissa is m/z, and the ordinate is the number of polypeptides. Most of the polypeptides had an m/z of 400–1,200, and as m/z gradually increased, the number of identified polypeptides gradually decreased, as expected (Figures 2(e) and 2(f)). In the final correlation analysis, the normalized protein abundance values were used to analyze the correlation between the samples. Figures 2(g) and 2(h) are the correlation map and correlation coefficient matrix analysis of serum and urine

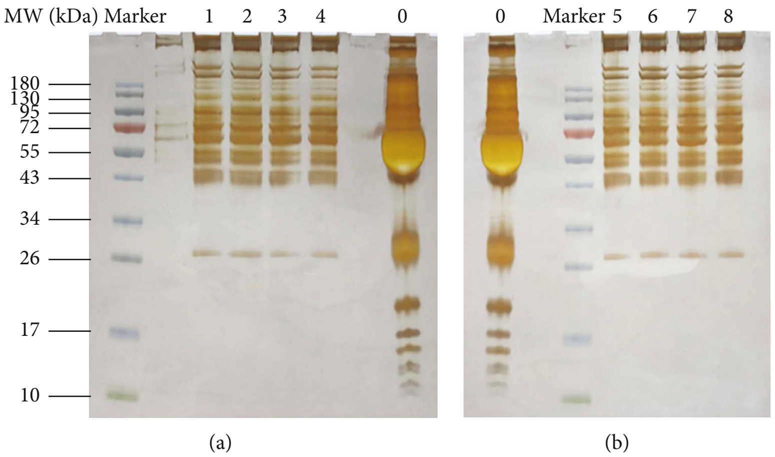


Figure 1: SDS-PAGE electrophoresis of removal high-abundance proteins. 0: nothing was done; 1: healthy control; 2: stage Ia1; 3: stage Ia2; 4: stage Ia; 5: stage Ib; 6: stage II; 7: stage III; 8: stage IV.

samples. Correlation between the samples was very high, and the correlation coefficient was close to 1.

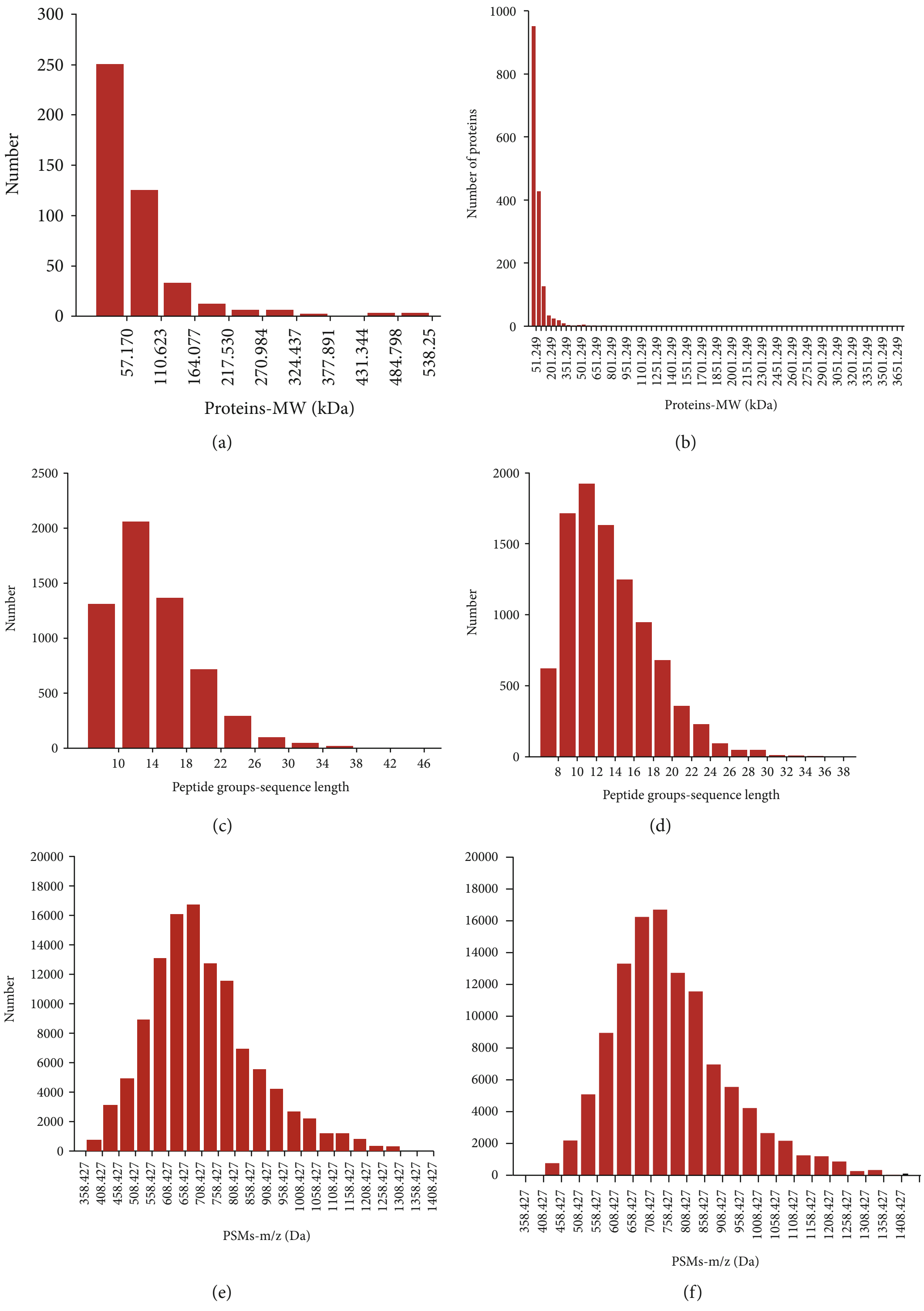
3.3. Identification of Differentially Expressed Proteins in the Urine. In the urine, 1,161 proteins were identified. Urine samples from healthy controls were compared to urine samples from different lung adenocarcinoma groups to obtain differential protein levels. Proteins that were upregulated by >1.5-fold or downregulated by <0.67-fold in urine were considered significant. Based on this rule, levels of 461 urine proteins were increased, and levels of 332 proteins were decreased (Table 2 and Table S3, urine). Changes in urine proteins are indicated on a heat map (Figure 3(a)). We had carried on the thorough analysis to the result and discovered the expression of the prostate-specific antigen, and prostatic acid phosphatase decreased significantly in the metastatic group (stages III and IV) and neutrophil defensin 1 increased significantly in the metastatic group compared to the nonmetastatic group (stages I and II) (Figures 3(c)–3(e)). Changes in these proteins were consistent with those reported in the literature.

3.4. Identification of Differentially Expressed Proteins in the Serum. A total of 441 proteins were identified in serum samples. Serum samples from healthy controls were compared to serum samples from different lung adenocarcinoma groups to obtain differential protein levels. Proteins that were upregulated by >1.3-fold or downregulated by <0.77-fold in serum were considered significant. Based on this rule, 425 proteins increased, and 73 proteins decreased (Table 3 and Table S4 serum), and then we made the heat map for the changed in serum proteins (Figure 3(b)). Elongation factor 1-alpha 2 (Q05639), proteasome subunit alpha type (B2RDG0), and a spermatogenesis-associated protein (A0A0R4J2F1) were significantly increased (>1.3-fold) in all stages of lung adenocarcinoma compared with those of healthy controls (Figures 4(a)–4(c)). Transmembrane protein 143 (Q96AN5), cadherin 5 (Q59EA3), fibronectin 1 (A0A024R462), and collectin-11 (Q9BWP8) were significantly lower (<0.77-fold) in adenocarcinoma cancer stages III and IV compared with the level of expression in stages I and II (untransferred) (Figures 4(d)–4(g)). ELISAs were used to verify the four proteins, and the results were consistent with those of the LC-MS/MS (Figures 4(h)–4(k)).

3.5. Bioinformatics Analysis of Differential Proteins in Serum Samples. Significantly differentially expressed proteins from the LC-MS/MS were analyzed by Blast2Go and clusterProfiler software with Human as the background library, followed by GO annotations, including a biological process (BP), a cellular component (CC), and a molecular function (MF) (Figures 5(a)–5(c)).

In the BP analysis (Figure 5(a)), the biological processes associated with cell proliferation, including DNA replication, chromosome assembly, and organization, as well as beta-catenin-TCF complex assembly, were significantly altered relative to normal samples. Changes in biological processes also changed cellular components and molecular functions. In the CC analysis (Figure 5(b)), significant changes in cancer patients included the interaction of cellular genetic material and DNA proteins. In the MF analysis (Figure 5(c)), changes in histone binding were most pronounced, because histidine is abundantly present in the cell chromatin, and the protein regulates changes in genetic material by binding to histones. Taken together, the GO analysis demonstrated significant changes in protein function in lung cancer patients, particularly those associated with cell proliferation, differentiation, and metastasis.

3.6. Bioinformatics Analysis of the Differential Proteins in Urine Samples. In urine samples, we used Blast2Go and clusterProfiler software to analyze differentially expressed proteins with human as the background library, followed by GO annotations, including BP, CC, and MF (Figures 6(a)–6(c)). The results of the BP analysis were similar to those for the serum samples, which are consistent with changes in DNA replication and chromosome assembly, and they verify the rapid infinite value added by cancer cells (Figure 6(a)). However, in CC and MF analyses, urine samples still exhibited changes different from serum. In the CC



(a)

(b)

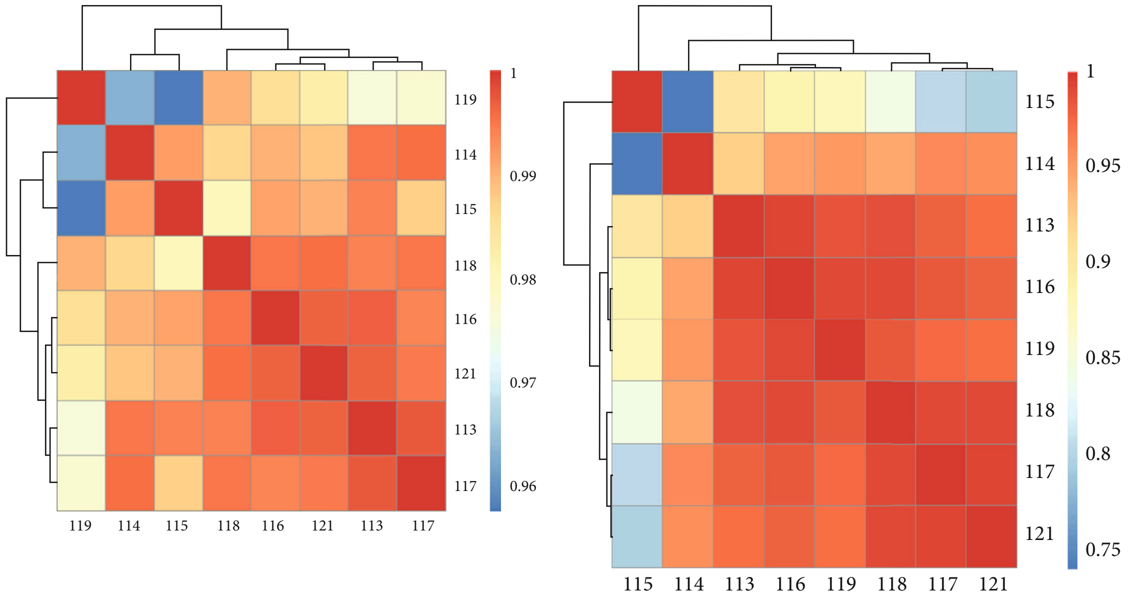
(c)

(d)

(e)

(f)

Figure 2: Continued.



(g)

(h)

Figure 2: (a) Molecular weight distribution of serum. (b) Molecular weight distribution of urine. (c) The number of peptides of different lengths in the serum specimen. (d) The number of peptides of different lengths in the urine specimen. (e) The distribution of PSMs *m*/*z* in the serum specimen. (f) The distribution of PSMs *m*/*z* in the urine specimen. (g) The diagram of correlation coefficient array in the serum specimen. (h) The diagram of correlation coefficient array in the urine specimen.

Table 2: Differential protein statistics of urine samples from different stages of lung adenocarcinoma.

	113/121	114/121	115/121	116/121	117/121	118/121	119/121
Upregulated	142	74	47	33	38	77	50
Downregulated	10	42	197	14	6	26	37

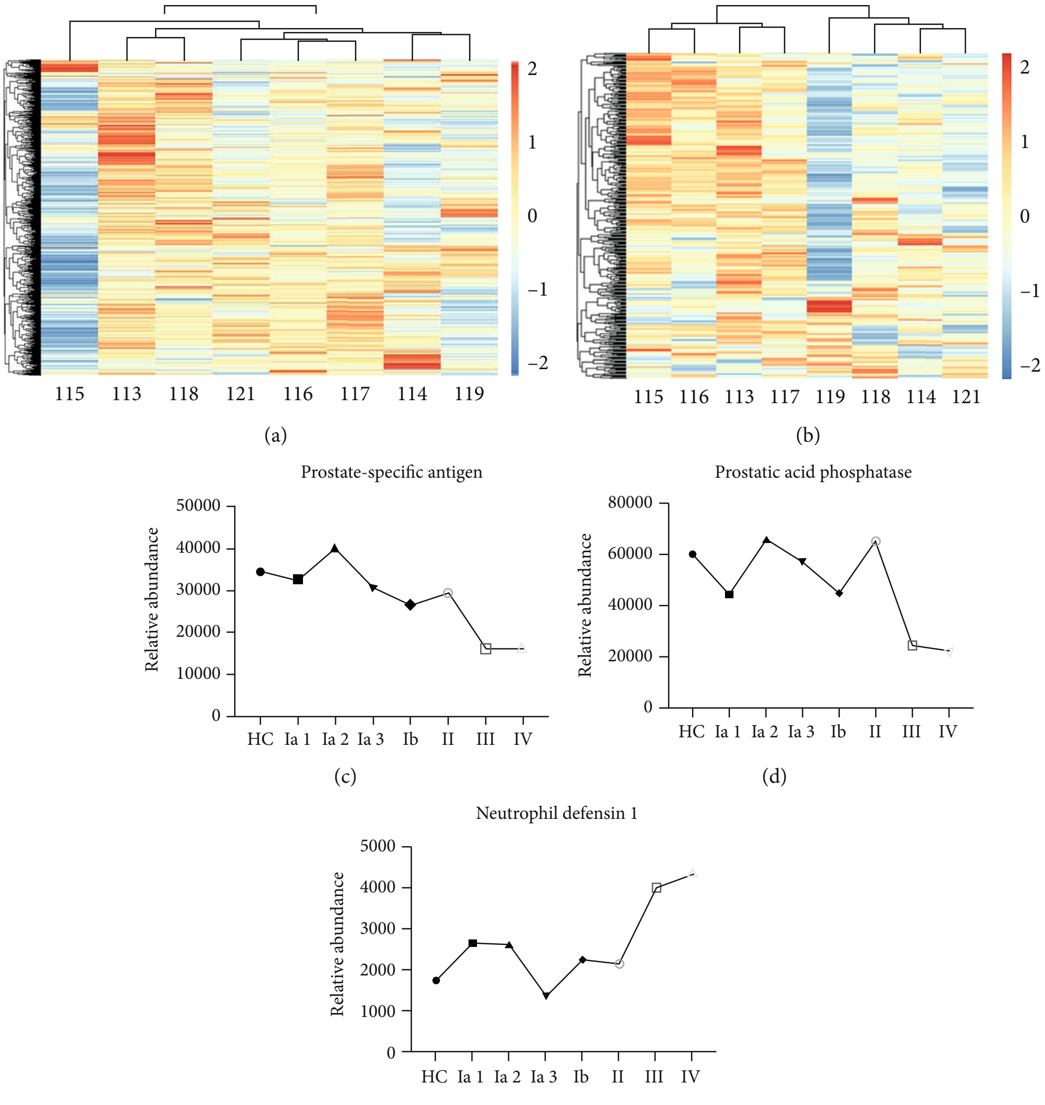
121: control group; 113: stage Ia 1; 114: stage Ia2; 115: stage Ia3; 116: stage Ib; 117: stage II; 118: stage III; 119: stage IV.

analysis (Figure 6(b)), the result of urine analysis differed from the serum result, mainly reflecting the fact that in the urine, the vesicle lumens changed significantly in the lung cancer group. In the MF analysis (Figure 6(c)), changes in histone binding and protein heteromerization were consistent with the results in the serum.

4. Discussion

Using iTRAQ8-plex labeling combined with liquid chromatography-tandem mass spectrometry as a quantitative proteomics analysis method of analyzing the serum and urine samples of healthy controls and stage I, II, III, and IV non-small-cell lung adenocarcinoma patients, we identified a total of 441 proteins in serum samples and 1,161 proteins in urine. In the serum samples, elongation factor 1-alpha 2, proteasome subunit alpha type, and the spermatogenesis-associated protein increased significantly in all stages of lung adenocarcinoma compared with healthy controls. In stage III and IV patients, the expression levels of transmembrane protein 143, cadherin 5, fibronectin 1, and collectin-11 decreased significantly compared with those in stages I and II. In urine samples, the prostate-specific antigen and prostatic acid phosphatase levels were significantly decreased, whereas neutrophil defensin 1 was significantly elevated in stage III and IV lung adenocarcinoma patients. These results are consistent with previous reports verifying that the method is reliable. Moreover, these differentiated proteins may serve as potential diagnostic markers for the early diagnosis of lung adenocarcinoma. In addition, combining the markers in serum and urine may distinguish different stages of lung adenocarcinoma.

Due to the rapid division and proliferation of cancer cells, the synthesis and metabolism of nucleic acids and proteins are increased compared with those in normal cells. Thus, the associated protein content is increased [26, 27] relative to normal cells. Elongation factor 1-alpha 2 has been reported in the literature as a potential marker for cancer [17]. The proteasome subunit alpha type is significantly increased in all cancers, consistent with literature reports [17, 27, 28]. Protein ubiquitination and subsequent proteolysis and degradation by the proteasome are important mechanisms in the cell cycle, cell growth and differentiation, gene transcription, signal transduction, and apoptosis [29]. Proteasomes hydrolyze cells and control apoptosis. Proteosomes bind to the p21 protein and inhibit p21 protein activity



(a)

(b)

(c)

(d)

(e)

Figure 3: Statistical analysis of differentially expressed proteins and urine protein markers with significant differences between early and late stage lung adenocarcinomas. (a) The heat map of changed in urine proteins. (b) The heat map of changed in serum proteins. (c) Prostate-specific antigen expression in lung adenocarcinoma groups was significantly lower in the metastatic groups than in the non-metastasis groups. (d) Prostatic acid phosphatase expression of the metastatic groups was significantly lower than that of the nonmetastatic groups in lung adenocarcinoma. (e) The expression of neutrophil defensin in the lung adenocarcinoma groups was significantly higher than that in the metastatic group than that in the nonmetastatic groups.

Table 3: Differential protein statistics of serum samples from different stages of lung adenocarcinoma.

iTRAQ	113/121	114/121	115/121	116/121	117/121	118/121	119/121
Upregulated	93	41	98	78	51	35	29
Downregulated	5	4	4	8	5	8	39

121: control group; 113: stage Ia 1; 114: stage Ia2; 115: stage Ia3; 116: stage Ib; 117: stage II; 118: stage III; 119: stage IV.

[30], leading to cell cycle disorders and, ultimately, cancer. Therefore, the increased expression of this protein in cancer cells inhibits the tumor-suppressing effect of p21. The proteasome subunit alpha type may be significantly associated with the protein metabolism in the body, and its expression is associated with cancer [30, 31]. The spermatogenesis-associated protein is a spermatogenesis-related protein that is mainly found in mammalian tissues and is significantly elevated in prostate cancer tissues [32], testicular cancer tissues [32, 33], and breast cancer tissues [34]. The expression of spermatogenesis-associated 6 (SPATA6) is significantly elevated in testicular cancer, and inhibition of SPATA6

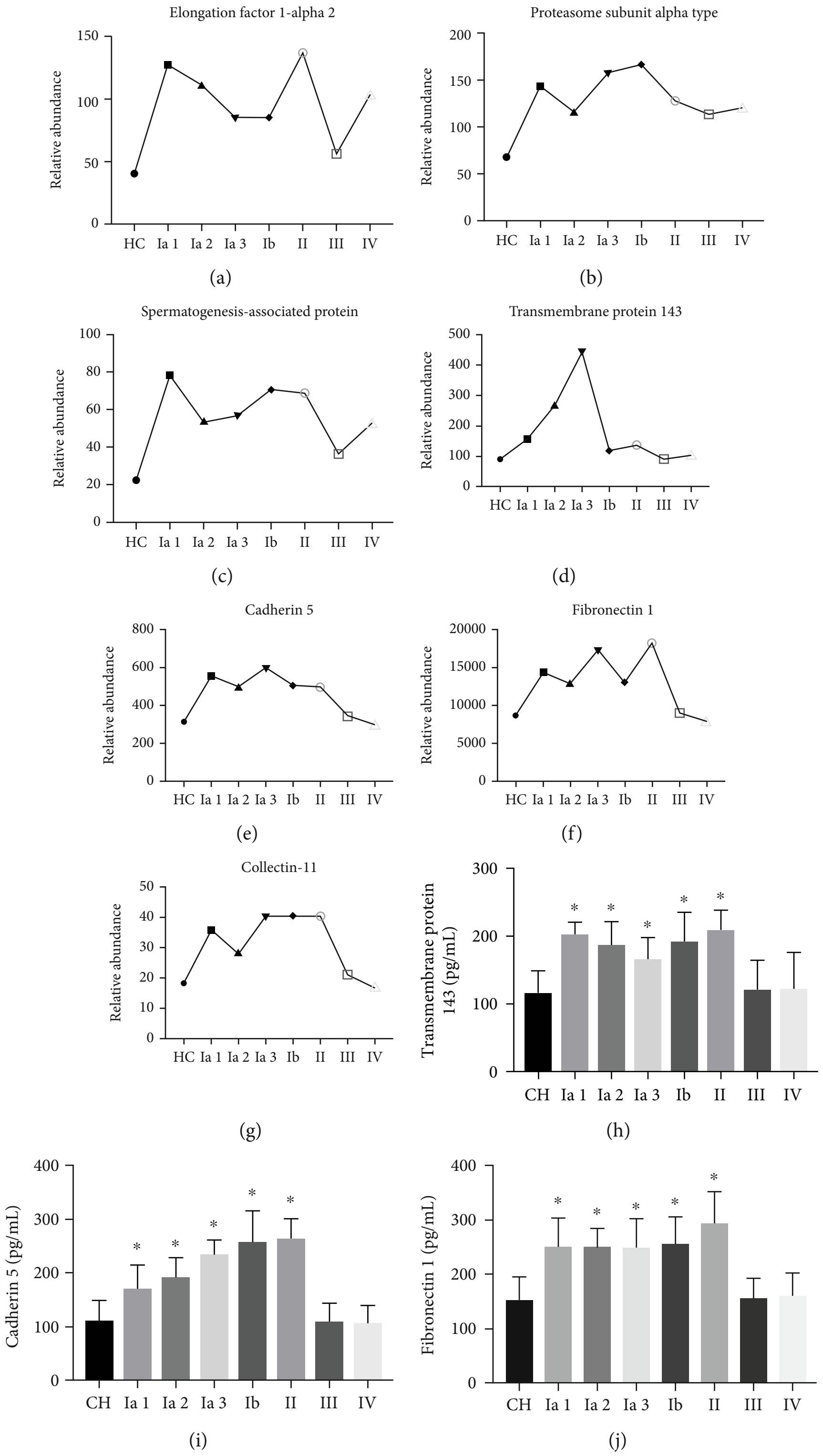
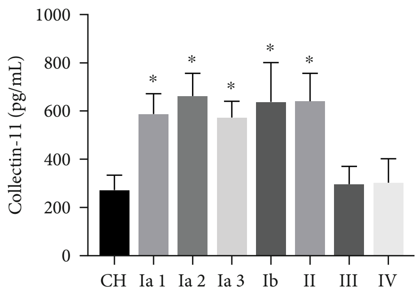
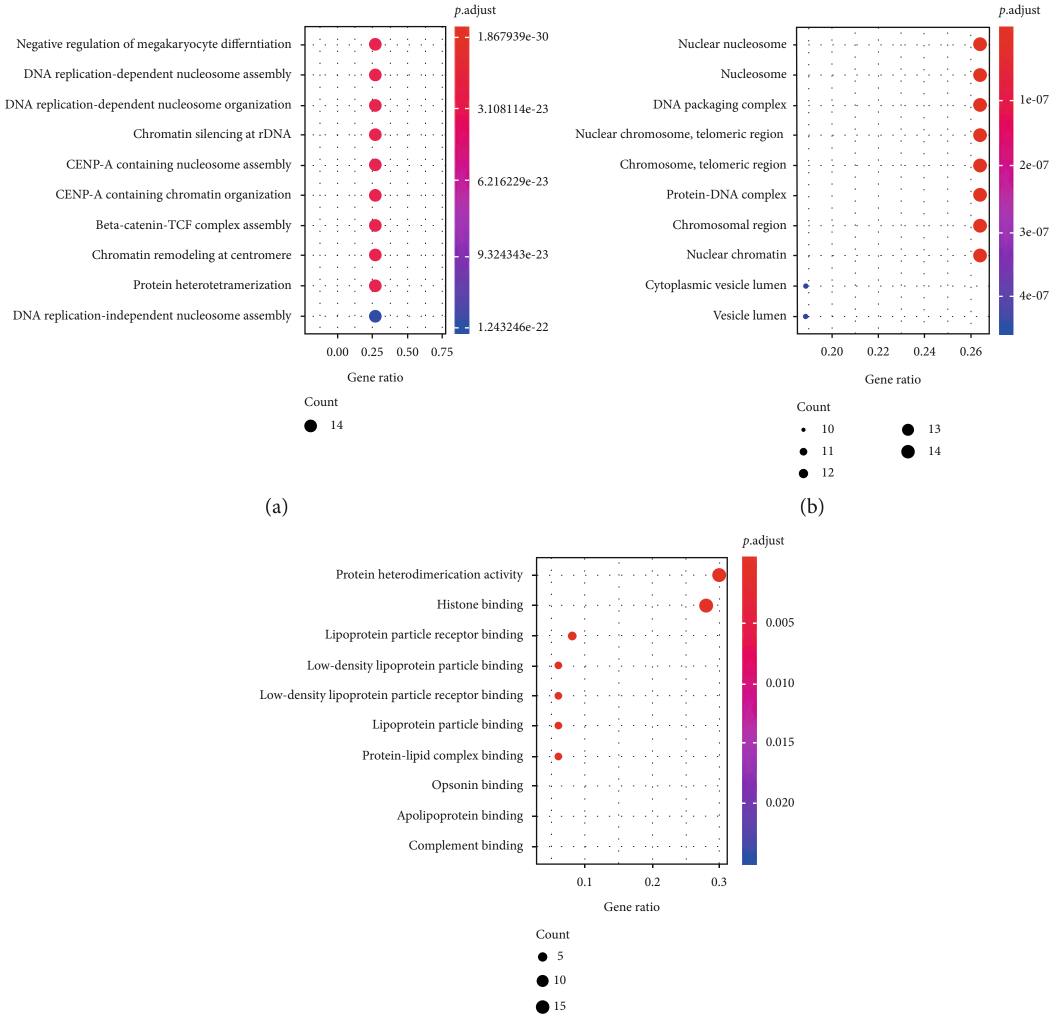


Figure 4: Continued.



(k)

Figure 4: Differences the proteins in serum samples. (a–c) The level of expression elongation factor 1-alpha 2, proteasome subunit alpha type, and spermatogenesis-associated in each group. (d–g) LC-MS/MS of transmembrane protein 143, cadherin 5, fibronectin 1, and collectin-11; (h, k) ELISA of transmembrane protein 143, cadherin 5, fibronectin 1, and collectin-11. $^*P < 0.05$.



(a)

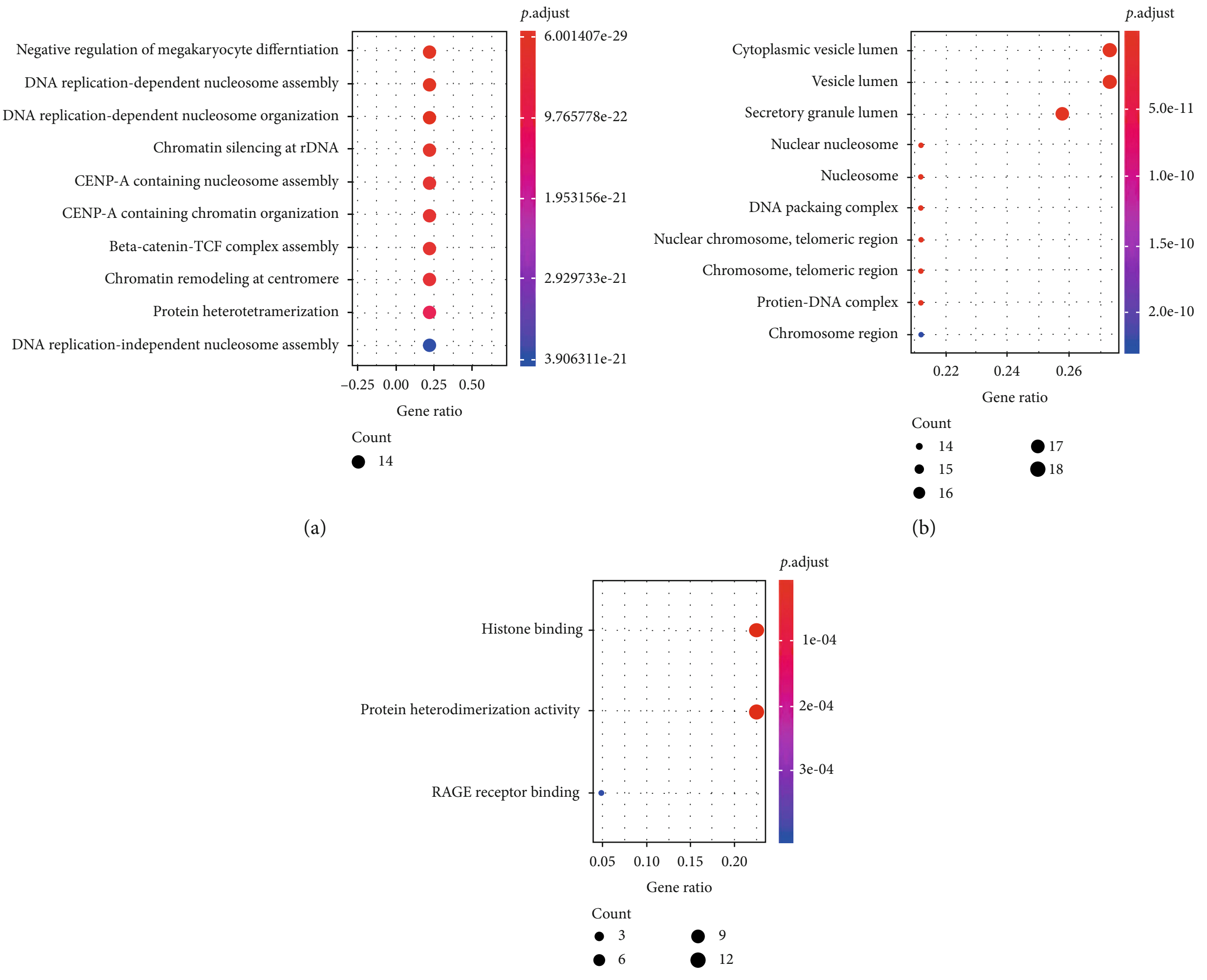
(b)

(c)

Figure 5: The GO analysis the protein markers in the serum samples. (a) Serum biological process (BP). (b) Serum cellular component (CC). (c) Serum molecular function (MF).

expression can cause cancer cells to die [35]. SPATA20 is significantly elevated in cholangiocarcinoma [36]. Our results are consistent with those reported in the literature.

The prostate-specific antigen (PSA) and prostatic acid phosphatase are markers for the diagnosis of prostate cancer [37]; elevated total PSA (tPSA) and free PSA (fPSA) and



(a)

(b)

(c)

Figure 6: The GO analysis the protein markers in the urine samples. (a) Urine Biological process (BP). (b) Urine cellular component (CC). (c) Urine molecular function (MF).

decreased tPSA/fPSA indicate prostate cancer [38]. We demonstrated a significant downregulation of PSA in metastatic lung cancer urine relative to nontransfer of lung cancer. Neutrophil defensin 1 protein is important in the neutrophil defense system and has antitumor effects [39]. In advanced metastatic lung cancer, the metabolism and activity of the tumor are increased, and neutrophil defensin 1 is part of its defense system. Thus, neutrophil defensin 1 can be used as an indicator of cancer prognosis.

Transmembrane protein 143 was present at different levels in different stages of lung cancer and may be important in the early diagnosis and prognosis of cancer. These results are novel and have not been reported. E-cadherin acts as an invasion suppressor and a classical tumor suppressor gene. E-cadherin is a biomarker for cancer [18, 40] and is attenuated and reduced in many cancers [41–43], especially in lung and breast cancers. E-cadherin loss in tumor cells leads to decreased adhesion between tumor cells, which favors epithelial-mesenchymal transition [44] and promotes the ability of tumor cells to invade and metastasize. In metastatic lung cancer (stages III and IV), E-cadherin protein levels were significantly lower than those in nonmetastatic stages (I and II). Fibronectin can regulate cancer and the migration of cancer cells, which is closely associated with the prognosis of tumor formation and development. However, the mechanisms underlying this relationship are unclear [45]. Fibronectin is a biomarker [18, 45–48], which is upregulated in cancer, and our results are consistent with the literature. Fibronectin 1 can bind to cancer cells and favors metastasis and invasion [48, 49]. Cancer prognosis studies demonstrate that the higher the level of fibronectin 1 in vivo, the worse the prognosis and survival rate [45]. When cancer occurs, fibronectin 1 expression increases, promoting adhesion and invasion of cancer cells and increasing the damage to normal tissues. Therefore, fibronectin 1 is also an important indicator of cancer prognosis. After cancer treatment, if fibronectin 1 levels are still high, the prognosis is poor.

Collectin-11 decreased significantly in stage III and IV adenocarcinomas compared to earlier stages. Collectins are a family of collagenous calcium-dependent defense lectins in animals [50]. Collectins are humoral molecules of the innate immune system that modulate inflammatory and allergic responses, the adaptive immune system, and clearance of apoptotic cells [51, 52]. Collectin-11 is directly related to cancer and the human immune system. In stage I and stage II lung cancer, collectin-11 is elevated because the human immune system is still in the early stage of cancer development. In stage III and stage IV cancer, collectin-11 is decreased, by the massive proliferation and metastasis of cancer cells, a process that destroys the normal immune system. This report is the first to identify changes in collectin-11 in lung cancer, which can be used as a marker for staging non-small-cell lung cancer.

Based on the biological analysis of differential proteins in serum and urine, the protein changes in serum and urine were similar, indicating that the biological processes of blood and urine were interrelated and consistent. In the CC analysis, the changes in urine vesicles in the lung cancer group were more pronounced compared to those in the blood. The results of serum and urine are partially consistent, but urine and serum also have unique characteristics, so the combination provides a better reference for the selection of biomarkers. In summary, these differentiated proteins may be potential diagnostic markers for lung adenocarcinoma and may serve as a basis for the early diagnosis of lung adenocarcinoma. Further research is required to verify the experimental results.

Data Availability

The datasets used and/or analyzed during the present study are available from Tables S1, S2, S3, and S4.

Ethical Approval

The present study was approved by the ethics committee of the Chinese PLA General Hospital (Number S2018-007-001).

Consent

Informed consent was obtained from each patient with available follow-up information.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Authors' Contributions

QZ, CBW, and LW participated in the design and conception of the present study. MRZ, YK, XBW, and WL were involved in data acquisition, analysis of the clinical, and analysis the data. The manuscript was written by QZ, and revision of the manuscript was done by QZ, SC, and MRZ. All authors read and approved the final manuscript.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 81902067 and 81272115), the Natural Science Foundation of Shenzhen University General Hospital (No. SUGH2018QD037), the Shenzhen Science and Technology Innovation Committee (No. JCY20180305124849781), the Beijing Natural Science Foundation (No. 7204306), and the fellowship of China Postdoctoral Science Foundation (Nos. 2020M670047ZX, 2020M670048ZX, and 2020M682888.)

Supplementary Materials

Supplementary 1. Table S1: the results of serum reversed-phase elution test and searching in the SEQUEST database.

Supplementary 2. Table S2: the results of urine reversed-phase elution test and searching in the SEQUEST database.

Supplementary 3. Table S3: a further study on the LC-MS analysis results of urine in Table S2. Urine samples from healthy controls were compared to urine samples from different lung adenocarcinoma groups to obtain differential protein levels.

Supplementary 4. Table S4: a further study on the LC-MS analysis results of serum in Table S1. Serum samples from healthy controls were compared to serum samples from different lung adenocarcinoma groups to obtain differential protein levels.

References

[1] A. Ghosh and H. Yan, "Stability analysis at key positions of EGFR related to non-small cell lung cancer," *Current Bioinformatics*, vol. 15, no. 3, pp. 260–267, 2020.

[2] T. Zou, X. Mao, J. Yin et al., "Emerging roles of RAC1 in treating lung cancer patients," *Clinical Genetics*, vol. 91, no. 4, pp. 520–528, 2017.

[3] L. A. Byers and C. M. Rudin, "Small cell lung cancer: where do we go from here?," *Cancer*, vol. 121, no. 5, pp. 664–672, 2015.

[4] F. Yuan, L. Lu, and Q. Zou, "Analysis of gene expression profiles of lung cancer subtypes with machine learning algorithms," *Biochimica Et Biophysica Acta-Molecular Basis of Disease*, vol. 1866, no. 8, article 165822, 2020.

[5] K. Mizuno, H. Mataki, N. Seki, T. Kumamoto, K. Kamikawaji, and H. Inoue, "MicroRNAs in non-small cell lung cancer and idiopathic pulmonary fibrosis," *Journal of Human Genetics*, vol. 62, no. 1, pp. 57–65, 2017.

[6] F. A. Shepherd, J. Crowley, P. Van Houtte et al., "The International Association for the Study of Lung Cancer Lung Cancer Staging Project: proposals for the revision of the clinical and pathologic staging of small cell lung cancer in the forthcoming Eighth Edition of the TNM Classification for Lung Cancer," *Journal of Thoracic Oncology*, vol. 2, no. 12, pp. 1067–1077, 2007.

[7] A. S. Panja, A. Nag, B. Bandopadhyay, and S. Maiti, "Protein stability determination (PSD): a tool for proteomics analysis," *Current Bioinformatics*, vol. 14, no. 1, pp. 70–77, 2019.

[8] H.-J. Sung, J.-M. Ahn, Y.-H. Yoon et al., "Quiescin sulfhydryl oxidase 1 (QSOX1) secreted by lung cancer cells promotes cancer metastasis," *International Journal of Molecular Sciences*, vol. 19, no. 10, article 3213, 2018.

[9] A. S. Al-Wajeeh, M. N. Ismail, S. M. Salhimi, I. A. Khalid, and A. B. A. Latiff, "Identification of glycobiomarker candidates for

breast cancer using LTQ-orbitrap fusion technique," *International Journal of Pharmacology*, vol. 13, no. 5, pp. 425–437, 2017.

[10] R. Chen, Y. Liu, H. Zhuang et al., "Quantitative proteomics reveals that long non-coding RNA MALAT1 interacts with DBC1 to regulate p53 acetylation," *Nucleic Acids Research*, vol. 45, no. 17, pp. 9947–9959, 2017.

[11] S. Arshid, M. Tahir, B. Fontes et al., "High performance mass spectrometry based proteomics reveals enzyme and signaling pathway regulation in neutrophils during the early stage of surgical trauma," *PROTEOMICS - Clinical Applications*, vol. 11, no. 1-2, article 1600001, 2017.

[12] C. A. Granville and P. A. Dennis, "An overview of lung cancer genomics and proteomics," *American Journal of Respiratory Cell and Molecular Biology*, vol. 32, no. 3, pp. 169–176, 2005.

[13] A. C. Iliopoulos, G. Beis, P. Apostolou, and I. Papasotiriou, "Complex networks, gene expression and cancer complexity: a brief review of methodology and applications," *Current Bioinformatics*, vol. 15, no. 6, pp. 629–655, 2020.

[14] M. F. Sharpnack, N. Ranbaduge, A. Srivastava et al., "Proteogenomic analysis of surgically resected lung adenocarcinoma," *Journal of Thoracic Oncology*, vol. 13, no. 10, pp. 1519–1529, 2018.

[15] Y. Xie, S. Li, L. Sun et al., "Fungal immunomodulatory protein from Nectria haematococca suppresses growth of human lung adenocarcinoma by inhibiting the PI3K/Akt pathway," *International Journal of Molecular Sciences*, vol. 19, no. 11, article 3429, 2018.

[16] J. Wang, H. Wang, X. Wang, and H. Chang, "Predicting drug-target interactions via FM-DNN learning," *Current Bioinformatics*, vol. 15, no. 1, pp. 68–76, 2020.

[17] R. Jakhar, M. Dangi, A. Khichi, and A. K. Chhillar, "Relevance of molecular docking studies in drug designing," *Current Bioinformatics*, vol. 15, no. 4, pp. 270–278, 2020.

[18] F. Li, D. Zhao, S. Yang et al., "ITRAQ-based proteomics analysis of triptolide on human A549 lung adenocarcinoma cells," *Cellular Physiology and Biochemistry*, vol. 45, no. 3, pp. 917–934, 2018.

[19] P. Widlak, M. Pietrowska, J. Polanska et al., "Serum mass profile signature as a biomarker of early lung cancer," *Lung Cancer*, vol. 99, pp. 46–52, 2016.

[20] W. Tang, S. Wan, Z. Yang, A. E. Teschendorff, and Q. Zou, "Tumor origin detection with tissue-specific miRNA and DNA methylation markers," *Bioinformatics*, vol. 34, no. 3, pp. 398–406, 2018.

[21] P. Cifani and A. Kentsis, "Towards comprehensive and quantitative proteomics for diagnosis and therapy of human disease," *Proteomics*, vol. 17, no. 1-2, 2017.

[22] Z. Chen, L. Long, K. Wang et al., "Identification of nasopharyngeal carcinoma metastasis-related biomarkers by iTRAQ combined with 2D-LC-MS/MS," *Oncotarget*, vol. 7, no. 23, pp. 34022–34037, 2016.

[23] T. Basak, A. Bhat, D. Malakar, M. Pillai, and S. Sengupta, "In-depth comparative proteomic analysis of yeast proteome using iTRAQ and SWATH based MS," *Molecular BioSystems*, vol. 11, no. 8, pp. 2135–2143, 2015.

[24] M. Sanda and R. Goldman, "Data independent analysis of IgG glycoforms in samples of unfractionated human plasma," *Analytical Chemistry*, vol. 88, no. 20, pp. 10118–10125, 2016.

[25] M. P. Weekes, P. Tomasec, E. L. Huttlin et al., "Quantitative temporal viromics: an approach to investigate host-pathogen interaction," *Cell*, vol. 157, no. 6, pp. 1460–1472, 2014.

[26] Y. Sun, S. Liu, Z. Qiao et al., "Systematic comparison of exosomal proteomes from human saliva and serum for the detection of lung cancer," *Analytica Chimica Acta*, vol. 982, pp. 84–95, 2017.

[27] H. Song, H. Xiong, J. Che et al., "Gel-based chemical cross-linking analysis of 20S proteasome subunit-subunit interactions in breast cancer," *J Huazhong Univ Sci Technol Med Sci*, vol. 36, no. 4, pp. 564–570, 2016.

[28] Q. Yang, M. H. Roehrl, and J. Y. Wang, "Proteomic profiling of antibody-inducing immunogens in tumor tissue identifies PSMA1, LAP3, ANXA3, and maspin as colon cancer markers," *Oncotarget*, vol. 9, no. 3, pp. 3996–4019, 2018.

[29] L. C. Cantley and B. G. Neel, "New insights into tumor suppression: PTEN suppresses tumor formation by restraining the phosphoinositide 3-kinase/AKT pathway," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 8, pp. 4240–4245, 1999.

[30] Z. Shi, Z. Li, Z. J. Li et al., "Cables1 controls p21/Cip1 protein stability by antagonizing proteasome subunit alpha type 3," *Oncogene*, vol. 34, no. 19, pp. 2538–2545, 2015.

[31] T. Wang, T. Chen, A. Thakur et al., "Association of PSMA4 polymorphisms with lung cancer susceptibility and response to cisplatin-based chemotherapy in a Chinese Han population," *Clinical & Translational Oncology*, vol. 17, no. 7, pp. 564–569, 2015.

[32] K. K. Wong, F. A. Hussain, S. K. Loo, and J. I. López, "Cancer/-testis antigen SPATA19 is frequently expressed in benign prostatic hyperplasia and prostate cancer," *APMIS*, vol. 125, no. 12, pp. 1092–1101, 2017.

[33] W. Xu, P. Fang, Z. Zhu et al., "Cigarette smoking exposure alters pebp1 DNA methylation and protein profile involved in MAPK signaling pathway in mice testis," *Biology of Reproduction*, vol. 89, no. 6, p. 142, 2013.

[34] G. Kazemi-Oula, S. Ghafouri-Fard, M. B. Mobasheri, L. Geranpayeh, and M. H. Modarressi, "Upregulation of RHOXF2 and ODF4 expression in breast cancer tissues," *Cell Journal*, vol. 17, no. 3, pp. 471–477, 2015.

[35] S. Huo, W. Du, P. Shi, Y. Si, and S. Zhao, "The role of spermatogenesis-associated protein 6 in testicular germ cell tumors," *International Journal of Clinical and Experimental Pathology*, vol. 8, no. 8, pp. 9119–9125, 2015.

[36] J. Shen, W. Wang, J. Wu et al., "Comparative proteomic profiling of human bile reveals SSP411 as a novel biomarker of cholangiocarcinoma," *PLoS One*, vol. 7, no. 10, article e47476, 2012.

[37] S. M. Hanash, E. J. Ostrin, and J. F. Fahrmann, "Blood based biomarkers beyond genomics for lung cancer screening," *Transl Lung Cancer Res*, vol. 7, no. 3, pp. 327–335, 2018.

[38] G. Berx, A. M. Cleton-Jansen, F. Nollet et al., "E-cadherin is a tumour/invasion suppressor gene mutated in human lobular breast cancers," *The EMBO Journal*, vol. 14, no. 24, pp. 6107–6115, 1995.

[39] K. F. Becker, M. J. Atkinson, U. Reich et al., "E-cadherin gene mutations provide clues to diffuse type gastric carcinomas," *Cancer Research*, vol. 54, pp. 3845–3852, 1994.

[40] W. J. F. de Leeuw, G. Berx, C. B. J. Vos et al., "Simultaneous loss of E-cadherin and catenins in invasive lobular breast cancer and lobular carcinoma in situ," *The Journal of Pathology*, vol. 183, no. 4, pp. 404–411, 1997.

[41] K. Polyak and R. A. Weinberg, "Transitions between epithelial and mesenchymal states: acquisition of malignant and stem cell traits," *Nature Reviews. Cancer*, vol. 9, no. 4, pp. 265–273, 2009.

[42] J. Xiao, W. Yang, B. Xu et al., "Expression of fibronectin in esophageal squamous cell carcinoma and its role in migration," *BMC Cancer*, vol. 18, no. 1, p. 976, 2018.

[43] H.-Y. Wu, F.-L. Yang, L.-H. Li et al., "Ergosterol peroxide from marine fungus *Phoma sp*. induces ROS-dependent apoptosis and autophagy in human lung adenocarcinoma cells," *Scientific Reports*, vol. 8, no. 1, article 17956, 2018.

[44] N. El-Hachem, N. Habel, T. Naiken et al., "Uncovering and deciphering the pro-invasive role of HACE1 in melanoma cells," *Cell Death and Differentiation*, vol. 25, no. 11, pp. 2010–2022, 2018.

[45] F. Di Modugno, S. Spada, B. Palermo et al., "hMENA isoforms impact NSCLC patient outcome through fibronectin/β1 integrin axis," *Oncogene*, vol. 37, no. 42, pp. 5605–5617, 2018.

[46] F. Haun, S. Neumann, L. Peintner et al., "Identification of a novel anoikis signalling pathway using the fungal virulence factor gliotoxin," *Nature Communications*, vol. 9, no. 1, article 3524, 2018.

[47] J. K. Van De Wetering, L. M. G. Van Golde, and J. J. Batenburg, "Collectins: players of the innate immune system," *European Journal of Biochemistry*, vol. 271, no. 7, pp. 1229–1249, 2004.

[48] G. Gupta and A. Surolia, "Collectins: sentinels of innate immunity," *BioEssays*, vol. 29, no. 5, pp. 452–464, 2007.

[49] A. Nayak, E. Dodagatta-Marri, A. G. Tsolaki, and U. Kishore, "An insight into the diverse roles of surfactant proteins, SP-A and SP-D in innate and adaptive immunity," *Frontiers in Immunology*, vol. 3, p. 131, 2012.

[50] J. E. Heller, "Prostatic acid phosphatase: its current clinical status," *The Journal of Urology*, vol. 137, no. 6, pp. 1091–1103, 1987.

[51] L. G. Gomella, X. S. Liu, E. J. Trabulsi et al., "Screening for prostate cancer: the current evidence and guidelines controversy," *The Canadian Journal of Urology*, vol. 18, no. 5, pp. 5875–5883, 2011.

[52] J. Bingham, H. Clarke, M. Spangehl, A. Schwartz, C. Beauchamp, and B. Goldberg, "The alpha defensin-1 biomarker assay can be used to evaluate the potentially infected total joint arthroplasty," *Clinical Orthopaedics and Related Research*, vol. 472, no. 12, pp. 4006–4009, 2014.

Hindawi
BioMed Research International
Volume 2021, Article ID 6639698, 10 pages
https://doi.org/10.1155/2021/6639698

Research Article

Identifying the Immunological Gene Signatures of Immune Cell Subtypes

Yu-Hang Zhang[1,2] **Zhandong Li**[3] **Tao Zeng**[4] **WenCong Lu**[5] **Tao Huang**[6] **and Yu-Dong Cai**[1]

[1]*School of Life Sciences, Shanghai University, Shanghai 200444, China*
[2]*Channing Division of Network Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA*
[3]*College of Food Engineering, Jilin Engineering Normal University, Changchun, China*
[4]*Bio-Med Big Data Center, CAS Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, Shanghai 200031, China*
[5]*Department of Chemistry, College of Sciences, Shanghai University, Shanghai 200444, China*
[6]*Key Laboratory of Tissue Microenvironment and Tumor, Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, Shanghai 200031, China*

Correspondence should be addressed to Tao Huang; tohuangtao@126.com and Yu-Dong Cai; cai_yud@126.com

Received 24 December 2020; Revised 25 January 2021; Accepted 10 February 2021; Published 19 February 2021

Academic Editor: Luis Alberto Morales Quintana

The immune system is a complicated defensive system that comprises multiple functional cells and molecules acting against endogenous and exogenous pathogenic factors. Identifying immune cell subtypes and recognizing their unique immunological functions are difficult because of the complicated cellular components and immunological functions of the immune system. With the development of transcriptomics and high-throughput sequencing, the gene expression profiling of immune cells can provide a new strategy to explore the immune cell subtyping. On the basis of the new profiling data of mouse immune cell gene expression from the Immunological Genome Project (ImmGen), a novel computational pipeline was applied to identify different immune cell subtypes, including $\alpha\beta$ T cells, B cells, $\gamma\delta$ T cells, and innate lymphocytes. First, the profiling data was analyzed by a powerful feature selection method, Monte-Carlo Feature Selection, resulting in a feature list and some informative features. For the list, the two-stage incremental feature selection method, incorporating random forest as the classification algorithm, was applied to extract essential gene signatures and build an efficient classifier. On the other hand, a rule learning scheme was applied on the informative features to construct quantitative expression rules. A group of gene signatures was found as qualitatively related to the biological processes of four immune cell subtypes. The quantitative expression rules can efficiently cluster immune cells. This work provides a novel computational tool for immune cell quantitative subtyping and biomarker recognition.

1. Introduction

The immune system is a complicated defensive system that comprises multiple functional cells and molecules acting against endogenous and exogenous pathogenic factors [1–3]. With organism evolution, the immune system gradually becomes complicated and finally forms layered defensive mechanisms, including innate and adaptive immune systems, in advanced creatures such as mammals [4, 5]. Both these immune systems are complicated and constitute cellular (immune cells) and noncellular components (immunoregulatory molecules) [6]. The cellular component is diverse on structural and functional levels [6]. In particular, each single immune response could involve multiple subtypes of immune cells, and each immune cell subtype may play various important roles in multiple immune responses, thereby constituting a complicated regulatory network on the cellular level [6].

Identifying the subtypes of immune cells and recognizing their unique immunological functions are difficult due to the complicated cellular components and immunological functions of the immune system. The only standards are the typical molecular markers recognized by cytobiology [7–9]. However, such biomarkers cannot accurately reflect the components of immune cells and reveal their immunoregulatory mechanisms in vivo. With the development of transcriptomics and high-throughput sequencing [10, 11], the gene expression profiling of immune cells can provide a new strategy to explore the complicated immunoregulatory mechanisms, e.g., detailed immune cell subtyping. Gene expression profiling with transcriptomic analysis can reflect the typical gene expression pattern of each cell subgroup. In accordance with the central dogma of molecular biology, cells with different gene expression patterns may have varying proteomic features and biological functions and therefore must be clustered into different cell subtypes [12, 13]. Differentially expressed genes or transcripts may be potential biomarkers for the identification of a given cell subgroup/subtype. Therefore, transcriptomic sequencing is a novel technique for immune cell subtyping through the recognition of cell subgroups and their respective biomarkers and functions.

The Immunological Genome Project (ImmGen) [14, 15] is aimed at establishing a systematic panorama of gene expression and regulatory networks of all immune cells by using a mouse model. Initiated in 2008, this collaborative study [15] has analyzed the differentiation, maturation, active responses, effector stages, tissue localizations, and genetic variations of more than 250 subtypes of immune cells in mouse models. A systematic immunoregulatory network in mice, which encompasses the innate and adaptive immune systems, has been established by this project through systematic quality check control and standardized analyzed conditions. The analyzed data and constructed networks can be accessed by using a dedicated data browser, and the raw sequencing and microarray data can be accessed in a public database [16]. This project provides reliable resources for immune cell gene expression profiling for further exploration and research.

As reported by various previous publications, the immune system is composed by multiple immune cell subtypes which are impossible for us to study one by one. For further analyses on the mouse immune cell gene expression profiling, we focused on four basic subtypes of immune cells: $\alpha\beta$ T cells [17], B cells [18], $\gamma\delta$ T cells [19], and innate lymphocytes [20]. Among them, $\alpha\beta$ T cells are the majority of T cells with $\alpha\beta$ TCR, contributing to immune-mediated cell death as adaptive immune responders [21]. As for $\gamma\delta$ T cells, as quite a functioning minority of T cells, which are different from $\alpha\beta$ T cells mainly functioning for adaptive immune responses, there are three major functions for $\gamma\delta$ T cells: (1) regulating other immune cells for central and peripheral immune responses [22, 23], (2) contributing to thermogenesis to maintain body temperature [23], and (3) regulating autoimmune responses [24]. B cells are the main participator for antibody medicated adaptive humoral immune responses with complicated activation processes relying either on T cells or not [25]. Innate lymphocytes are a group of immune cells that participate in the innate immune responses with cytotoxic natural killer (NK) cells in the circulating system and innate lymphoid cells (ILCs) in the tissue-resident microenvironment [26]. Therefore, as introduced above, the four major subgroups of immune cells have quite different biological functions, controlling the basic biological functions of the immune system: innate and adaptive immune responses. Therefore, considering the significance of such four subgroups of cells in the immune system, we selected such immune cell subtypes for detailed classification studies on the murine transcriptomic level in this study.

A new batch of profiling data for mouse immune cell gene expression has been released by ImmGen on the Gene Expression Omnibus (GEO) database provided by NCBI [16]. On the basis of these data, a novel computational pipeline was applied to distinguish different immune cell subtypes including $\alpha\beta$ T cells, B cells, $\gamma\delta$ T cells, and innate lymphocytes. These four subtypes of immune cells are the major effective immune cells in mouse immune systems, contribute to different immune responses, and constitute a complicated immunoregulatory system by playing unique roles and interacting with each other. The powerful feature selection method, the Monte-Carlo Feature Selection (MCFS) [27], was first applied to the profiling data. We obtained a feature list and some informative features. Of the feature list, the two-stage incremental feature selection (IFS) [28] method with random forest (RF) [29] as the classification algorithm was executed to extract essential gene signatures and build an efficient RF classifier. Furthermore, some quantitative expression rules were constructed on the informative features via a rule learning scheme. Extensive analysis on gene signatures and rules were performed by literature review. All in all, we recognized the typical gene signatures and rules of each key immune cell subtype, which were helpful to explore the complicated regulatory mechanisms of immune systems.

2. Materials and Methods

2.1. Dataset. The system-wide mouse RNA-Seq data released by the Immunological Genome Project (ImmGen) were downloaded from GEO (Gene Expression Omnibus) under accession number of GSE109125 [30]. The reads were mapped to the mouse reference genome (mm10), and then the uniquely mapped reads were assigned to genes according to GENCODE annotation (vM12). The genes were quantified as counts per million (CPM) using edgeR [31]. A total of 49,480 genes and 112 samples were obtained from the four types of cells: 46 $\alpha\beta$ T cells, 33 B cells, 13 $\gamma\delta$ T cells, and 20 innate lymphocytes. The original dataset included more samples and cell types (46 $\alpha\beta$T cells, 33 B cells, 20 innate lymphocytes, 13 $\gamma\delta$ T cells, 12 stromal cells, 11 stem cells, 8 dendritic cells, 7 macrophages, 6 granulocytes, and 1 mast cell). Since the sample sizes of many cell types was extremely small, we only kept the top four cell types with enough samples (46 $\alpha\beta$ T cells, 33 B cells, 20 innate lymphocytes, and 13 $\gamma\delta$ T cells). The gene expression signatures were identified for such four major types of immunological cells.

2.2. Feature Selection. The MCFS [27] was first used to identify interpretable information about gene discrimination among the different groups of immune cells. Then, the two-stage IFS [28] method was applied to obtain genes with strong classification ability to improve the component recognition for the immune system.

2.2.1. Monte-Carlo Feature Selection. MCFS is a classic feature selection method for distinguishable features and ranks the features through guided sampling. For specific steps, multiple feature subsets with m features were arbitrarily chosen from the original M features ($m < < M$), the bootstrap dataset was trained for each specific feature subset, and the generated p decision trees were evaluated. The $p \times t$ decision trees were obtained by repeating the above steps t times. The accuracy weights of generated multiple decision trees provide a relative importance (RI) score for each feature, which was calculated as below.

$$\mathrm{RI}_f = \sum_{\tau=1}^{pt} (\mathrm{wAcc})^u \mathrm{IG}\left(n_f(\tau)\right) \left(\frac{\mathrm{no.in}\, n_f(\tau)}{\mathrm{no.in}\, \tau}\right)^v, \qquad (1)$$

where wAcc is the weighted accuracy and $n_f(\tau)$ is a node of feature f in decision tree τ. The information gain of $n_f(\tau)$ is expressed as $\mathrm{IG}(n_f(\tau))$, $\mathrm{no.in}\, n_f(\tau)$ is the number of training samples in $n_f(\tau)$, and u and v are the two weighting factors. Here, we adopted the MCFS program obtained from http://www.ipipan.eu/staff/m.draminski/mcfs.html. For convenience, default parameters were used.

After obtaining the RI scores of all features, we ranked all features in a list with the decreasing order of their RI scores. In addition to the feature list, the MCFS method also yields some most important features, called informative features in this study, which are some top features in the list. These features are accessed by a permutation test on class labels and one-sided Student's t-test.

2.2.2. Two-Stage Incremental Feature Selection. IFS is a feature selection method that accurately distinguishes samples from different classes by screening a set of optimal features. The features in the ranked list from MCFS can be sorted in a descending order according to their RI scores as mentioned above. Such feature list can help the classification algorithm in producing optimal performance. The original IFS must test all possible feature subsets, which are constructed from the feature list, to filter out the optimal feature subset that can identify samples' classes with best performance. Here, due to the large number of features (~50000), inducing lots of time to test all feature subsets, we designed a two-stage IFS method. In the first stage, we constructed candidate feature subsets with a large step size. Taking 10-step size as an example, in N feature subsets $F = [F_1^1, F_2^1, \cdots, F_N^1]$, the i-th feature subset contains $i \times 10$ high-ranked features, denoted as $F_i^1 = [f_1, f_2, \cdots, f_{i\times10}]$. Classifiers on samples with each feature subset were learned, which were further tested with 10-fold cross-validation [32–37]. Then, the feature interval containing the feature subset with the highest performance was determined and denoted as [min, max]. In the second stage, a series of feature subsets which contained top min, min + 1,…, max-1, max features were constructed. Likewise, a classifier for each feature subset was built and evaluated by 10-fold cross-validation. Accordingly, the classifier with the best performance can be found and termed as the optimal classifier. The corresponding feature subset is defined as the optimal feature subset. In this study, we selected RF to construct classifiers.

2.3. Random Forest. RF [29] is a classic machine learning algorithm, which contains a large number of decision tree classifiers, that is RF is an assemble classification algorithm. It is widely used in computational biology as one of the most common machine learning methods [38–43]. The output sample class/category of RF is determined by these tree classifiers (i.e., decision trees) in an aggregating vote manner. A RF consists of multiple decision trees with subtle differences. Thus, the mean of the predictions of all decision trees is usually taken as the final consensus results. Although this approach can lead to interpretability loss and slight increase in the model bias, it can avoid overfitting and improve the performance robustness. To quickly implement RF, the tool "RandomForest" in Weka [44, 45] was employed. Such a tool was executed with its default parameters.

2.4. Rule Learning Scheme. The IFS method with RF is helpful to construct a powerful classifier. However, such a classifier is absolutely a black-box classifier. It is very hard to capture the classification principle from such a classifier. Thus, we further employed a rule learning scheme to extract classification rules from the cell expression data.

To save time, we directly used the informative features yielded by the MCFS method. These features were first processed by the Johnson reducer algorithm [46, 47]. Some nonessential features were discarded, and the remaining features had the similar classification ability to the original informative features. Then, the repeated incremental pruning to produce the error reduction (RIPPER) algorithm [48] was applied on the remaining features to construct rules. The RIPPER algorithm is a specific method for constructing rule-based classifiers. The main frame of the RIPPER algorithm is based on IF-ELSE rules and consists of two parts: rule generation and rule optimization. The rule generation is a two-layer loop: the outer loop generates a rule each time after pruning and adds it to the rule pool, and the inner loop adds a predecessor to the rule each time. The rule optimization constructs alternatives based on the rules in the pool and finally selects the optimal rule to update the rule pool.

The above procedures are implemented and integrated in the MCFS program downloaded from http://www.ipipan.eu/staff/m.draminski/mcfs.html. We directly used it to produce rules.

2.5. Performance Measurement. The Matthew Correlation Coefficient (MCC) [49–57] is a common method used to evaluate the performance of dissimilar classifiers. This variable correlation coefficient calculates the correlation between the target and prediction classes with return value between -1 and +1. MCC considers true and false positives and negatives and is generally considered as a balanced measurement, even when the sample categories have different sizes. In this study, MCC within 10-fold cross-validation was used to evaluate classification performance.

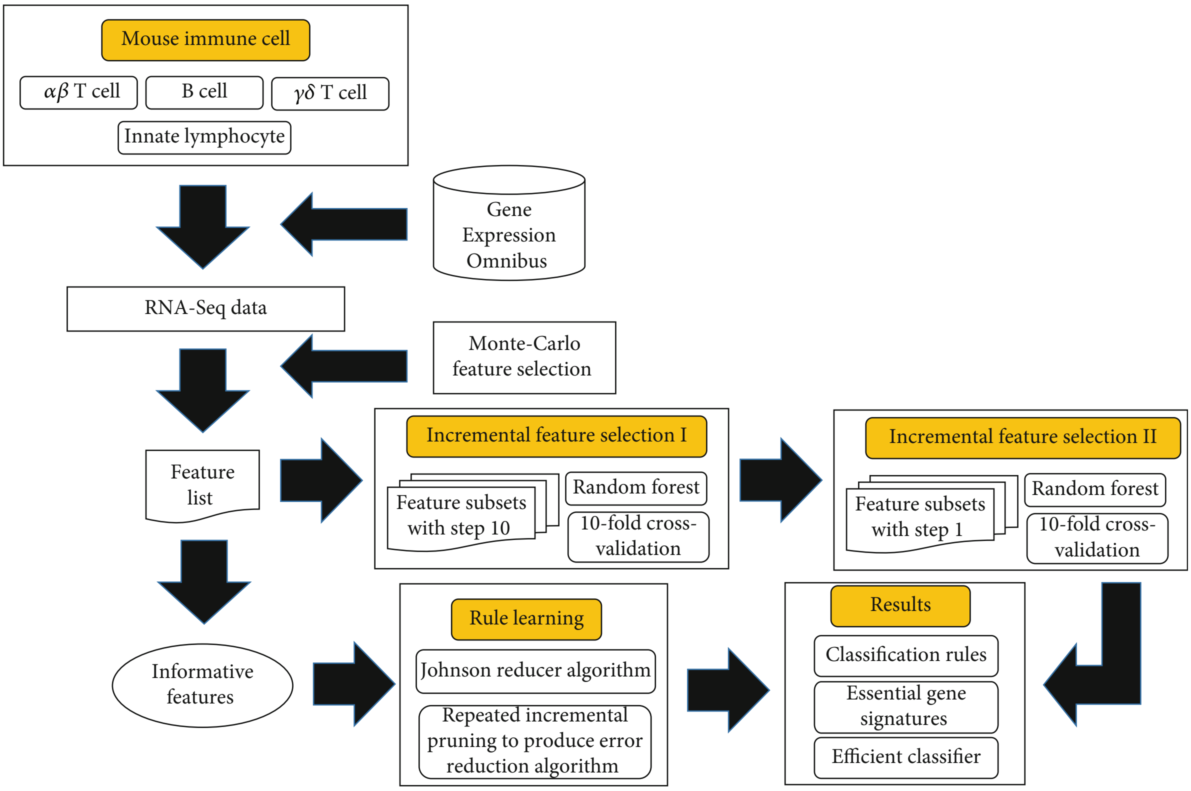


Figure 1: Entire procedures of the computational analysis on RNA-Seq data of mouse immunological cells. The data is retrieved from the Gene Expression Omnibus and is analyzed by the Monte-Carlo Feature Selection method. One feature list and some informative features are produced. A two-stage incremental feature selection method with random forest as the classification algorithm was applied on the feature list to extract essential gene signatures and one efficient classifier. Furthermore, a rule learning scheme is executed on the informative features for constructing classification rules.

Besides, we also employed accuracy on each cell type and overall accuracy (ACC) to fully evaluate the performance of different classifiers.

3. Results

In this study, we adopted several computational methods to analyze the RNA-Seq data of mouse immunological cells. The entire procedures are illustrated in Figure 1. The purpose was to extract essential gene signatures and rules for different immunological cell types. This section gives detailed results of each step of the procedures.

3.1. Results of the MCFS Method. The RNA-Seq data was first analyzed by the MCFS method. Accordingly, each feature was assigned a RI score. Then, a feature list was constructed with the decreasing order of their RI scores, which are provided in Table S1. Moreover, some informative features were also yielded by the MCFS method, which were the top 84 features in the list provided in Table S1.

3.2. Results of the IFS with Random Forest. A two-stage IFS method, incorporating RF as the classification algorithm, was applied to the feature list. In the first stage, we ran IFS with a step size of 10 on the feature list from MCFS. A RF classifier was built based on each constructed feature subset. Then, all classifiers were assessed by 10-fold cross-validation. The predicted results were counted as MCCs, accuracies on four types, and ACCs, which are available in Table S2. For an easy observation, we plotted the obtained MCCs on a coordinate system with the number of used features as the x-axis, as shown in Figure 2(a). It can be seen that when the top ten features were adopted, the RF classifier gave a perfection prediction with MCC = 1. Thus, we determined the min = 1 and max = 50 to do the second IFS stage. Feature subsets containing the top 1-50 features were built, on each of which a RF classifier was set up. Each classifier was evaluated by 10-fold cross-validation. The predicted results are provided in Table S3. Figure 2(b) listed the performance of the RF classifier based on the top ten feature subsets. The RF with the top six features yielded the perfect classification. Thus, such RF classifier was called the optimal classifier, and the corresponding feature subset was termed as the optimal feature subset.

3.3. Results of the Rule Learning. Besides the RF black-box classifier, we also used a rule learning scheme to give a clearer description on the classification procedure, thereby evidently elaborating the differences on four immunological cell types.

According to the MCFS results, 84 informative features were obtained (see the first 84 features in Table S1). Then, the Johnson reducer algorithm was applied on these features to further select the most essential features. The RIPPER algorithm followed to extract rules with the

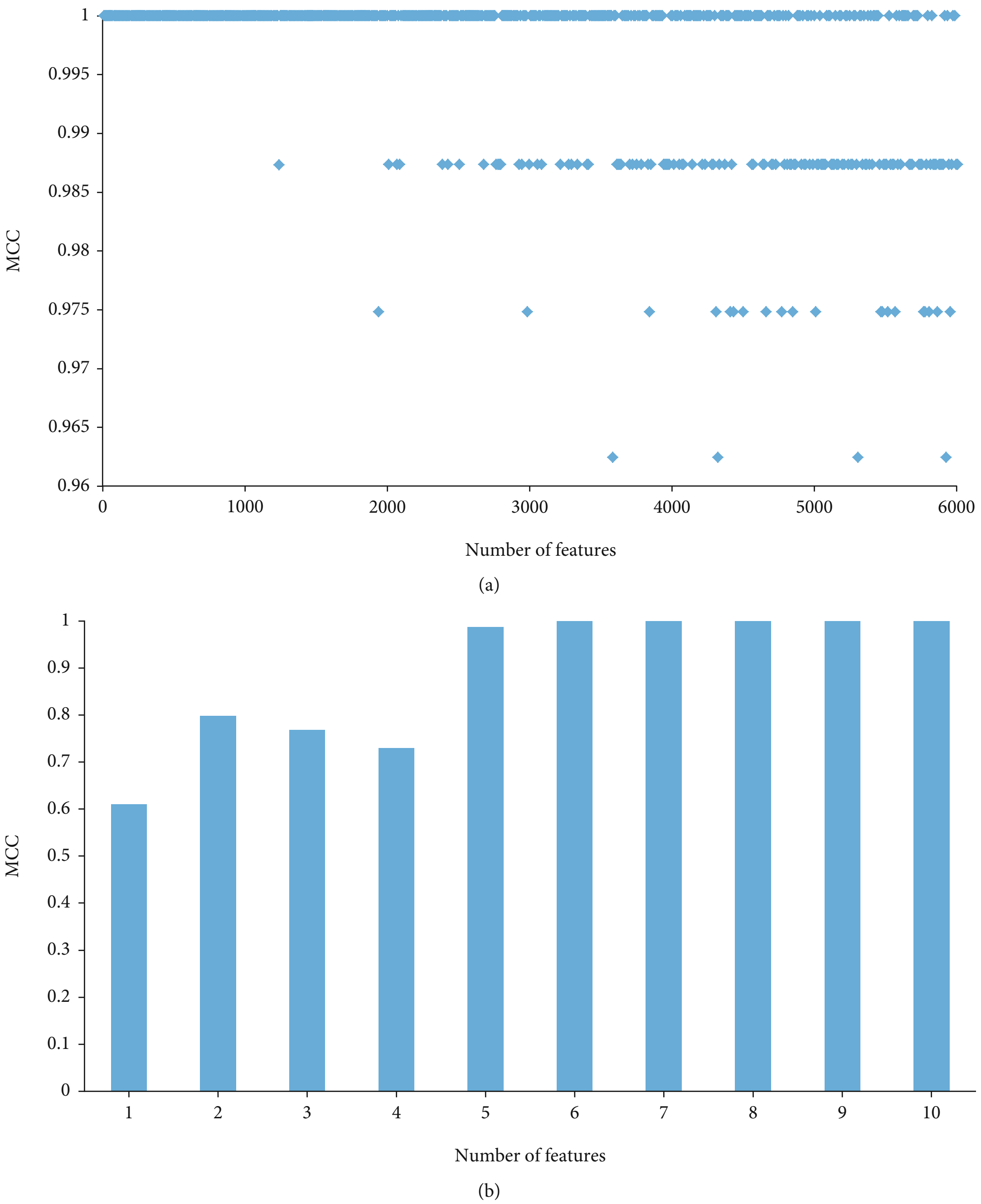


Figure 2: Performance curve of incremental feature selection (IFS) with random forest (RF): (a) performance of RF classifiers with different numbers of features in the first stage of IFS method; (b) performance of RF classifiers with different numbers of features in the second stage of IFS method.

remaining features, resulting in four rules, which are listed in Table 1. To indicate the utility of these rules, two measurements, support and accuracy, were calculated for each rule, which are also listed in Table 1. It can be seen that each rule can cover several immunological cells, and the efficiency of each rule was quite high.

Furthermore, to elaborate the utility of the procedures for constructing the rules, we did the 10-fold cross-validation three times. The accuracies on four cell types are shown in Figure 3. Except the accuracy on the $\gamma\delta$ T cell (82.05%), other accuracies were all no less than 90%. The ACC was 93.15% and MCC was 0.903. All these indicated that such a rule learning scheme was quite effective to extract efficient rules, also indicating the reliability of the rules in Table 1.

4. Discussion

We analyzed the following four typical cell subtypes in the mouse immune system: $\alpha\beta$ T cells, B cells, $\gamma\delta$ T cells, and innate lymphocytes, to screen detailed immune genes and establish standards for cell subgrouping. Basing on the gene expression profiling of individual cells, we performed

Table 1: Classification rules from RIPPER.

Rules	Criteria	Patients	Support[a]	Accuracy[b]
Rule 1	Tcrg-V4 ≥ 62.7560	$\gamma\delta$ T cells	10.71%	100%
Rule 2	Aifm2 ≥ 14.9503	Innate lymphocytes	17.86%	95.00%
Rule 3	Abcb9 ≤ 10.6151	B cells	29.46%	96.97%
Rule 4	Others	$\alpha\beta$ T cells	43.75%	93.88%

[a]Support is defined as the proportion of immunological cells satisfying the rule. [b]Accuracy is defined as the proportion of correctly predicted immunological cells among the cells satisfying the rules.

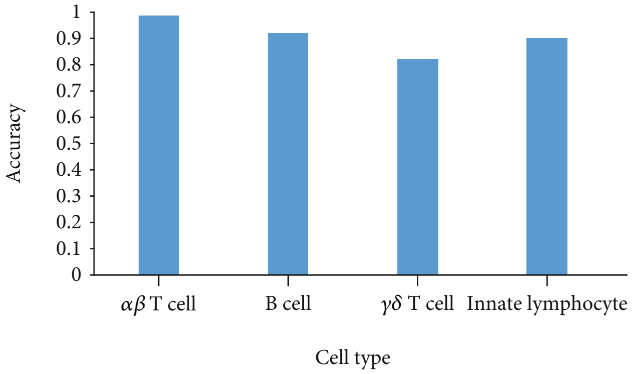


Figure 3: Performance of the rule learning scheme with 10-fold cross-validation three times. The accuracy on each cell type is quite high.

qualitative prediction on cell subtypes, identification of candidate immune cell-associated genes (noted as ImmGen-associated genes), and quantitative screening for the detailed recognition criteria of each cell subtype in a rule manner. According to recent publications, all identified ImmGen-associated genes and quantitative rules can be supported and confirmed by existing experiments and analysis, thus validating the efficacy and accuracy of our prediction. The detailed analysis of high-ranked ImmGen-associated genes and corresponding quantitative rules can be seen below.

4.1. Cell Type-Specific Function of ImmGen-Associated Genes.

With the MCFS method, we ranked features (genes) in a list (Table S1). Here, we selected the top ten genes, which are listed in Table 2, for detailed analysis.

The top gene in the ranked feature list is *Ighv1-72*, which encodes the variable region in the heavy chain of immunoglobulin [58]. According to recent publications, this gene participates in antigen-responding antibody synthesis [58]. All biological processes involving antibody synthesis mostly occur in one of our candidate cell subtypes, i.e., B cells, but not in the other cell subtypes [59–61]. Therefore, the expression pattern of *Ighv1-72* in B cells may be different from those in the other three cell subtypes. This finding validates the potential distinguishing role of *Ighv1-72*.

The next high-ranked gene is *Cd5*, which encodes a famous cluster of differentiation. In the mouse immune system, Cd5 is a T-cell surface glycoprotein that regulates T cell inhibition [76, 77]. According to recent publications, *Cd5* is expressed in $\alpha\beta$ T cells and $\gamma\delta$ T cells and is a potential biomarker for T cell subgroups [62, 63]. Although *Cd5* has a specific role of encoding protein T cell surface glycoprotein, its protein products are found on the surface of a specific subgroup of B cells [64, 65]. This finding implies that this biomarker may distinguish innate lymphocytes from the other three immune cell subtypes.

The next predicted gene is *Klrb1b*, which encodes a specific lectin-like receptor on the surface of natural killer cells. Our predicted gene *Nlrb1b* (*Klrb1b*) encodes a functional subunit of a receptor-ligand system in NK cells and may regulate an MHC-independent immune surveillance mechanism [66]. For its cell subtype specific expression pattern, *Nlrb1b* plays an irreplaceable role in natural killer cells and T cells [66], i.e., a key subtype of innate lymphocyte. Hence, Nlrb1b may be a potential marker for innate lymphocytes.

Phka1 exhibits a differential expression pattern among the different cell subtypes. Encoding an alpha chain of the phosphorylase kinase, this gene has a differential expression pattern in different T cells, B cells, and innate lymphocyte subtypes and even under different cell activation status [67–69]. Therefore, *Phka1* is definitely a potential biomarker for the distinction of four immune cell subtypes due to its substantially biological functions and alternative expression patterns in different immune cell subtypes under various immune conditions.

Trdv5, *Trbj1-2*, *Trbj1-3*, *Tcrg-V4*, and *Trbj1-7* are the feature genes encoding different regions of the T cell receptor (TCR) [78, 79]. The differential expression pattern of these genes in four cell subtypes indicates or reflects that of T cell receptors in different cell subtypes. $\alpha\beta$ T and $\gamma\delta$ T cells have a high expression of T cell receptors [80, 81]. These two cell subtypes can be further distinguished according to feature genes due to the differential expression pattern of Tcrg-V4, which encodes a unique region of the gamma chain [82]. A high expression pattern of Tcrg-V4 can be found in $\gamma\delta$ T cells but not in $\alpha\beta$ T cells, thus confirming the distinguishing capacity of our predicted gene signatures [72, 73]. For the remaining cell subtypes, B cells and innate lymphocytes, the former does not have the expression of all TCR-associated genes. A specific subtype of innate lymphocyte, namely, natural killer T cells, has a unique expression pattern of T cell receptors [70, 71]. All identified natural killer T cells with T cell receptor expression would also have the specific $\alpha\beta$ T cell receptors but not the $\gamma\delta$ T cell receptors [71]. Therefore, some subgroups of innate lymphocytes may also have alternative expression patterns of Trdv5, Trbj1-2, Trbj1-3, and Trbj1-7 but not Tcrg-V4. This finding reflects the distinguishing effects of our predicted ImmGen-associated genes involved in T cell receptors.

Table 2: Information of top ten genes selected by Monte-Carlo Feature Selection method.

Rank	Gene	RI score	Cell types	Reference
1	*Ighv1-72*	0.7660	B cells	[58–61]
2	*Cd5*	0.5902	$\alpha\beta$ T cells/$\gamma\delta$ T cells and B cells (with another specific pattern)	[62–65] (for B cells)
3	*Klrb1b*	0.5720	Innate lymphocytes	[66]
4	*Phka1*	0.5535	$\alpha\beta$ T cells, $\gamma\delta$ T cells, B cells, and innate lymphocytes (with different expression level)	[67–69]
5	*Trdv5*	0.5223	$\alpha\beta$ T cells/$\gamma\delta$ T cells and innate lymphocytes (with another specific pattern)	[70, 71]
6	*Trbj1-2*	0.4789		
7	*Trbj1-3*	0.4752		
9	*Trbj1-7*	0.4454		
8	*Tcrg-V4*	0.4738	$\gamma\delta$ T cells	[72, 73]
10	*EBF1*	0.4403	B cells	[74, 75]

Various T cell receptor-associated genes can still be found in the top-ranked genes of our feature gene list, thereby implying the unique differential capacity of the T cell receptor expression pattern and validating the efficacy and accuracy of our prediction approach. In addition to T cell receptor coding genes, we also identified a unique B cell recognizing gene named *EBF1*. Acting as a transcription factor, this gene contributes to the maintenance of B cell identity and prevention of alternative fates in committed cells, such as transferring to the T cell lineage [74, 75]. Therefore, the high expression pattern of EBF1 can only be identified in B cells, implying its potential as a biomarker for this cell type.

Owing to the limitation of the article length, we cannot individually analyze the discriminative genes. However, the above-mentioned high-ranked genes have cell type-specific expression patterns in immune cell subtypes, thus validating the efficacy and accuracy of our prediction and analysis.

4.2. Cell Type-Specific Expression Pattern of ImmGen-Associated Rules. Besides the gene signatures, we also obtained some classification rules via a rule learning scheme (Table 1). They were analyzed as follows. The analysis was based on the expression level measured by Fragments Per Kilobase of transcript per Million mapped reads (FPKM).

The first identified quantitative parameter is *Tcrg-V4*. As a specific encoding gene for the $\gamma\delta$ T cell receptors, this gene may have high expression in $\gamma\delta$ T cells. In accordance with our predicted expression rules, the expression abundance of *Tcrg-V4* is higher than 62.755978 (FPKM) [72, 73]. According to the Mouse Genome Informatics database [83], the expression level and relative expression quantity of *Tcrg-V4* in the T cell subgroup, $\gamma\delta$ T cells, basically conforms to our predicted rules [84, 85], thereby validating the efficacy and accuracy of our prediction.

The next classification rule of quantitative parameter involves a specific gene named *Aifm2*, which contributes to the identification of innate lymphocytes. According to the Mouse Genome Informatics database [83], this gene may have a unique higher expression pattern in mucosal tissues, which are full of innate immune cells, than in the blood system and lymphoid node, which are full of T cells and B cells [86, 87]. Therefore, *Aifm2* may be highly expressed in innate lymphocytes with a threshold of approximately 15 FPKM.

Another quantitative parameter is *Abcb9*, whose low expression (lower than 10 FPKM) is indicative of B cells rather than T cells or innate lymphocytes. Mucosal tissues are full of innate immune cells, and thymus tissues are full of T cells. By contrast, the anatomic area, the spleen, is full of mature B cells and thus has low expression level of *Abcb9* (<5 FPKM) [83].

The cells that do not follow the three rules mentioned above may be $\alpha\beta$ T cells. Thus, all typical expression patterns can be set up for corresponding immune cell subtypes and have been confirmed by recent studies, thereby validating the efficacy and accuracy of our analysis.

5. Conclusions

By using our newly presented computational approach, we identified a group of signature genes that are qualitatively related to the biological processes of four immune cell subtypes. We also set up a set of quantitative expression rules for the detailed clustering of immune cells based on the absolute expression levels measured by FPKM. This work provides a novel computational tool for the quantitative subtyping of immune cells and biomarker recognition.

Data Availability

The data used to support the findings of this study have been deposited in the Gene Expression Omnibus repository (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE109125).

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Authors' Contributions

Yu-Hang Zhang and Zhandong Li contributed equally to this work.

Acknowledgments

This work was supported by the Shanghai Municipal Science and Technology Major Project [2017SHZDZX01], the National Key R&D Program of China [2018YFC0910403, 2017YFC1201200], the National Natural Science Foundation of China [31701151], the Shanghai Sailing Program [16YF1413800], the Youth Innovation Promotion Association of Chinese Academy of Sciences (CAS) [2016245], the Strategic Leading Science & Technology Programme of the Chinese Academy of Sciences (XDB38050200), and the Fund of the Key Laboratory of Tissue Microenvironment and Tumor of Chinese Academy of Sciences [202002].

Supplementary Materials

Supplementary 1. Table S1: ranked features by MCFS

Supplementary 2. Table S2: performance of IFS with RF when using different numbers of features ranked by MCFS with step size of 10

Supplementary 3. Table S3: performance of IFS with RF when using different numbers of features ranked by MCFS with step size of 1.

References

[1] M. Hasegawa and N. Inohara, "Regulation of the gut microbiota by the mucosal immune system in mice," *International Immunology*, vol. 26, no. 9, pp. 481–487, 2014.

[2] S. C. Latet, V. Y. Hoymans, P. L. van Herck, and C. J. Vrints, "The cellular immune system in the post-myocardial infarction repair process," *International Journal of Cardiology*, vol. 179, pp. 240–247, 2015.

[3] J. Parkin and B. Cohen, "An overview of the immune system," *Lancet*, vol. 357, no. 9270, pp. 1777–1789, 2001.

[4] N. Labrecque and N. Cermakian, "Circadian clocks in the immune system," *Journal of Biological Rhythms*, vol. 30, no. 4, pp. 277–290, 2015.

[5] S. F. Martin, "Adaptation in the innate immune system and heterologous innate immunity," *Cellular and Molecular Life Sciences*, vol. 71, no. 21, pp. 4115–4130, 2014.

[6] O. Osborn and J. M. Olefsky, "The cellular and signaling networks linking the immune system and metabolism in disease," *Nature Medicine*, vol. 18, no. 3, pp. 363–374, 2012.

[7] J. Goverman, T. Hunkapiller, and L. Hood, "A speculative view of the multicomponent nature of T cell antigen recognition," *Cell*, vol. 45, no. 4, pp. 475–484, 1986.

[8] T. A. Springer, "Adhesion receptors of the immune system," *Nature*, vol. 346, no. 6283, pp. 425–434, 1990.

[9] T. A. Springer, M. L. Dustin, T. K. Kishimoto, and S. D. Marlin, "The lymphocyte function-associated LFA-1, CD2, and LFA-3 molecules: cell adhesion receptors of the immune system," *Annual Review of Immunology*, vol. 5, no. 1, pp. 223–252, 1987.

[10] M. Guerau-de-Arellano, H. Alder, H. G. Ozer, A. Lovett-Racke, and M. K. Racke, "miRNA profiling for biomarker discovery in multiple sclerosis: from microarray to deep sequencing," *Journal of Neuroimmunology*, vol. 248, no. 1-2, pp. 32–39, 2012.

[11] A. A. Alizadeh and L. M. Staudt, "Genomic-scale gene expression profiling of normal and malignant immune cells," *Current Opinion in Immunology*, vol. 12, no. 2, pp. 219–225, 2000.

[12] I. San Segundo-Val and C. S. Sanz-Lozano, "Introduction to the gene expression analysis," *Methods in Molecular Biology*, vol. 1434, pp. 29–43, 2016.

[13] C. Pilarsky, L. K. Nanduri, and J. Roy, "Gene expression analysis in the age of mass sequencing: an introduction," *Methods in Molecular Biology*, vol. 1381, pp. 67–73, 2016.

[14] C. C. Kim and L. L. Lanier, "Beyond the transcriptome: completion of act one of the Immunological Genome Project," *Current Opinion in Immunology*, vol. 25, no. 5, pp. 593–597, 2013.

[15] The Immunological Genome Project Consortium, T. S. P. Heng, M. W. Painter et al., "The Immunological Genome Project: networks of gene expression in immune cells," *Nature Immunology*, vol. 9, no. 10, pp. 1091–1094, 2008.

[16] E. Clough and T. Barrett, "The Gene Expression Omnibus database," *Methods in Molecular Biology*, vol. 1418, pp. 93–110, 2016.

[17] C. D. Castro, C. T. Boughter, A. E. Broughton, A. Ramesh, and E. J. Adams, "Diversity in recognition and function of human $\gamma\delta$ T cells," *Immunological Reviews*, vol. 298, no. 1, pp. 134–152, 2020.

[18] D. Nemazee, "Mechanisms of central tolerance for B cells," *Nature Reviews Immunology*, vol. 17, no. 5, pp. 281–294, 2017.

[19] R. M. Rezende, A. J. Lanser, S. Rubino et al., "$\gamma\delta$ T cells control humoral immune response by inducing T follicular helper cell differentiation," *Nature Communications*, vol. 9, no. 1, pp. 1–13, 2018.

[20] K. Neumann, K. Karimi, J. Meiners et al., "A proinflammatory role of type 2 innate lymphoid cells in murine immune-mediated hepatitis," *The Journal of Immunology*, vol. 198, no. 1, pp. 128–137, 2017.

[21] M. S. Cruz, A. Diamond, A. Russell, and J. M. Jameson, "Human $\alpha\beta$ and $\gamma\delta$ T cells in skin immunity and disease," *Frontiers in Immunology*, vol. 9, p. 1304, 2018.

[22] B. Silva-Santos, S. Mensurado, and S. B. Coffelt, "$\gamma\delta$ T cells: pleiotropic immune effectors with therapeutic potential in cancer," *Nature Reviews Cancer*, vol. 19, no. 7, pp. 392–404, 2019.

[23] A. C. Kohlgruber, S. T. Gal-Oz, N. M. LaMarche et al., "$\gamma\delta$ T cells producing interleukin-17A regulate adipose regulatory T cell homeostasis and thermogenesis," *Nature Immunology*, vol. 19, no. 5, pp. 464–474, 2018.

[24] D. Liang, H. Shao, W. K. Born, R. L. O'Brien, H. J. Kaplan, and D. Sun, "High level expression of A2ARs is required for the enhancing function, but not for the inhibiting function, of $\gamma\delta$ T cells in the autoimmune responses of EAU," *PLoS One*, vol. 13, no. 6, article e0199601, 2018.

[25] S. Garaud, L. Buisseret, C. Solinas et al., "Tumor-infiltrating B cells signal functional humoral immune responses in breast cancer," *JCI insight*, vol. 5, no. 18, article e129641, 2019.

[26] E. R. Kansler and M. O. Li, "Innate lymphocytes—lineage, localization and timing of differentiation," *Cellular & Molecular Immunology*, vol. 16, no. 7, pp. 627–633, 2019.

[27] M. Draminski, A. Rada-Iglesias, S. Enroth, C. Wadelius, J. Koronacki, and J. Komorowski, "Monte Carlo feature selection for supervised classification," *Bioinformatics*, vol. 24, no. 1, pp. 110–117, 2008.

[28] H. A. Liu and R. Setiono, "Incremental feature selection," *Applied Intelligence*, vol. 9, no. 3, pp. 217–230, 1998.

[29] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[30] H. Yoshida, C. A. Lareau, R. N. Ramirez et al., “The cis-regulatory atlas of the mouse immune system,” *Cell*, vol. 176, no. 4, pp. 897–912.e20, 2019, e20.

[31] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, “edgeR: a bioconductor package for differential expression analysis of digital gene expression data,” *Bioinformatics*, vol. 26, no. 1, pp. 139-140, 2010.

[32] R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” in *International joint Conference on artificial intelligence*, Lawrence Erlbaum Associates Ltd, San Francisco, CA, USA, 1995.

[33] J.-P. Zhou, L. Chen, and Z.-H. Guo, “iATC-NRAKEL: an efficient multi-label classifier for recognizing anatomical therapeutic chemical classes of drugs,” *Bioinformatics*, vol. 36, no. 5, pp. 1391–1396, 2020.

[34] J. Che, L. Chen, Z. H. Guo, S. Wang, and Aorigele, “Drug target group prediction with multiple drug networks,” *Combinatorial Chemistry & High Throughput Screening*, vol. 23, no. 4, pp. 274–284, 2020.

[35] Y. Zhu, B. Hu, L. Chen, and Q. Dai, “iMPTCE-Hnetwork: A Multilabel Classifier for Identifying Metabolic Pathway Types of Chemicals and Enzymes with a Heterogeneous Network,” *Computational and Mathematical Methods in Medicine*, vol. 2021, Article ID 6683051, 12 pages, 2021.

[36] H. Liu, B. Hu, L. Chen, and L. Lu, “Identifying protein subcellular location with embedding features learned from networks,” *Current Proteomics*, vol. 17, 2020.

[37] S. Wang, Q. Zhang, J. Lu, and Y. D. Cai, “Analysis and prediction of nitrated tyrosine sites with the mRMR method and support vector machine algorithm,” *Current Bioinformatics*, vol. 13, no. 1, pp. 3–13, 2018.

[38] X. Zhao, L. Chen, and J. Lu, “A similarity-based method for prediction of drug side effects with heterogeneous information,” *Mathematical Biosciences*, vol. 306, pp. 136–144, 2018.

[39] X. Zhao, L. Chen, Z. H. Guo, and T. Liu, “Predicting drug side effects with compact integration of heterogeneous networks,” *Current Bioinformatics*, vol. 14, no. 8, pp. 709–720, 2019.

[40] H. Liang, L. Chen, X. Zhao, and X. Zhang, “Prediction of drug side effects with a refined negative sample selection strategy,” *Computational and Mathematical Methods in Medicine*, vol. 2020, Article ID 1573543, 16 pages, 2020.

[41] X. Zhang, L. Chen, Z. H. Guo, and H. Liang, “Identification of human membrane protein types by incorporating network embedding methods,” *IEEE Access*, vol. 7, pp. 140794–140805, 2019.

[42] Y. Jia, R. Zhao, and L. Chen, “Similarity-based machine learning model for predicting the metabolic pathways of compounds,” *IEEE Access*, vol. 8, pp. 130687–130696, 2020.

[43] J. Li, L. Lu, Y. H. Zhang et al., “Identification of synthetic lethality based on a functional network by using machine learning algorithms,” *Journal of Cellular Biochemistry*, vol. 120, no. 1, pp. 405–416, 2019.

[44] E. Frank, M. Hall, L. Trigg, G. Holmes, and I. H. Witten, “Data mining in bioinformatics using Weka,” *Bioinformatics*, vol. 20, no. 15, pp. 2479–2481, 2004.

[45] I. H. Witten and E. Frank, *Data mining: practical machine learning tools and techniques*, Kaufmann, San Francisco, Morgan, 2nd edition, 2005.

[46] D. S. Johnson, “Approximation algorithms for combinatorial problems,” *Journal of Computer and System Sciences*, vol. 9, no. 3, pp. 256–278, 1974.

[47] A. Ohrn, *Discernibility and rough sets in medicine: tools and applications, in Department of Computer and Information Science*, Norwegian University of Science and Technology, Trondheim, 1999.

[48] W. W. Cohen, “Fast effective rule induction,” in *The Proceeding of Proceeding of the Twelfth International. Conference of Machine Learning*, pp. 115–123, Tahoe City, CA, USA, July 9–12, 1995.

[49] B. Matthews, “Comparison of the predicted and observed secondary structure of T4 phage lysozyme,” *Biochimica et Biophysica Acta (BBA)-Protein Structure*, vol. 405, no. 2, pp. 442–451, 1975.

[50] H. Cui and L. Chen, “A binary classifier for the prediction of EC numbers of enzymes,” *Current Proteomics*, vol. 16, no. 5, pp. 381–389, 2019.

[51] L. Chen, C. Chu, Y. H. Zhang et al., “Identification of drug-drug interactions using chemical interactions,” *Current Bioinformatics*, vol. 12, no. 6, pp. 526–534, 2017.

[52] L. Chen, S. Wang, Y. H. Zhang et al., “Identify key sequence features to improve CRISPR sgRNA efficacy,” *IEEE Access*, vol. 5, pp. 26582–26590, 2017.

[53] Y.-H. Zhang, H. Li, T. Zeng et al., “Identifying transcriptomic signatures and rules for SARS-CoV-2 infection,” *Frontiers in Cell and Developmental Biology*, vol. 8, p. 627302, 2021.

[54] Y.-H. Zhang, Z. Li, T. Zeng et al., “Detecting the multiomics signatures of factor-specific inflammatory effects on airway smooth muscles,” *Frontiers in Genetics*, vol. 11, p. 599970, 2021.

[55] X. Pan, H. Li, T. Zeng et al., “Identification of protein subcellular localization with network and functional embeddings,” *Frontiers in Genetics*, vol. 11, p. 626500, 2021.

[56] Y.-H. Zhang, Z. Li, T. Zeng et al., “Distinguishing glioblastoma subtypes by methylation signatures,” *Frontiers in Genetics*, vol. 11, p. 604336, 2020.

[57] L. Chen, Z. Li, T. Zeng et al., “Identifying robust microbiota signatures and interpretable rules to distinguish cancer subtypes,” *Frontiers in Molecular Biosciences*, vol. 7, p. 604794, 2020.

[58] N. Kono, L. Sun, H. Toh et al., “Deciphering antigen-responding antibody repertoires by using next-generation sequencing and confirming them through antibody-gene synthesis,” *Biochemical and Biophysical Research Communications*, vol. 487, no. 2, pp. 300–306, 2017.

[59] P. M. Glassman, L. Abuqayyas, and J. P. Balthasar, “Assessments of antibody biodistribution,” *Journal of Clinical Pharmacology*, vol. 55, Suppl 3, pp. S29–S38, 2015.

[60] J. W. Larrick, P. W. H. I. Parren, J. S. Huston et al., “Antibody engineering and therapeutics conference. The annual meeting of the antibody society, Huntington Beach, CA, December 7-11, 2014,” *MAbs*, vol. 6, no. 5, pp. 1115–1123, 2014.

[61] L. Presta, “Antibody engineering for therapeutics,” *Current Opinion in Structural Biology*, vol. 13, no. 4, pp. 519–525, 2003.

[62] H. Yokozeki, K. Watanabe, K. Igawa, Y. Miyazaki, I. Katayama, and K. Nishioka, “Gammadelta T cells assist alphabeta T cells in the adoptive transfer of contact hypersensitivity to para-phenylenediamine,” *Clinical and Experimental Immunology*, vol. 125, no. 3, pp. 351–359, 2001.

[63] T. Sugie, H. Kubota, M. Sato, E. Nakamura, M. Imamura, and N. Minato, "NK 1+ CD4- CD8- alphabeta T cells in the peritoneal cavity: specific T cell receptor-mediated cytotoxicity and selective IFN-gamma production against B cell leukemia and myeloma cells," *Journal of Immunology*, vol. 157, no. 9, pp. 3925–3935, 1996.

[64] V. L. Palmer, V. K. Nganga, M. E. Rothermund, G. A. Perry, and P. C. Swanson, "Cd1d regulates B cell development but not B cell accumulation and IL10 production in mice with pathologic CD5(+) B cell expansion," *BMC Immunology*, vol. 16, no. 1, p. 66, 2015.

[65] R. R. Hardy and K. Hayakawa, "Perspectives on fetal derived CD5+ B1 B cells," *European Journal of Immunology*, vol. 45, no. 11, pp. 2978–2984, 2015.

[66] Q. Zhang, M. M. A. Rahim, D. S. J. Allan et al., "Mouse Nkrp1-Clr gene cluster sequence and expression analyses reveal conservation of tissue-specific MHC-independent immunosurveillance," *PLoS One*, vol. 7, no. 12, article e50561, 2012.

[67] W. R. Osborne and C. R. Scott, "The metabolism of deoxyguanosine and guanosine in human B and T lymphoblasts. A role for deoxyguanosine kinase activity in the selective T-cell defect associated with purine nucleoside phosphorylase deficiency," *The Biochemical Journal*, vol. 214, no. 3, pp. 711–718, 1983.

[68] R. M. Goldblum, F. C. Schmalstieg, J. A. Nelson, and G. C. Mills, "Adenosine deaminase (ADA) and other enzyme abnormalities in immune deficiency states," *Birth Defects Original Article Series*, vol. 14, no. 6A, pp. 73–84, 1978.

[69] R. B. Trelease, R. A. Henderson, and J. B. Park, "A qualitative process system for modeling NF-kappaB and AP-1 gene regulation in immune cell biology research," *Artificial Intelligence in Medicine*, vol. 17, no. 3, pp. 303–321, 1999.

[70] A. Rodríguez-Caballero, A. C. García-Montero, P. Bárcena et al., "Expanded cells in monoclonal TCR-alphabeta+/CD4 +/NKa+/CD8-/+dim T-LGL lymphocytosis recognize hCMV antigens," *Blood*, vol. 112, no. 12, pp. 4609–4616, 2008.

[71] M. Verykokakis, M. D. Boos, A. Bendelac, E. J. Adams, P. Pereira, and B. L. Kee, "Inhibitor of DNA binding 3 limits development of murine slam-associated adaptor protein-dependent "innate" gammadelta T cells," *PLoS One*, vol. 5, no. 2, article e9303, 2010.

[72] T. Washburn, E. Schweighoffer, T. Gridley et al., "Notch activity influences the $\alpha\beta$ versus $\gamma\delta$ T cell lineage decision," *Cell*, vol. 88, no. 6, pp. 833–843, 1997.

[73] L. Riera-Sans and A. Behrens, "Regulation of alphabeta/gammadelta T cell development by the activator protein 1 transcription factor c-Jun," *Journal of Immunology*, vol. 178, no. 9, pp. 5690–5700, 2007.

[74] R. Nechanitzky, D. Akbas, S. Scherer et al., "Transcription factor EBF1 is essential for the maintenance of B cell identity and prevention of alternative fates in committed cells," *Nature Immunology*, vol. 14, no. 8, pp. 867–875, 2013.

[75] I. Gyory, S. Boller, R. Nechanitzky et al., "Transcription factor Ebf1 regulates differentiation stage-specific signaling, proliferation, and survival of B cells," *Genes & Development*, vol. 26, no. 7, pp. 668–682, 2012.

[76] M. Bamberger, A. M. Santos, C. M. Gonçalves et al., "A new pathway of CD5 glycoprotein-mediated T cell inhibition dependent on inhibitory phosphorylation of Fyn kinase," *The Journal of Biological Chemistry*, vol. 286, no. 35, pp. 30324–30336, 2011.

[77] W. Luo, H. Van de Velde, I. von Hoegen, J. R. Parnes, and K. Thielemans, "Ly-1 (CD5), a membrane glycoprotein of mouse T lymphocytes and a subset of B cells, is a natural ligand of the B cell surface protein Lyb-2 (CD72)," *Journal of Immunology*, vol. 148, no. 6, pp. 1630–1634, 1992.

[78] E. Ruggiero, J. P. Nicolay, R. Fronza et al., "High-resolution analysis of the human T-cell receptor repertoire," *Nature Communications*, vol. 6, no. 1, p. 8081, 2015.

[79] D. R. Thapa, R. Tonikian, C. Sun et al., "Longitudinal analysis of peripheral blood T cell receptor diversity in patients with systemic lupus erythematosus by next-generation sequencing," *Arthritis Research & Therapy*, vol. 17, no. 1, p. 132, 2015.

[80] K. Ohshima, K. Karube, R. Kawano et al., "Classification of distinct subtypes of peripheral T-cell lymphoma unspecified, identified by chemokine and chemokine receptor expression: analysis of prognosis," *International Journal of Oncology*, vol. 25, no. 3, pp. 605–613, 2004.

[81] B. M. Hall, "T cells: soldiers and spies–the surveillance and control of effector T cells by regulatory T cells," *Clinical Journal of the American Society of Nephrology*, vol. 10, no. 11, pp. 2050–2064, 2015.

[82] S. Huck, P. Dariavach, and M. P. Lefranc, "Variable region genes in the human T-cell rearranging gamma (TRG) locus: V-J junction and homology with the mouse genes," *The EMBO Journal*, vol. 7, no. 3, pp. 719–726, 1988.

[83] C. J. Bult, J. T. Eppig, J. A. Blake, J. A. Kadin, J. E. Richardson, and the Mouse Genome Database Group, "Mouse genome database 2016," *Nucleic Acids Research*, vol. 44, no. D1, pp. D840–D847, 2016.

[84] J. S. Heilig and S. Tonegawa, "Diversity of murine gamma genes and expression in fetal and adult T lymphocytes," *Nature*, vol. 322, no. 6082, pp. 836–840, 1986.

[85] C. Hetzer-Egger, M. Schorpp, A. Haas-Assenbaum, R. Balling, H. Peters, and T. Boehm, "Thymopoiesis requires Pax9 function in thymic epithelial cells," *European Journal of Immunology*, vol. 32, no. 4, pp. 1175–1181, 2002.

[86] G. Diez-Roux, S. Banfi, M. Sultan et al., "A high-resolution anatomical atlas of the transcriptome in the mouse embryo," *PLoS Biology*, vol. 9, no. 1, article e1000582, 2011.

[87] S. Magdaleno, P. Jensen, C. L. Brumwell et al., "BGEM: an in situ hybridization database of gene expression in the embryonic and adult mouse nervous system," *PLoS Biology*, vol. 4, no. 4, article e86, 2006.

Hindawi
BioMed Research International
Volume 2021, Article ID 5542224, 10 pages
https://doi.org/10.1155/2021/5542224

Review Article

Recent Advances in Predicting Protein S-Nitrosylation Sites

Qian Zhao,[1] Jiaqi Ma,[1] Fang Xie,[1] Yu Wang,[1] Yu Zhang,[1] Hui Li,[1] Yuan Sun,[2] Liqi Wang,[1] Mian Guo,[3] and Ke Han[1]

[1]*School of Computer and Information Engineering, Heilongjiang Provincial Key Laboratory of Electronic Commerce and Information Processing, Harbin University of Commerce, Harbin 150028, China*
[2]*School of Pharmacy, Center of Pharmaceutical Engineering and Technology, Harbin University of Commerce, Harbin 150076, China*
[3]*Department of Neurosurgery, The Second Affiliated Hospital of Harbin Medical University, Harbin 150086, China*

Correspondence should be addressed to Ke Han; thruster@163.com

Received 7 January 2021; Revised 24 January 2021; Accepted 25 January 2021; Published 11 February 2021

Academic Editor: Quan Zou

Protein S-nitrosylation (SNO) is a process of covalent modification of nitric oxide (NO) and its derivatives and cysteine residues. SNO plays an essential role in reversible posttranslational modifications of proteins. The accurate prediction of SNO sites is crucial in revealing a certain biological mechanism of NO regulation and related drug development. Identification of the sites of SNO in proteins is currently a very hot topic. In this review, we briefly summarize recent advances in computationally identifying SNO sites. The challenges and future perspectives for identifying SNO sites are also discussed. We anticipate that this review will provide insights into research on SNO site prediction.

1. Introduction

Protein S-nitrosylation (SNO) is one of the most important and common posttranslational modifications (PTMs), as shown in Figure 1, incorporating the covalent modification of nitric oxide (NO) and its derivatives and cysteine residues [1]. Numerous studies have shown that S-nitrosylation regulates multiple physiological and pathological processes, such as the immune response [2], cellular senescence [3], transcription, and posttranslational regulation [4]. In addition, abnormalities in protein S-nitrosylation and other posttranslational modifications can also lead to many diseases, such as Alzheimer's disease [5–7] and breast cancer [8]. In recent years, through molecular recognition and labelling of SNO sites in proteins, many large-scale proteomics experimental screenings have been completed, and the number of SNO proteins verified by experiments is also increasing [9, 10]. As to other protein posttranslational modification sites [11–18], the predicted SNO sites are time-wasting, strenuous, and extortionate through large-scale experimental screening methods. With continuous breakthroughs in sequence and structural biology, computational biology using machine learning has become an indispensable part of drug development [19–36].

As an alternative to biochemical experiments, identifying SNO sites in biological sequences with the least cost and efficiency in recent years is a focus of current research. To help researchers understand the development of this field, this review will use Chou's five-step rule as the literature selection criteria [37]: (1) how to select or construct an effective fiducial marker dataset subcellular location to train and test predictors, (2) how to express the sample with an effective formula that can truly reflect the intrinsic correlation between the sample and predicted target, (3) how to introduce or develop powerful algorithms to make predictions, (4) how to correctly conduct cross-validation tests to objectively evaluate the expected prediction accuracy, and (5) how to build a user-friendly web server for forecasters. In addition, to help researchers overcome the overall development of this field, this review briefly introduces early research on the identification of SNO sites using biochemical methods.

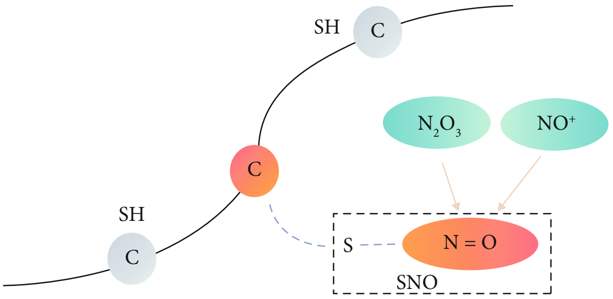


Figure 1: A schematic diagram of protein S-nitrosylation sites. Protein fragments have many residues, of which C (cysteine) is depicted as a circle. When NO and cysteine residues are covalently modified, SNO is formed, which is represented by a warm color, and the rest is gray.

2. Materials

High-quality datasets are the cornerstone of scientific research [38, 39]. With the development of proteomics and the advancement of research by scientists, the number of experimentally identified SNO sites is also increasing. In the process of predicting S-nitrosylation sites, the dynamic changes of the database and the dataset are sorted in this part.

2.1. Database. UniProt [40] (Universal Protein Resource) is a high-quality, extensive, and open-access database of protein sequences and functional annotations created and maintained by the UniProt Consortium, namely, EBI, SIB (Swiss Institute of Bioinformatics), and PIR (Protein Information Resource), an association of three institutions. It mainly includes three parts: the UniProtKB knowledge base, UniParc archive library, and UniRef reference sequence set. The UniProt database collects cysteine SNO sites from different species. With the continuous addition of a large number of experimentally verified SNO sites, the dataset used by scientists to predict SNO sites is also updated accordingly [41, 42].

dbSNO [43] (database of cysteine S-nitrosylation) is the first database specifically designed to integrate experimentally determined SNO sites and their structure or function information. SNO peptide sequences collected from different sources are heterogeneous, so dbSNO maps the identity of these sequences to UniProtKB protein entry. In addition, the dbSNO database also provides powerful structural and functional analysis functions to help researchers better understand the structural correlation and shared motifs of these SNO peptide sequences. The dbSNO database is divided into two versions: the first version ended in April 2012, and this version contains 43,000 experimentally verified SNO peptide sequences collected in numerous published studies using text mining methods; the second version is dbSNO2.0. In this version, dbSNO2.0 is also expanded to explore the structural environment of low SNO sites and the regulatory network resources of S-nitrosylation proteins. In SNO site prediction experiments, many scientists have also used the S-nitrosylation peptide sequence of the dbSNO database [41, 44].

PRISMOID [45] is a newly established database focusing on posttranslational modification and mutations with functional impact. Compared with traditional databases that focus on protein sequences, PRISMOID has added the real 3D structure of proteins and is equipped with various friendly operation interactions for information visualization. This database is the first version and contains 37 kinds of PTM annotation data (323 nitrosylation sites) manually compiled and is expected to be updated at least every 6 months. In addition, PRISMOID also integrates information such as the protein secondary structure and protein disordered regions to facilitate researchers to carry out scientific research.

2.2. Datasets. With the continuous in-depth understanding of the characteristics of S-nitrosylation, an increasing number of SNO peptide sequences have been identified, and the datasets used to predict SNO sites are based on previous studies. Dynamic changes are taking place. Therefore, the datasets commonly used by researchers for the detection of S-nitrosylation sites are chronologically explained in this section.

SNOSID, the first bioinformatics tool for predicting S-nitrosylation sites, was developed by Hao et al. [46]. In this study, they used S-nitrosoglutathione-treated rat cerebellar lysates. In 56 of the proteins, 68 cysteine sites were designated, and the initial limited 65 positive and negative samples were selected in the random sampling process. Xue et al. [47] also developed a predictor GPS-SNO for predicting SNO sites. They collected 363 experimentally verified S-nitrosylation sites published on PubMed using nitrosylated or nitrosylation as keywords and then integrated the public database SysPTM [48] and two large-scale S-nitrosylation site surveys [46, 49]. Finally, 504 positive sites and 2581 negative sites were obtained through sequence identity threshold setting [50] and protein sequence alignment [51]. A year later, Li et al. [52] used GPS-SNO datasets to develop a method for predicting SNO sites using SVM. Before long, Lee et al. [53] also developed a tool for predicting SNO sites, SNOSite. In this study, the training set and test set are from Chen et al. [54] and GPS-SNO data, respectively. Chen et al. used a high-throughput S-alkylating biotin conversion method in SNAP/L cysteine-stimulated mouse endothelial cells to obtain 586 positive sites and 2728 negative sites. In addition, since the data came from different datasets, the test set and training set may have the same homology. Therefore, they first defined SNO sequences with more than 30% identity as homologous sequences and then used BLAST 2 [55] to compare the fragment sequences. A test set containing 479 positive sites and 2501 negative sites was finally obtained.

In 2012, Li et al. [56] developed a method to predict and analyse SNO sites using minimal-redundancy-maximal-relevance and incremental feature selection. The dataset used in this experiment had three sources. The first source of SNO sites was the UniProt database [57] (version 2011_07) and the second from GPS-SNO and the third from large-scale S-nitrosylation site surveys [58–61] at that time to obtain the remaining two datasets. Finally, a training set (784 positive sites and 1568 negative sites) and a test set (43 positive sites and 121 negative sites) were obtained. In 2013, Xu et al. [41] developed iSNO-PseAAC, a tool for predicting S-nitrosylation sites. They randomly selected 438 proteins from the dbSNO database, and the sequence identity of these proteins was less than 40%. After comparison with the annotations in the dbSNO database, 731 positive sites and 810 negative sites verified by experiments were collected in the UniProt database [62] (version 2012_08). Xu et al. [63] improved on the basis of iSNO-PseAAC and developed iSNO-AAPair. This experiment used the original data of S-nitrosylation sites from dbSNO (version 1.0) and the UniProt database (version 2012_08). By using Chou's peptide formula [64–67], sequence identity setting, and random selection, 2300 SNO-positive and SNO-negative sites verified by experiments were obtained as training sets, and 81 positive and 100 negative sites were obtained as test sets.

In 2014, Jia et al. [68] developed a bioinformatics tool named iSNO-ANBPB used to predict SNO sites. In this experiment, they used the dataset constructed by Li et al. [52] and iSNO-PseAAC and obtained 1229 positive sites and 1223 negative sites by sequence identity setting and clustering. Soon, Zhang et al. [44] also developed the experimental tool PSNO. To reach a consensus assessment with previous experiments, they first constructed a training set containing 731 positive and 810 negative loci and a test set containing 53 positive and 103 negative sites from the dbSNO database. In addition, the 2302 positive sites selected from the GPS-SNO dataset and the 81 positive sites and 100 negative sites selected from the iSNO-PseAAC dataset were used as the test set.

After a brief stagnation, Xie et al. [69] used deep learning technology to develop a bioinformatics tool, DeepNitro, to predict SNO sites. They searched the relevant literature published before June 30, 2015, from PubMed and obtained a training set containing 20862 sites (3409 positive sites and 17453 negative sites) through residue modification and sequence clustering. To reach a consensus with previous research, they collected the latest data and eliminated the repeated sequences in previous work. Finally, an independent test set was built (485 positive sites and 4947 negative sites).

In 2019, Li et al. [70] predicted S-nitrosylation sites by multifeature fusion. In this study, they used 731 positive sites and 810 negative sites of iSNO-PseAAC and iSNO-AAPair as the training set and 43 positive sites and 121 negative sites of Li et al. as the test set. At the same time, Hasan et al. [71] developed PreSNO and used the DeepNitro dataset. To avoid overestimation of the prediction model, CD-HIT was used to screen homology and eliminate SNO sequences with the same window. Furthermore, to avoid prediction bias, they adopted the method of randomly taking and merging the sequences to balance the number of SNO-positive and SNO-negative sites.

3. Research Review

For protein S-nitrosylation site prediction, the traditional method is based on biochemical methods, but the SNO sites predicted are time-wasting, strenuous, and extortionate. With continuous breakthroughs in sequence and structural biology, computing methods have gradually become the mainstream of current research. This method is low cost and efficient. This section focuses on computational methods based on machine learning or deep learning to provide researchers with a systematic understanding of the development of this field. Traditional biochemical methods are also briefly introduced.

3.1. Biochemical Methods. Jaffrey and Snyder [72] invented biotin switch assay (BSA) technology. This method first converts nitrosylated cysteine residues into biotinylated cysteine residues and then detects biotin or specific proteins by Western blot [73] to detect the proteins labelled by biotin. BSA not only greatly improves the feasibility of SNO protein identification but also promotes the improvement of high-throughput identification of SNO sites. In 2005, Gao et al. [74] proposed using BSA and protein sequencing technology to identify endogenous SNO sites. The method is simple and rapid and can meet the needs of separation, purification, and identification of SNO proteins.

In 2006, Hao et al. [46] extended the original biotin method and proposed a new improved method, SNOSID. SNOSID introduced a protein hydrolysis and digestion step before capturing the antibiotin protein. This step was not like the previous complete separation of the peptide fragment of SNO protein but the selective separation of the residues containing the SNO site before. SNOSID also introduced the machine learning algorithm SVM for the first time. In addition, the original limited 65 positive samples and 65 negative samples as training data, but the prediction results were not ideal.

Although SNOSID technology can identify the target proteins and target sites of S-nitrosylation, the degree of protein nitrosamine cannot be accurately measured. With the advancement of proteomics technology, Wu et al. [75] and Fares (2014) developed a technology combining BSA with an isotope-coded affinity tag (ICAT). This technique was the first to achieve large-scale identification of S-nitrosylation residues but is disadvantaged by its use of isotopes.

3.2. Computational Biology Methods. With the continuous emergence of massive biological sequences in the postgene era, traditional biochemical sequencing methods are far from being able to meet the needs of development. However, machine learning algorithms cannot directly deal with biological sequence data. Therefore, how to use discrete models or a certain way to express biological sequences and fully express their sequence information or key pattern features has become the focus and content of research in computational biology [76–84]. Since Chou proposed the pseudoamino acid composition [85, 86] or PseAAC [87], computational biology based on machine learning or deep learning has also developed rapidly. The following introduces the software and server based on Chou's five-step rule to predict protein S-nitrosylation sites through algorithms. See Table 1 for details.

3.2.1. GPS-SNO. Xue et al. [47] developed GPS-SNO1.0, a tool for predicting protein S-nitrosylation sites using the

Table 1: List of 13 predictors for predicting the SNO sites in protein sequences.

No.	Name	Link	Time	Refs
1	SNOSID	Not provided	2006	[46]
2	GPS-SNO	http://sno.biocuckoo.org/	2010	[47]
3	CPR-SNO	http://math.cau.edu.cn/CPR-SNO	2011	[52]
4	SNOSite	http://csb.cse.yzu.edu.tw/SNOSite	2011	[53]
5	Li et al.	Not provided	2012	[56]
6	iSNO-PseAAC	http://app.aporc.org/iSNO-PseAAC	2013	[41]
7	iSNO-AAPair	http://app.aporc.org/iSNO-AAPair	2013	[63]
8	iSNO-ANBPB	Not provided	2014	[68]
9	PSNO	http://59.73.198.144:8088/PSNO	2014	[44]
10	DeepNitro	http://deepnitro.renlab.org	2018	[69]
11	PreSNO	http://kurata14.bio.kyutech.ac.jp/PreSNO	2019	[71]
12	Li et al.	Not provided	2019	[70]

GPS3.0 algorithm. The software is developed on the basis of the GPS2.0 algorithm [88] previously proposed. In this study, they first used the amino acid substitution matrix to calculate the nitrosylation peptide sequence and obtain the corresponding score. Then, k-means clustering, peptide selection (PS), weight training (WT), and matrix mutation (MaM) were used to improve the performance. The accuracy of the experiment under low threshold conditions was 75.80%, the sensitivity was 53.57%, and the specificity was 80.14%. In addition, the prediction ability of GPS-SNO on 485 potential S-nitrosylation low positions was also tested, and 371 positions of these targets were successfully predicted. GPS-SNO can be obtained for free from the website http://sno.biocuckoo.org/.

3.2.2. CPR-SNO. GPS-SNO has initially explored its ability on S-nitrosylated substrates. Although good results have been achieved, there is still room for improvement. Li et al. [52] developed CPR-SNO. In this study, they used SVM as a classifier and used the coding scheme based on coupling mode to realize the prediction system. In the performance evaluation, the F-score is used to identify the effective coding scheme, and referencing the work of Xue et al. [47], tenfold cross-validation is used for verification. In addition, this research solves the problem of existing coding schemes not being able to provide enough information to predict SNO sites. By using the F-score to identify effective coupling modes, they proved that some coupling modes are not related to S-nitrosylation. The CPR-SNO server is no longer in use.

3.2.3. SNOSite. Although traditional research on the characteristics and mechanism of S-nitrosylation has made great progress, the understanding of its substrate specificity is still insufficient. In 2011, Lee et al. [53] made a breakthrough on this issue and developed a new bioinformatics tool, SNOSite, for predicting SNO sites. In this study, they used maximal dependence decomposition (MDD) to serialize the nitrosylation sites into different subgroups and used SVM to generate a prediction model for each MDD cluster motif. By using fivefold cross-validation, the SVM using MDD clustering achieves 90% accuracy. SNOSite can be used for free on the website http://csb.cse.yzu.edu.tw/SNOSite/.

3.2.4. mRMR and IFS Method. Feature selection is useful for machine learning-based biosequence analysis [22, 89–105], including SNO prediction. Li et al. [56] developed a predictor based on the nearest neighbour algorithm [106] (NNA), which uses maximum relevance minimum redundancy [107] (mRMR) for incremental feature selection [108–110] (IFS). In this work, they generated 666 features from the peptide sequences used in the experiment and then used mRMR to rank the relevance and redundancy of these features in order of importance. For the obtained feature rankings, the best features are determined through IFS, and then, these features are constructed into different feature sets. Finally, the predictive evaluation performance of each feature set is generated by NNA. The best feature combination composed of 67 features is selected through the above method, and an accuracy of 0.61607 is obtained in the test set. In addition, this experiment also shows that the characteristics of the site far from the central cysteine can help determine the S-nitrosylation site. There is no online server for this predictor.

3.2.5. iSNO-PseAAC and iSNO-AAPair. Xu et al. [41] proposed a new SNO site predictor iSNO-PseAAC. In this study, they used PseAAC to represent protein sequence information, constructed as a 21×20 position-specific amino acid propensity (PSAAP) matrix, and finally used the conditional random field (CRF) algorithm to construct a predictor for predicting SNO sites. The cross-validation test of iSNO-PseAAC on an independent dataset also achieved a success rate of over 90%. iSNO-PseAAC can be obtained for free on the website http://app.aporc.org/iSNO-PseAAC/. However, iSNO-PseAAC simply considers the positional orientation of each group of amino acids when predicting variables but does not consider any correlation between them. The amino acids in all proteins are processed individually. However, there must be some connection between them in physiology or mechanism. To solve this problem, Xu et al. [63] made improvements on the basis of iSNO-PseAAC, added related influences when predicting protein SNO sites, and released a new SNO site prediction tool iSNO-AAPair. It considers the coupling effects of all pairs formed by the closest residues along the protein chain and the pairs formed by the closest residues. The predictor was cross-validated on the latest benchmark test set and achieved good performance. iSNO-AAPair can be obtained for free on the website http://app.aporc.org/iSNO-AAPair/.

3.2.6. iSNO-ANBPB. Jia et al. [68] proposed an iSNO-ANBPB predictor based on support vector machines. In this study, they constructed four feature extraction schemes and combined Chou's pseudoamino acid composition for model evaluation. The cross-validation of the basic SVM showed that the combination scheme using ANBPB for feature extraction obtained the best test results. In addition, studies [56] have shown that examples of the static charge of amino acids in cysteine residues and the secondary structure of amino acids play a key role in the prediction of SNO sites. Therefore, in addition to feature extraction, this study also considered the physical and chemical information in the peptide sequence. There is no online server for this predictor.

3.2.7. PSNO. In 2014, Zhang et al. [44] proposed a new bioinformatics tool, PSNO, for predicting SNO sites. In this study, they studied various derived features of the experimental sequence and integrated them into PseAAC to represent the experimental sample. In addition, to prevent the increase in the amount of information from increasing the difficulty of feature dimensions and predictors [111], they used relative entropy to discard noisy features from the high-level space and then optimize the optimal feature subset. However, the features of the optimal subset are different, so IFS is used here to rank these features, and a classifier based on 10-fold cross-validation is constructed for each of the optimal feature subsets. Finally, the *k*-nearest neighbour algorithm is used to predict the input sample and discriminate the prediction samples. In 10-fold cross-validation, the accuracy of PSNO was 75.67%, and the accuracy of MCC was 0.5119. With the completion of the whole-genome sequencing project, the gap in the sequence structure is rapidly expanding. In the absence of a protein structure, sequence-based prediction represented by PSNO can become a powerful supplement to replace structure-based prediction. The server provided by the software is now invalid.

3.2.8. DeepNitro. Since Hinton et al. [112] proposed the hierarchical training strategy to solve the gradient diffusion problem in 2006, deep learning technology has also been widely used in computational biology [113–126] and drug discovery [127–134]. In 2018, Xie et al. [69] used the deep learning algorithm for the first time to develop the S-nitrosylation site prediction bioinformatics tool DeepNitro. DeepNitro is an eight-layer neural network. The first layer is the data input layer, which is used to assign prediction and training values to neurons; the second to seventh layers are fully connected layers, of which the second to fourth layers use the dropout algorithm to improve the generalization ability of unknown data.

In the process of neural network design, to solve the problem of gradient diffusion in the training process, the ReLU function was used as the activation function, and the log-likelihood probability was used as the loss function to optimize the weights and other parameters in the neural network. In the process of backpropagation, a minibatch gradient descent algorithm is used to update the network parameters. Compared with traditional optimization algorithms, the momentum method is superior in optimizing parameters such as weights, so the momentum method was selected as the optimization function. In addition, L1 and L2 regular terms are introduced as hyperparameters to prevent overfitting. For the last layer, the softmax algorithm is used to obtain the probability distribution of the prediction results. Finally, through principal component analysis (PCA), DeepNitro obtained an AUC value of 0.7437 on the test set. DeepNitro uses deep learning algorithms, new encoding algorithms, and a position-specific scoring matrix [135] (PSSM) to greatly improve the accuracy of nitrosation site prediction and provides a free website server (http://deepnitro. http://renlab.org/) for academic research.

3.2.9. PreSNO. In 2019, Hasan et al. [71] proposed a prediction tool, PreSNO, for predicting protein SNO sites by an ensemble algorithm. The focus of the study was the use of four different coding schemes, including the composition of profile-based amino acids (CPA), *K*-space spectral amino acid composition (SAC), tripeptide composition from the PSSM (TCP), and physical-chemical properties of amino acids (PPA). The four coding schemes use SVM and random forest to calculate the probability score and then multiply it by weight to calculate the prediction effect of PreSNO. Through 5-fold cross-validation, PreSNO also achieved excellent performance. The predictor can be obtained for free on the website (http://kurata14.bio.kyutech.ac.jp/PreSNO/).

3.2.10. Multiple Features Combination Method. Soon, Li et al. [70] proposed a method to predict protein S-nitrosylation sites using multifeature mixing. This work improves prediction performance by extracting nine sequence features, such as parallel correlation pseudoamino acid composition (PC-PseAAC), general parallel correlation pseudoamino acid composition [136], and ANBPB. Then, the importance of amino acids is evaluated by subtracting the given amino acids from the information gain [137] (IG), and finally, the max-relevance-max-distance [138] (MRMD) generates a feature subset with lower redundancy and strong correlation with the target category. In the cross-validation of the test set, the ACC and MCC of this method were 73.17% and 0.3788, respectively, which becomes a useful supplement to the existing SNO identification tools.

4. Concluding Remarks and Perspectives

Many physiological and pathological studies of SNO have been reported in recent years. Therefore, accurate prediction of SNO sites will pave the way to speed up related drug development.

Several exciting computational methods have been proposed to predict SNO. Although these works promoted research on SNO and facilitated the prediction of SNO sites, the following challenges should be considered in future works.

Although many predictors have been developed to predict SNO sites, some corresponding indicators have greatly improved the space. This is because existing methods were trained on the basis of an imbalanced dataset. To solve this problem, it is necessary to collect many more positive SNO sites to enlarge the number of SNO sites in the dataset and balance it. In addition, the focus of future research in this field is to use these new technologies and methods to predict more nitrosylated target proteins and sites to reveal the mechanism by which nitrosylation regulates various physiological processes.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

The work was supported by the Natural Science Foundation of China (32072259, 31901114, and 81773161), the Natural Science Foundation of Heilongjiang Province (LH2020F009), the Support Program for Young Creative Talents of Harbin University of Commerce (2019CX01 and 2019CX22), the University Nursing Program for Young Scholar with Creative Talents in Heilongjiang Province (UNPYSCT-2020212), the Chinese Postdoctoral Science Foundation (2017M621294), and the first batch of industry-university cooperative education project of the Ministry of Education in 2019 (201901240006).

References

[1] M. W. Foster, D. T. Hess, and J. S. Stamler, "Protein *S*-nitrosylation in health and disease: a current perspective," *Trends in Molecular Medicine*, vol. 15, no. 9, pp. 391–404, 2009.

[2] B. Cui, Q. Pan, D. Clarke et al., "*S*-nitrosylation of the zinc finger protein SRG1 regulates plant immunity," *Nature Communications*, vol. 9, no. 1, article 4226, 2018.

[3] S. Rizza, S. Cardaci, C. Montagna et al., "S-Nitrosylation drives cell senescence and aging in mammals by controlling mitochondrial dynamics and mitophagy," *Proceedings of the National Academy of Sciences*, vol. 115, no. 15, pp. E3388–E3397, 2018.

[4] R. González, A. Cruz, G. Ferrín et al., "Nitric oxide mimics transcriptional and post-translational regulation during α-tocopherol cytoprotection against glycochenodeoxycholate-induced cell death in hepatocytes," *Journal of Hepatology*, vol. 55, no. 1, pp. 133–144, 2011.

[5] G. Liu and Q. Jiang, "Alzheimer's diseaseCD33rs3865444 variant does not contribute to cognitive performance," *Proceedings of the National Academy of Sciences*, vol. 113, no. 12, pp. E1589–E1590, 2016.

[6] Q. Jiang, S. Jin, Y. Jiang et al., "Alzheimer's disease variants with the genome-wide significance are significantly enriched in immune pathways and active in immune cells," *Molecular Neurobiology*, vol. 54, no. 1, pp. 594–600, 2017.

[7] T. S. Wijasa, M. Sylvester, N. Brocke-Ahmadinejad et al., "Quantitative proteomics of synaptosome S-nitrosylation in Alzheimer's disease," *Journal of Neurochemistry*, vol. 152, no. 6, pp. 710–726, 2020.

[8] D. Mishra, V. Patel, and D. Banerjee, "Nitric oxide and S-nitrosylation in cancers: emphasis on breast cancer," *Breast Cancer: Basic and Clinical Research*, vol. 14, article 117822341988268, 2020.

[9] S. L. Cook and G. P. Jackson, "Characterization of tyrosine nitration and cysteine nitrosylation modifications by metastable atom-activation dissociation mass spectrometry," *Journal of the American Society for Mass Spectrometry*, vol. 22, no. 2, pp. 221–232, 2011.

[10] B. Li, J. Tang, Q. Yang et al., "NOREVA: normalization and evaluation of MS-based metabolomics data," *Nucleic Acids Research*, vol. 45, no. W1, pp. W162–W170, 2017.

[11] K. Chen, Z. Wei, Q. Zhang et al., "WHISTLE: a high-accuracy map of the human N6-methyladenosine (m6A) epitranscriptome predicted using a machine learning approach," *Nucleic Acids Research*, vol. 47, no. 7, article e41, 2019.

[12] S. Amanat, A. Ashraf, W. Hussain, N. Rasool, and Y. D. Khan, "Identification of lysine carboxylation sites in proteins by integrating statistical moments and position relative features via general PseAAC," *Current Bioinformatics*, vol. 15, no. 5, pp. 396–407, 2020.

[13] W. Bao, D.-S. Huang, and Y.-H. Chen, "MSIT: malonylation sites identification tree," *Current Bioinformatics*, vol. 15, no. 1, pp. 59–67, 2020.

[14] W. Chen and K. Liu, "Analysis and comparison of RNA pseudouridine site prediction tools," *Current Bioinformatics*, vol. 15, no. 4, pp. 279–286, 2020.

[15] M. A. M. Hasan, M. K. Ben Islam, J. Rahman, and S. Ahmad, "Citrullination site prediction by incorporating sequence coupled effects into PseAAC and resolving data imbalance issue," *Current Bioinformatics*, vol. 15, no. 3, pp. 235–245, 2020.

[16] H. Long, Z. Sun, M. Li, H. Y. Fu, and M. C. Lin, "Predicting protein phosphorylation sites based on deep learning," *Current Bioinformatics*, vol. 15, no. 4, pp. 300–308, 2020.

[17] H. Zhu, X. Du, and Y. Yao, "ConvsPPIS: identifying protein-protein interaction sites by an ensemble convolutional neural network with feature graph," *Current Bioinformatics*, vol. 15, no. 4, pp. 368–378, 2020.

[18] G. Wang, X. Luo, J. Wang et al., "MeDReaders: a database for transcription factors that bind to methylated DNA," *Nucleic Acids Research*, vol. 46, no. D1, pp. D146–D151, 2018.

[19] W.-Z. Zhong and S.-F. Zhou, "Molecular science for drug development and biomedicine," *International Journal of Molecular Sciences*, vol. 15, no. 11, pp. 20072–20078, 2014.

[20] W. Xue, F. Yang, P. Wang et al., "What contributes to serotonin-norepinephrine reuptake inhibitors' dual-targeting mechanism? The key role of transmembrane domain 6 in human serotonin and norepinephrine transporters revealed by molecular dynamics simulation," *ACS Chemical Neuroscience*, vol. 9, no. 5, pp. 1128–1140, 2018.

[21] H. Xu, J. Zhou, S. Lin, W. Deng, Y. Zhang, and Y. Xue, "PLMD: an updated data resource of protein lysine modifications," *Journal of Genetics and Genomics*, vol. 44, no. 5, pp. 243–250, 2017.

[22] H. Lv, F.-Y. Dao, D. Zhang et al., "iDNA-MS: an integrated computational tool for detecting DNA modification sites in multiple genomes," *iScience*, vol. 23, no. 4, article 100991, 2020.

[23] Q. Zou, G. Lin, X. Jiang, X. Liu, and X. Zeng, "Sequence clustering in bioinformatics: an empirical study," *Briefings in Bioinformatics*, vol. 21, no. 1, pp. 1–10, 2018.

[24] J. Shao and B. Liu, "ProtFold-DFG: protein fold recognition by combining Directed Fusion Graph and PageRank algorithm," *Briefings in Bioinformatics*, 2020.

[25] B. Liu, C. Li, and K. Yan, "DeepSVM-fold: protein fold recognition by combining support vector machines and pairwise sequence similarity scores generated by deep learning networks," *Briefings in Bioinformatics*, vol. 21, no. 5, pp. 1733–1741, 2020.

[26] Y.-d. Liao and Z.-r. Jiang, "MoABank: an integrated database for drug mode of action knowledge," *Current Bioinformatics*, vol. 14, no. 5, pp. 446–449, 2019.

[27] A. Munir, S. I. Malik, and K. A. Malik, "Proteome mining for the identification of putative drug targets for human pathogen Clostridium tetani," *Current Bioinformatics*, vol. 14, no. 6, pp. 532–540, 2019.

[28] N. Srivastava, B. N. Mishra, and P. Srivastava, "In-silico identification of drug lead molecule against pesticide exposed-neurodevelopmental disorders through network-based computational model approach," *Current Bioinformatics*, vol. 14, no. 5, pp. 460–467, 2019.

[29] X. Zhao, L. Chen, Z.-H. Guo, and T. Liu, "Predicting drug side effects with compact integration of heterogeneous networks," *Current Bioinformatics*, vol. 14, no. 8, pp. 709–720, 2019.

[30] R. Jakhar, M. Dangi, A. Khichi, and A. K. Chhillar, "Relevance of molecular docking studies in drug designing," *Current Bioinformatics*, vol. 15, no. 4, pp. 270–278, 2020.

[31] Z. Li, T. Zhang, H. Lei et al., "Research on gastric cancer's drug-resistant gene regulatory network model," *Current Bioinformatics*, vol. 15, no. 3, pp. 225–234, 2020.

[32] J. Liu, X. Lian, F. Liu et al., "Identification of novel key targets and candidate drugs in oral squamous cell carcinoma," *Current Bioinformatics*, vol. 15, no. 4, pp. 328–337, 2020.

[33] J. Wang, H. Wang, X. Wang, and H. Chang, "Predicting drug-target interactions *via* FM-DNN learning," *Current Bioinformatics*, vol. 15, no. 1, pp. 68–76, 2020.

[34] Q. Jiang, G. Wang, S. Jin, Y. Li, and Y. Wang, "Predicting human microRNA-disease associations based on support vector machine," *International Journal of Data Mining and Bioinformatics*, vol. 8, no. 3, pp. 282–293, 2013.

[35] G. Liu, Y. Hu, S. Jin, and Q. Jiang, "Genetic variant rs763361 regulates multiple sclerosisCD226gene expression," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 114, no. 6, pp. E906–E907, 2017.

[36] W. Chen, P. Feng, T. Liu, and D. Jin, "Recent advances in machine learning methods for predicting heat shock proteins," *Current Drug Metabolism*, vol. 20, no. 3, pp. 224–228, 2019.

[37] K.-C. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 273, no. 1, pp. 236–247, 2011.

[38] Y. W. Zhao, Z. D. Su, W. Yang, H. Lin, W. Chen, and H. Tang, "IonchanPred 2.0: a tool to predict ion channels and their types," *International Journal of Molecular Sciences*, vol. 18, no. 9, article 1838, 2017.

[39] Z. Y. Liang, H. Y. Lai, H. Yang et al., "Pro54DB: a database for experimentally verified sigma-54 promoters," *Bioinformatics*, vol. 33, no. 3, pp. 467–469, 2017.

[40] U Consortium, "UniProt: a worldwide hub of protein knowledge," *Nucleic Acids Research*, vol. 47, no. D1, pp. D506–D515, 2019.

[41] Y. Xu, J. Ding, L.-Y. Wu, and K. C. Chou, "iSNO-PseAAC: predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition," *PLoS One*, vol. 8, no. 2, article e55844, 2013.

[42] Y. Wang, S. Zhang, F. Li et al., "Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics," *Nucleic Acids Research*, vol. 48, no. D1, pp. D1031–D1041, 2020.

[43] Y.-J. Chen, C.-T. Lu, M.-G. Su et al., "dbSNO 2.0: a resource for exploring structural environment, functional and disease association and regulatory network of protein S-nitrosylation," *Nucleic Acids Research*, vol. 43, no. D1, pp. D503–D511, 2015.

[44] J. Zhang, X. Zhao, P. Sun, and Z. Ma, "PSNO: predicting cysteine S-nitrosylation sites by incorporating various sequence-derived features into the general form of Chou's PseAAC," *International Journal of Molecular Sciences*, vol. 15, no. 7, pp. 11204–11219, 2014.

[45] F. Li, C. Fan, T. T. Marquez-Lago et al., "PRISMOID: a comprehensive 3D structure database for post-translational modifications and mutations with functional impact," *Briefings in Bioinformatics*, vol. 21, no. 3, pp. 1069–1079, 2020.

[46] G. Hao, B. Derakhshan, L. Shi, F. Campagne, and S. S. Gross, "SNOSID, a proteomic method for identification of cysteine S-nitrosylation sites in complex protein mixtures," *Proceedings of the National Academy of Sciences*, vol. 103, no. 4, pp. 1012–1017, 2006.

[47] Y. Xue, Z. Liu, X. Gao et al., "GPS-SNO: computational prediction of protein S-nitrosylation sites with a modified GPS algorithm," *PLoS One*, vol. 5, no. 6, article e11290, 2010.

[48] H. Li, X. Xing, G. Ding et al., "SysPTM: a systematic resource for proteomic research on post-translational modifications," *Molecular & Cellular Proteomics*, vol. 8, no. 8, pp. 1839–1849, 2009.

[49] P. Han and C. Chen, "Detergent-free biotin switch combined with liquid chromatography/tandem mass spectrometry in the analysis of S-nitrosylated proteins," *Rapid Communications in Mass Spectrometry*, vol. 22, no. 8, pp. 1137–1145, 2008.

[50] W. Li and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics*, vol. 22, no. 13, pp. 1658-1659, 2006.

[51] S. F. Altschul, T. L. Madden, A. A. Schäffer et al., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[52] Y.-X. Li, Y.-H. Shao, L. Jing, and N. Y. Deng, "An efficient support vector machine approach for identifying protein S-nitrosylation sites," *Protein and Peptide Letters*, vol. 18, no. 6, pp. 573–587, 2011.

[53] T.-Y. Lee, Y.-J. Chen, T.-C. Lu, H. D. Huang, and Y. J. Chen, "SNOSite: exploiting maximal dependence decomposition to

identify cysteine S-nitrosylation with substrate site specificity," *PLoS One*, vol. 6, no. 7, article e21849, 2011.

[54] Y.-J. Chen, W.-C. Ku, P.-Y. Lin, H. C. Chou, K. H. Khoo, and Y. J. Chen, "S-Alkylating labeling strategy for site-specific identification of theS-nitrosoproteome," *Journal of Proteome Research*, vol. 9, no. 12, pp. 6417–6439, 2010.

[55] T. A. Tatusova and T. L. Madden, "BLAST 2 sequences, a new tool for comparing protein and nucleotide sequences," *FEMS Microbiology Letters*, vol. 174, no. 2, pp. 247–250, 1999.

[56] B.-Q. Li, L.-L. Hu, S. Niu, Y. D. Cai, and K. C. Chou, "Predict and analyze S-nitrosylation modification sites with the mRMR and IFS approaches," *Journal of Proteomics*, vol. 75, no. 5, pp. 1654–1665, 2012.

[57] U Consortium, "The universal protein resource (UniProt) in 2010," *Nucleic Acids Research*, vol. 38, suppl_1, pp. D142–D148, 2010.

[58] P.-T. Doulias, J. L. Greene, T. M. Greco et al., "Structural profiling of endogenous S-nitrosocysteine residues reveals unique features that accommodate diverse mechanisms for protein S-nitrosylation," *Proceedings of the National Academy of Sciences*, vol. 107, no. 39, pp. 16958–16963, 2010.

[59] Q. Yang, Y. Wang, Y. Zhang et al., "NOREVA: enhanced normalization and evaluation of time-course and multi-class metabolomic data," *Nucleic Acids Research*, vol. 48, no. W1, pp. W436–W448, 2020.

[60] M. Liu, J. Hou, L. Huang et al., "Site-specific proteomics approach for study protein S-nitrosylation," *Analytical Chemistry*, vol. 82, no. 17, pp. 7160–7168, 2010.

[61] Q. Yang, B. Li, S. Chen et al., "MMEASE: online meta-analysis of metabolomic data by enhanced metabolite annotation, marker selection and enrichment analysis," *Journal of Proteomics*, vol. 232, article 104023, 2021.

[62] U Consortium, "Reorganizing the protein space at the Universal Protein Resource (UniProt)," *Nucleic Acids Research*, vol. 40, no. D1, pp. D71–D75, 2012.

[63] Y. Xu, X.-J. Shao, L.-Y. Wu, N. Y. Deng, and K. C. Chou, "iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins," *PeerJ*, vol. 1, article e171, 2013.

[64] K.-C. Chou, "Prediction of signal peptides using scaled window," *Peptides*, vol. 22, no. 12, pp. 1973–1979, 2001.

[65] J. Tang, J. Fu, Y. Wang et al., "ANPELA: analysis and performance assessment of the label-free quantification workflow for metaproteomic studies," *Briefings in Bioinformatics*, vol. 21, no. 2, pp. 621–636, 2020.

[66] Q. Yang, J. Hong, Y. Li et al., "A novel bioinformatics approach to identify the consistently well-performing normalization strategy for current metabolomic studies," *Briefings in Bioinformatics*, vol. 21, no. 6, pp. 2142–2152, 2020.

[67] B. Liu, X. Gao, and H. Zhang, "BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA, and protein sequences at sequence level and residue level based on machine learning approaches," *Nucleic Acids Research*, vol. 47, no. 20, article e127, 2019.

[68] C. Jia, X. Lin, and Z. Wang, "Prediction of protein S-nitrosylation sites based on adapted normal distribution bi-profile Bayes and Chou's pseudo amino acid composition," *International Journal of Molecular Sciences*, vol. 15, no. 6, pp. 10410–10423, 2014.

[69] Y. Xie, X. Luo, Y. Li et al., "DeepNitro: prediction of protein nitration and nitrosylation sites by deep learning," *Genomics, Proteomics & Bioinformatics*, vol. 16, no. 4, pp. 294–306, 2018.

[70] T. Li, R. Song, Q. Yin, M. Gao, and Y. Chen, "Identification of *S*-nitrosylation sites based on multiple features combination," *Scientific Reports*, vol. 9, no. 1, article 3098, 2019.

[71] M. M. Hasan, B. Manavalan, M. S. Khatun, and H. Kurata, "Prediction of S-nitrosylation sites by integrating support vector machines and random forest," *Mol Omics*, vol. 15, no. 6, pp. 451–458, 2019.

[72] S. R. Jaffrey and S. H. Snyder, "The biotin switch method for the detection of S-nitrosylated proteins," *Science's STKE*, vol. 2001, no. 86, p. pl1, 2001.

[73] C. Lindermayr, G. Saalbach, and J. Durner, "Proteomic identification of S-nitrosylated proteins in Arabidopsis," *Plant Physiology*, vol. 137, no. 3, pp. 921–930, 2005.

[74] C. Gao, H. Guo, J. Wei, Z. Mi, P. Y. Wai, and P. C. Kuo, "Identification of S-nitrosylated proteins in endotoxin-stimulated RAW264.7 murine macrophages," *Nitric Oxide*, vol. 12, no. 2, pp. 121–126, 2005.

[75] C. Wu, T. Liu, W. Chen et al., "Redox regulatory mechanism of transnitrosylation by thioredoxin," *Molecular & Cellular Proteomics*, vol. 9, no. 10, pp. 2262–2275, 2010.

[76] K.-C. Chou, "Impacts of bioinformatics to medicinal chemistry," *Medicinal Chemistry*, vol. 11, no. 3, pp. 218–234, 2015.

[77] X. Liang, W. Zhu, Z. Lv, and Q. Zou, "Molecular computing and bioinformatics," *Multidisciplinary Digital Publishing Institute*, vol. 24, no. 13, article 2358, 2019.

[78] Z. Lv, C. Ao, and Q. Zou, "Protein function prediction: from traditional classifier to deep learning," *Proteomics*, vol. 19, no. 14, article 1900119, 2019.

[79] Z. Lv, P. Wang, Q. Zou, and Q. Jiang, "Identification of sub-Golgi protein localization by use of deep representation learning features," *Bioinformatics*, 2020.

[80] G. Liu, S. Jin, Y. Hu, and Q. Jiang, "Disease status affects the association between rs4813620 and the expression of Alzheimer's disease susceptibility geneTRIB3," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 115, no. 45, pp. E10519–E10520, 2018.

[81] X. Zhao, Q. Jiao, H. Li et al., "ECFS-DEA: an ensemble classifier-based feature selection for differential expression analysis on expression profiles," *BMC Bioinformatics*, vol. 21, no. 1, pp. 43–43, 2020.

[82] K. Liu and W. Chen, "iMRM:a platform for simultaneously identifying multiple kinds of RNA modifications," *Bioinformatics*, vol. 36, no. 11, pp. 3336–3342, 2020.

[83] D. Zhang, Z. C. Xu, W. Su et al., "iCarPS: a computational tool for identifying protein carbonylation sites by novel encoded features," *Bioinformatics*, 2020.

[84] H. Lv, F.-Y. Dao, Z.-X. Guan, H. Yang, Y. W. Li, and H. Lin, "Deep-Kcr: accurate detection of lysine crotonylation sites using deep learning method," *Briefings in Bioinformatics*, 2020.

[85] K. C. Chou, "Prediction of protein cellular attributes using pseudo-amino acid composition," *Proteins*, vol. 43, no. 3, pp. 246–255, 2001.

[86] J. Hong, Y. Luo, Y. Zhang et al., "Protein functional annotation of simultaneously improved stability, accuracy and false discovery rate achieved by a sequence-based deep learning," *Briefings in Bioinformatics*, vol. 21, no. 4, pp. 1437–1447, 2020.

[87] J. Hong, Y. Luo, M. Mou et al., "Convolutional neural network-based annotation of bacterial type IV secretion system effectors with enhanced accuracy and reduced false discovery," *Briefings in Bioinformatics*, vol. 21, no. 5, pp. 1825–1836, 2020.

[88] Y. Xue, J. Ren, X. Gao, C. Jin, L. Wen, and X. Yao, "GPS 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy," *Molecular & Cellular Proteomics*, vol. 7, no. 9, pp. 1598–1608, 2008.

[89] Z. Lv, D. Wang, H. Ding, B. Zhong, and L. Xu, "Escherichia coli DNA N-4-methycytosine site prediction accuracy improved by light gradient boosting machine feature selection technology," *Ieee Access*, vol. 8, pp. 14851–14859, 2020.

[90] Z. Lv, J. Zhang, H. Ding, and Q. Zou, "RF-PseU: a random forest predictor for RNA pseudouridine sites," *Frontiers in Bioengineering and Biotechnology*, vol. 8, p. 134, 2020.

[91] R. Su, J. Hu, Q. Zou, B. Manavalan, and L. Wei, "Empirical comparison and analysis of web-based cell-penetrating peptide prediction tools," *Briefings in Bioinformatics*, vol. 21, no. 2, pp. 408–420, 2020.

[92] R. Su, X. Liu, L. Wei, and Q. Zou, "Deep-Resp-Forest: a deep forest model to predict anti-cancer drug response," *Methods*, vol. 166, pp. 91–102, 2019.

[93] R. Su, X. Liu, G. Xiao, and L. Wei, "Meta-GDBP: a high-level stacked regression model to improve anticancer drug response prediction," *Briefings in Bioinformatics*, vol. 21, no. 3, pp. 996–1005, 2020.

[94] R. Su, H. Wu, B. Xu, X. Liu, and L. Wei, "Developing a multi-dose computational model for drug-induced hepatotoxicity prediction based on toxicogenomics data," *Ieee-Acm Transactions on Computational Biology and Bioinformatics*, vol. 16, no. 4, pp. 1231–1239, 2019.

[95] L. Wei, J. Hu, F. Li, J. Song, R. Su, and Q. Zou, "Comparative analysis and prediction of quorum-sensing peptides using feature representation learning and machine learning algorithms," *Briefings in Bioinformatics*, vol. 21, no. 1, pp. 106–119, 2018.

[96] L. Wei, M. Liao, Y. Gao, R. Ji, Z. He, and Q. Zou, "Improved and promising identification of human microRNAs by incorporating a high-quality negative set," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 11, no. 1, pp. 192–201, 2014.

[97] L. Wei, S. Wan, J. Guo, and K. K. L. Wong, "A novel hierarchical selective ensemble classifier with bioinformatics application," *Artificial Intelligence in Medicine*, vol. 83, pp. 82–90, 2017.

[98] L. Wei, P. Xing, J. Zeng, J. X. Chen, R. Su, and F. Guo, "Improved prediction of protein-protein interactions using novel negative samples, features, and an ensemble classifier," *Artificial Intelligence in Medicine*, vol. 83, pp. 67–74, 2017.

[99] Q. Jin, Z. Meng, T. D. Pham, Q. Chen, L. Wei, and R. Su, "DUNet: a deformable network for retinal vessel segmentation," *Knowledge-Based Systems*, vol. 178, pp. 149–162, 2019.

[100] L. Wei, C. Zhou, H. Chen, J. Song, and R. Su, "ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides," *Bioinformatics*, vol. 34, no. 23, pp. 4007–4016, 2018.

[101] G. Liu, Y. Hu, S. Jin, F. Zhang, Q. Jiang, and J. Hao, "Cis-eQTLs regulate reduced LST1 gene and NCR3 gene expression and contribute to increased autoimmune disease risk," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 113, no. 42, pp. E6321–E6322, 2016.

[102] G. Wang, Y. Wang, W. Feng et al., "Transcription factor and microRNA regulation in androgen-dependent and -independent prostate cancer cells," *BMC Genomics*, vol. 9, article S22, Supplement 2, 2008.

[103] Y. Zhao, F. Wang, and L. Juan, "MicroRNA promoter identification in Arabidopsis using multiple histone markers," *BioMed Research International*, vol. 2015, Article ID 861402, 10 pages, 2015.

[104] Y. Zhao, F. Wang, S. Chen, J. Wan, and G. Wang, "Methods of microRNA promoter prediction and transcription factor mediated regulatory network," *BioMed Research International*, vol. 2017, Article ID 7049406, 8 pages, 2017.

[105] W. Chen, P. Feng, and F. Nie, "iATP: a sequence based method for identifying anti-tubercular peptides," *Medicinal Chemistry*, vol. 16, no. 5, pp. 620–625, 2020.

[106] T. Denoeux, "A k-nearest neighbor classification rule based on Dempster-Shafer theory," in *Classic works of the Dempster-Shafer theory of belief functions*, pp. 737–760, Springer, 2008.

[107] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[108] Z. He, J. Zhang, X.-H. Shi et al., "Predicting drug-target interaction networks based on functional groups and biological features," *PLoS One*, vol. 5, no. 3, article e9603, 2010.

[109] F. Li, C. Li, M. Wang et al., "GlycoMine: a machine learning-based approach for predicting N-, C-and O-linked glycosylation in the human proteome," *Bioinformatics*, vol. 31, no. 9, pp. 1411–1419, 2015.

[110] F. Li, C. Li, J. Revote et al., "GlycoMine struct: a new bioinformatics tool for highly accurate mapping of the human N-linked and O-linked glycoproteomes by incorporating structural features," *Scientific Reports*, vol. 6, no. 1, pp. 1–16, 2016.

[111] J. Qian, D. Q. Miao, Z. Zhang, and W. Li, "Hybrid approaches to attribute reduction based on indiscernibility and discernibility relation," *International Journal of Approximate Reasoning*, vol. 52, no. 2, pp. 212–230, 2011.

[112] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[113] D. Quang, Y. Chen, and X. Xie, "DANN: a deep learning approach for annotating the pathogenicity of genetic variants," *Bioinformatics*, vol. 31, no. 5, pp. 761–763, 2015.

[114] Q. Yang, B. Li, J. Tang et al., "Consistent gene signature of schizophrenia identified by a novel feature selection strategy from comprehensive sets of transcriptomic data," *Briefings in Bioinformatics*, vol. 21, no. 3, pp. 1058–1068, 2020.

[115] Z. Hong, X. Zeng, L. Wei, and X. Liu, "Identifying enhancer–promoter interactions with neural network based on pre-trained DNA vectors and attention mechanism," *Bioinformatics*, vol. 36, no. 4, pp. 1037–1043, 2020.

[116] J. Tang, M. Mou, Y. Wang, Y. Luo, and F. Zhu, "MetaFS: performance assessment of biomarker discovery in metaproteomics," *Briefings in Bioinformatics*, 2020.

[117] S. Jin, X. Zeng, F. Xia, W. Huang, and X. Liu, "Application of deep learning methods in biological networks," *Briefings in Bioinformatics*, 2020.

[118] J. Tang, J. Fu, Y. Wang et al., "Simultaneous improvement in the precision, accuracy, and robustness of label- free proteome quantification by optimizing data manipulation chains*„" *Molecular & Cellular Proteomics*, vol. 18, no. 8, pp. 1683–1699, 2019.

[119] Y.-J. Tang, Y.-H. Pang, and B. Liu, "IDP-Seq2Seq: identification of intrinsically disordered regions based on sequence to sequence learning," *Bioinformaitcs*.

[120] C.-C. Li and B. Liu, "MotifCNN-fold: protein fold recognition based on fold-specific features extracted by motif-based convolutional neural networks," *Briefings in Bioinformatics*, vol. 21, no. 6, pp. 2133–2141, 2020.

[121] Z. Lv, H. Ding, L. Wang, and Q. Zou, "A convolutional neural network using dinucleotide one-hot encoder for identifying DNA N6-methyladenine sites in the rice genome," *Neurocomputing*, vol. 422, pp. 214–221, 2021.

[122] Z. Wang, W. He, J. Tang, and F. Guo, "Identification of highest-affinity binding sites of yeast transcription factor families," *Journal of Chemical Information and Modeling*, vol. 60, no. 3, pp. 1876–1883, 2020.

[123] H. Wang, Y. Ding, J. Tang, and F. Guo, "Identification of membrane protein types via multivariate information fusion with Hilbert-Schmidt Independence criterion," *Neurocomputing*, vol. 383, pp. 257–269, 2020.

[124] J. Li, Y. Pu, J. Tang, Q. Zou, and F. Guo, "DeepAVP: a dual-channel deep neural network for identifying variable-length antiviral peptides," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 10, pp. 3012–3019, 2020.

[125] Y. Shen, J. Tang, and F. Guo, "Identification of protein subcellular localization via integrating evolutionary and physicochemical information into Chou's general PseAAC," *Journal of Theoretical Biology*, vol. 462, pp. 230–239, 2019.

[126] G. Liu, F. Zhang, Y. Jiang et al., "Integrating genome-wide association studies and gene expression data highlights dysregulated multiple sclerosis risk pathways," *Multiple Sclerosis*, vol. 23, no. 2, pp. 205–212, 2017.

[127] X. Zeng, S. Zhu, Y. Hou et al., "Network-based prediction of drug–target interactions using an arbitrary-order proximity embedded deep forest," *Bioinformatics*, vol. 36, no. 9, pp. 2805–2812, 2020.

[128] Y. H. Li, X. X. Li, J. J. Hong et al., "Clinical trials, progression-speed differentiating features and swiftness rule of the innovative targets of first-in-class drugs," *Briefings in Bioinformatics*, vol. 21, no. 2, pp. 649–662, 2020.

[129] X. Fu, L. Cai, X. Zeng, and Q. Zou, "StackCPPred: a stacking and pairwise energy content-based prediction of cell-penetrating peptides and their uptake efficiency," *Bioinformatics*, vol. 36, no. 10, pp. 3028–3034, 2020.

[130] J. Yin, W. Sun, F. Li et al., "VARIDT 1.0: variability of drug transporter database," *Nucleic Acids Research*, vol. 48, no. D1, pp. D1042–D1050, 2020.

[131] M. Wang, X. Cui, B. Yu, C. Chen, Q. Ma, and H. Zhou, "SulSite-GTB: identification of protein S-sulfenylation sites by fusing multiple feature information and gradient tree boosting," *Neural Computing and Applications*, vol. 32, no. 17, pp. 13843–13862, 2020.

[132] Y. Ding, J. Tang, and F. Guo, "Identification of drug-side effect association via multiple information integration with centered kernel alignment," *Neurocomputing*, vol. 325, pp. 211–224, 2019.

[133] Y. Ding, J. Tang, and F. Guo, "The computational models of drug-target interaction prediction," *Protein and Peptide Letters*, vol. 27, no. 5, pp. 348–358, 2020.

[134] Y. Ding, J. Tang, and F. Guo, "Identification of drug-target interactions via fuzzy bipartite local model," *Neural Computing & Applications*, vol. 32, no. 14, pp. 10303–10319, 2020.

[135] V. Vacic, L. M. Iakoucheva, and P. Radivojac, "Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments," *Bioinformatics*, vol. 22, no. 12, pp. 1536-1537, 2006.

[136] D.-S. Cao, Q.-S. Xu, and Y.-Z. Liang, "propy: a tool to generate various modes of Chou's PseAAC," *Bioinformatics*, vol. 29, no. 7, pp. 960–962, 2013.

[137] B. Liu, X. Wang, L. Lin, Q. Dong, and X. Wang, "A discriminative method for protein remote homology detection and fold recognition combining Top-n-grams and latent semantic analysis," *BMC Bioinformatics*, vol. 9, no. 1, p. 510, 2008.

[138] Q. Zou, J. Zeng, L. Cao, and R. Ji, "A novel features ranking metric with application to scalable visual and bioinformatics data classification," *Neurocomputing*, vol. 173, pp. 346–354, 2016.

Hindawi
BioMed Research International
Volume 2021, Article ID 6690154, 16 pages
https://doi.org/10.1155/2021/6690154

Research Article

Drug-Target Interaction Prediction Based on Adversarial Bayesian Personalized Ranking

Yihua Ye,[1] **Yuqi Wen**,[2] **Zhongnan Zhang**,[1] **Song He**,[2] **and Xiaochen Bo**[2]

[1] *School of Informatics, Xiamen University, Xiamen 361005, China*
[2] *Department of Biotechnology, Beijing Institute of Radiation Medicine, Beijing 100850, China*

Correspondence should be addressed to Zhongnan Zhang; zhongnan_zhang@xmu.edu.cn, Song He; hes1224@163.com, and Xiaochen Bo; boxiaoc@163.com

Received 23 December 2020; Revised 17 January 2021; Accepted 23 January 2021; Published 10 February 2021

Academic Editor: Quan Zou

The prediction of drug-target interaction (DTI) is a key step in drug repositioning. In recent years, many studies have tried to use matrix factorization to predict DTI, but they only use known DTIs and ignore the features of drug and target expression profiles, resulting in limited prediction performance. In this study, we propose a new DTI prediction model named AdvB-DTI. Within this model, the features of drug and target expression profiles are associated with Adversarial Bayesian Personalized Ranking through matrix factorization. Firstly, according to the known drug-target relationships, a set of ternary partial order relationships is generated. Next, these partial order relationships are used to train the latent factor matrix of drugs and targets using the Adversarial Bayesian Personalized Ranking method, and the matrix factorization is improved by the features of drug and target expression profiles. Finally, the scores of drug-target pairs are achieved by the inner product of latent factors, and the DTI prediction is performed based on the score ranking. The proposed model effectively takes advantage of the idea of learning to rank to overcome the problem of data sparsity, and perturbation factors are introduced to make the model more robust. Experimental results show that our model could achieve a better DTI prediction performance.

1. Introduction

Drug repositioning is to discover new indications for existing drugs, which means that drug development based on approved drugs does not need to consider the safety and effectiveness of the original drug, effectively reducing the time of drug development process and cost. Prediction of drug-target interaction (DTI) which refers to the recognition of interactions between chemical compounds and the protein targets in the human body has become a key step in drug repositioning [1].

Due to the high cost of conducting animal experiments and clinical trials for a new drug [2], a large number of machine learning-based methods have been widely used in DTI prediction in recent years, and the cost of drug development has been greatly reduced through rapid screening of potential drug-target combinations [3, 4].

Existing machine learning-based methods often use the features of drugs and targets for prediction [5, 6]. They treat the prediction problem as a binary classification problem [7]. Drug-target pairs with interaction are considered positive samples, while pairs without interaction are treated as negative samples. The output of the binary classification is the label with higher prediction probability [8–10]. Bleakley and Yamanishi used a support vector machine (SVM) framework based on bipartite local models (BLM) to predict DTIs [11]. Mei et al. improved the original DTI prediction framework by integrate neighbor-based interaction-profile inferring (NII) into the existing BLM method [12]. Buza and Peška extended the BLM method to predict DTIs by using the hubness-aware regression technique [13]. Laarhoven et al. proposed a Gaussian interaction profiling (GIP) kernel to represent the interactions between drugs and targets [14] and then integrated the weighted nearest neighbor method into it to predict DTIs [15]. Chen et al. proposed a Random Walk with Restart-based method on the heterogeneous network to infer potential DTI [16]. Some studies constructed a heterogeneous network which integrates diverse drug-related information to predicted DTI [17, 18]. Thafar et al. utilized graph embedding for DTI prediction [19]. Zhao et al. integrated graph convolutional network and Deep Neural Network to predict DTI [20]. Since the number of positive samples is small, the machine learning-based methods can easily learn to predict unknown samples as negative to reduce the training penalty [3]. Recommendation system is aimed at obtaining accurate prediction results of unknown data even with a small amount of observed data. Considering the problem of data sparseness, learning to rank (LTR) in the recommendation system is able to accurately predict even with a small amount of known data. Therefore, in this study, we defined the DTI prediction problem as a ranking problem. The following paragraph introduces how we define the DTI prediction problem as a ranking problem.

LTR implies a scoring mechanism in which interacting drug-target pairs should have a higher score than those without interaction. In this way, samples with higher scores are treated as interacting drug-target pairs [21, 22]. Recently, there are some studies that apply the idea of LTR to predict DTI [23, 24]. Bagherian et al. showed that matrix factorization algorithms have outperformed other methods in DTI prediction [25]. Thus, we utilized matrix factorization of LTR to predict DTI in this study. Bayesian Personalized Ranking (BPR) which is a matrix factorization of LTR approach has been shown to be an excellent approach for various preference learning tasks even when data are sparse [26, 27].

However, the existing methods do not effectively combine the features of drug and target with the matrix factorization method. Thus, in this study, we propose a DTI prediction model in which BPR is the core and combined gene expression to improve the prediction performance. In the proposed model, the principle of ordering is that interacting drug-target pairs (i.e., positive samples) should be ranked before noninteracting drug-target pairs (i.e., negative samples). Firstly, a set of ternary partial orders is generated based on the positive samples and the negative samples. The set is divided into a training set and a test set. Next, the Adversarial Bayesian Personalized Ranking (ABPR) method is used to train the latent factors of drugs and targets, and the drug-drug similarity and target-target similarity are calculated based on their features, respectively, to improve the training of the latent factors. Finally, for each drug, the inner product of drug's latent factor and target's latent factor is used as the score for ranking. The top-ranked drug-target pairs are predicted with interaction, and the bottom-ranked drug-target pairs are predicted without interaction. This study has the following three contributions:

(i) Aiming at the existing problem of DTI prediction, the idea of matrix factorization of LTR is introduced to process a sparse matrix

(ii) BPR is not robust and vulnerable to adversarial perturbations on its parameters [28]. Perturbation factors are introduced to make the model more robust

(iii) This study also uses the drug and target expression profiles to calculate the drug-drug and target-target similarity, respectively, to improve the training of latent factors

Experimental results show that our method is significantly better than the traditional DTI prediction methods, such as Deep Neural Network (DNN) [8, 29], Generalized Matrix Factorization (GMF) [30], and other state-of-the-art LTR methods, like Neural Matrix Factorization (NeuMF) [30] and Adversarial Matrix Factorization (AMF) [28].

2. Data and Definition

2.1. Data Source. The Library of Integrated Network-Based Cellular Signatures (LINCS) project is a mutual fund project administered by the National Institutes of Health (NIH). This project uses L1000 technology to generate approximately one million gene expression profiles [31]. The L1000 technology uses the correlation between gene expressions to drastically reduce the amount of gene expression that needs to be measured, from more than 20,000 to 978. In this study, we use the drug perturbation and gene knockout transcriptome data from seven cell lines including A375, A549, HA1E, HCC515, HEPG2, PC3, and VCAP. There are three reasons to choose drug perturbation and gene knockout transcriptome data as feature data of drugs and targets: (1) both drug perturbation and gene knockout transcriptome data are from LINCS project and are processed by using L1000 technology. So they are naturally suited to be combined as the feature data. (2) There is a correlation between drug perturbation transcriptome data and the drug's target gene knockout transcriptome data. Pabon et al. have verified in their work that drug perturbation-induced mRNA expression profile correlates with the knockout-induced mRNA expression profile of the drug's target gene and/or genes on the same pathway(s) [32]. The correlation reveals drug-target interactions. Therefore, the correlation based on the expression profile suggests that we can treat the expression profiles as feature data for dual similarity regularization. (3) Transcriptome data can capture the complexity of drug activity in cells. So the use of information obtained from transcriptional profiling studies has a huge impact on multiple areas of the drug discovery including target identification, validation, compound selection, pharmacogenomics, biomarker development, clinical trial evaluation, and toxicology [33].

DrugBank is a comprehensive, freely available web resource containing detailed drug, drug-target, drug action, and drug interaction information about FDA-approved drugs as well as experimental drugs going through the FDA approval process [34]. To obtain complete DTI data, PubChem ID is used as the identifier of drug in the DrugBank and LINCS databases.

The data volume for the seven cell lines is listed in Table 1. The positive drug-target interactions from DrugBank are used to generate interacting drug-target pairs. To avoid treating unknown drug-target interactions in DrugBank as negative interactions, we constructed the nontarget

Table 1: Data volume of seven cell lines.

Cell line	Drug	Target	Nontarget	Interacting drug-target pair	Noninteracting drug-target pair
A375	520	363	2,754	796	1,432,080
A549	525	366	2,648	805	1,390,200
HA1E	533	372	2,707	818	1,442,831
HCC515	471	334	2,516	689	1,185,036
HEPG2	370	356	2,520	557	932,400
PC3	643	378	2,866	963	1,842,838
VCAP	521	377	3,003	809	1,564,563

set that any member of this set has no interaction record with any drug from the same cell line in DrugBank. That means the pair of a nontarget and a drug from the same cell line could be more likely to be treated as a negative sample.

2.2. Problem Definition. In this study, DTI prediction is defined as a ranking problem of drug-target scores.

Definition 1. $D^{\alpha} = \{d_1^{\alpha}, d_2^{\alpha}, d_3^{\alpha}, \cdots, d_m^{\alpha}\}$ represents the set of m drugs in cell line α, where $d_i^{\alpha} = \{d_{i,1}^{\alpha}, d_{i,2}^{\alpha}, \cdots, d_{i,978}^{\alpha}\}$ represents the expression profile of i-th drug.

Definition 2. $T^{\alpha} = \{t_1^{\alpha}, t_2^{\alpha}, t_3^{\alpha}, \cdots, t_n^{\alpha}\}$ represents the set of n targets and nontargets in cell line α, where $t_j^{\alpha} = \{t_{j,1}^{\alpha}, t_{j,2}^{\alpha}, \cdots, t_{j,978}^{\alpha}\}$ represents the expression profile of j-th target or nontarget.

Definition 3. Y^{α} represents the interaction relationship, and $y_{i,j}^{\alpha} \in \{0, 1\}$. If $y_{i,j}^{\alpha}$ =1, the pair of the drug d_i^{α} and target t_j^{α} is a positive sample; otherwise, $y_{i,j}^{\alpha} = 0$, and the pair of d_i^{α} and t_j^{α} is a negative sample.

As shown in Table 1, the numbers of drugs, targets, and interacting drug-target pairs in this study are all limited (for each cell line). Therefore, Y^{α} is a small-sized sparse matrix.

All combinations of drug and target with interactions in each cell line are used as positive samples; all drug and nontarget combinations are used to construct a negative sample candidate set. Since the number of negative samples is much larger than the number of positive samples in each cell line, we randomly sampled some negative samples from the negative sample candidate set to ensure that the number of selected negative samples is consistent with the number of positive samples within the same cell line.

Based on the known relationships of drug-target pairs, the score of drug-target pairs is sorted. The drug-target pairs with higher scores are more likely to interact. Conversely, the drug-target pairs with lower scores are more likely not to interact. Therefore, we transformed the DTI prediction problem into a problem that finds out a reasonable ranking strategy for a drug-target pair. In this paper, the methods are discussed in the same cell line, so the superscript α is omitted.

3. Methods

The proposed method (AdvB-DTI) is based on the method of BPR. Firstly, according to the interaction relationship Y, a ternary partial order set is generated as $H = \{H_i \mid 1 \le i \le m\}$, where $H_i = \{(d_i, t_j, t_k) \mid d_i \in D, t_j \in T, t_k \in T, y_{i,j} \in Y, y_{i,k} \in Y, y_{i,j} = 1, y_{i,k} = 0\}$. H_i combines the target t_j of one positive sample and the target t_k of the corresponding negative sample with the same drug d_i into a partially ordered triple (d_i, t_j, t_k), which means that (d_i, t_j) should be ranked before (d_i, t_k). Then, H is divided into two parts, the training set and test set. Next, based on the training set, BPR is used to train the latent factor matrix of drugs and targets (nontargets). F^D represents the latent factor matrix of the drug ($F^D \in \mathbb{R}^{m \times f}$, f is the size of latent factor), F^T represents target (nontarget) latent factor matrix ($F^T \in \mathbb{R}^{n \times f}$, f is the size of latent factor). Among them, $F_i^D \in \mathbb{R}^{1 \times f}$ represents the latent factor of drug d_i, and $F_j^T \in \mathbb{R}^{1 \times f}$ represents the latent factor of target (nontarget) t_j. $r_{i,j} = F_i^D \bullet F_j^T$ is the predicted score for ranking the interaction of d_i and t_j.

In order to improve the training of latent factors, we use the dual similarity regularization method based on the similarity theory to increase the latent distance between latent factors to increase the gap between the scores of different drug-target pairs.

Finally, gene expression data of LINCS project were treated as the features of drugs and targets to calculate drug-drug similarity and target-target similarity to improve training latent factors which represented key features of gene expression. Because the gene expression data are the observed values obtained from experiment, thus, the error between the observed value and the true value does exist. Therefore, latent factors of the drug and target (i.e., the model parameters) learned in this study can fluctuate within a certain range but the model's prediction results should be stable. Consequently, the perturbation factor Δ is introduced into the training process of F^D and F^T to make the trained model more robust. The overall process of model training is shown in Figure 1.

After the model is trained, calculate the value of $r_{i,j}$ for all drug-target pairs, and sort them in a descending order. The top-ranked drug-target pairs are predicted as the interaction, and the bottom ranked drug-target pairs are predicted as the noninteraction. The prediction process is shown in Figure 2. Next, we will introduce the related methods in detail.

3.1. Bayesian Personalized Ranking. BPR is a pairwise LTR method. It learns in an implicit feedback manner through personalized ranking and is widely used in the recommendation systems [26].

As shown in Table 1, the numbers of drugs, targets, and interacting drug-target pairs in this study are all limited (for each cell line). Since one partially ordered triple was generated based on one positive sample and the corresponding negative sample, the number of partially ordered triples is also limited. Therefore, what we faced in this study were not only a small amount of partially ordered triples but also

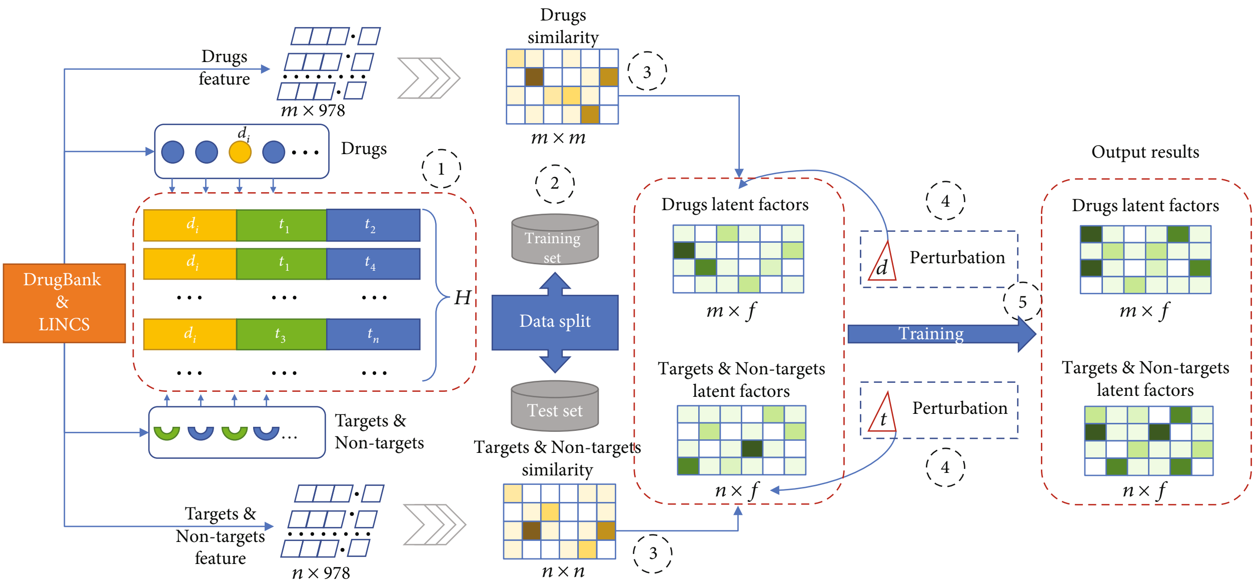


Figure 1: The flowchart of model training. (1) Generating ternary partial order set H. (2) Splitting H into a training set and a test set. (3) Calculating drug-drug and target-target similarity for improving latent factors. (4) Perturbation of latent factors for BPR. (5) Latent factor training.

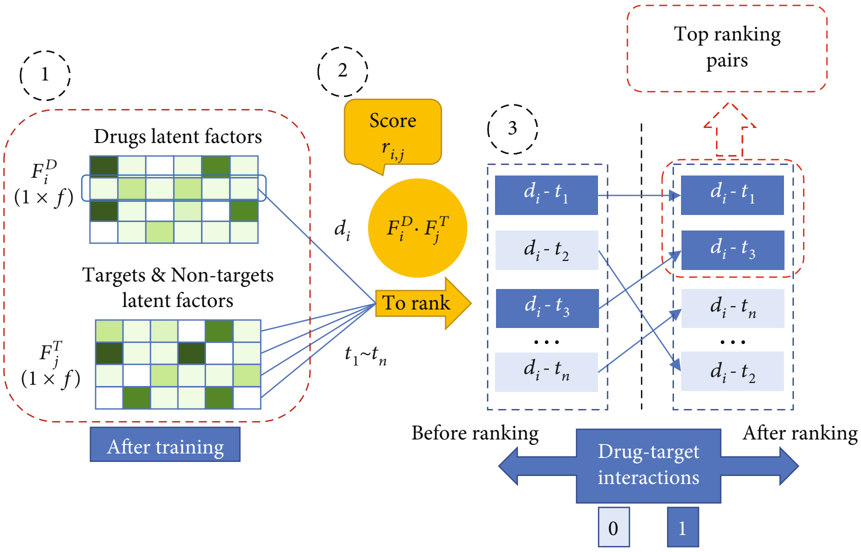


Figure 2: The flowchart of model prediction. (1) Latent factor matrix of F^D and F^T after training. (2) Calculating $r_{i,j}$ for ranking. (3) Ranking drug-target pairs.

high-dimensional data. BPR is able to accurately predict even with a small amount of known data [26]. And BPR could map both drugs and targets into a shared low-dimensional latent feature space and to use this representation to calculate the probability of drug-target interactions to overcome the problem of high dimensionality [27].

According to the study of [26], BPR was derived for solving the personalized ranking task that only positive observations are available. In the problem of DTI prediction, only positive drug-target interactions can be directly obtained from the DrugBank database which is a key challenge in the DTI prediction problem. Hence, these advantages make BPR suitable for the DTI prediction problem.

In this study, we use this method to rank the score of drug-target pairs.

For H_i of $d_i (1 \le i \le m)$, we have

$$p\left(\theta \mid t_j >_{d_i} t_k\right) \propto p\left(t_j >_{d_i} t_k \mid \theta\right) p(\theta), \qquad (1)$$

where θ denotes the parameters of the model and $t_j >_{d_i} t_k$ denotes that for d_i the possibility of interacting with t_j is greater than the possibility of interacting with t_k. Since the interaction of d_i and t_j has no interference on the interaction of d_i and t_k, all drug-target interactions are independent. The likelihood estimates for parameter θ are

$$\prod_{\left(d_i,t_j,t_k\right)\in H_i} p\left(t_j >_{d_i} t_k \mid \theta\right). \tag{2}$$

In order to calculate $p(t_j >_{d_i} t_k \mid \theta)$, we use the logistic sigmoid function [26]:

$$p\left(t_j >_{d_i} t_k \mid \theta\right) = \sigma\left(r_{i,j} - r_{i,k}\right), \tag{3}$$

where $\sigma(\bullet)$ is the logistic sigmoid function and $\sigma(x) = 1/(1 + e^{-x})$.

$(r_{i,j} - r_{i,k})$ captures the ranking relation between t_j and t_k with the given d_i. If t_j is more likely to interact with d_i than t_k, then $r_{i,j} \geq r_{i,k}$ and $(r_{i,j} - r_{i,k}) \geq 0$. Otherwise, $(r_{i,j} - r_{i,k}) \leq 0$. Any standard collaborative filtering model can be applied to predict the value of $(r_{i,j} - r_{i,k})$. Matrix factorization has been successfully applied in many studies [35–37]. Thus, the matrix factorization model is used in this study.

Next, consider $p(\theta)$ of formula (1). It is a Gaussian distribution with zero mean and variance-covariance matrix $\lambda_\theta I$ [26], where λ_θ is a model-specific regularization parameter and I is an identity matrix, so

$$p(\theta) \sim N(0, \lambda_\theta \mathrm{I}). \tag{4}$$

According to formulas (2)–(4), the maximum posterior probability of the BPR method can now be rewritten as

$$\begin{aligned}\max_{\theta} \mathscr{L} &= \ln p\left(\theta \mid t_j >_{d_i} t_k\right) = \ln p\left(t_j >_{d_i} t_k \mid \theta\right) p(\theta) \\ &= \sum_{\left(d_i,t_j,t_k\right)\in H_i} \ln p\left(t_j >_{d_i} t_k \mid \theta\right) - \lambda_\theta \|\theta\|^2 \\ &= \sum_{\left(d_i,t_j,t_k\right)\in H_i} \ln \sigma\left(r_{i,j} - r_{i,k}\right) - \lambda_\theta\left(\left\|F^D\right\|^2 + \left\|F^T\right\|^2\right),\end{aligned} \tag{5}$$

where $\|\bullet\|^2$ is an L2 regularization term.

From the maximum likelihood estimation for parameter θ in formula (5), an equivalent optimization objective formula can be obtained:

$$\begin{aligned}\min_{\theta} L_{\mathrm{BPR}}(H_i \mid \theta) &= \sum_{\left(d,t_i,t_j\right)\in H_i} - \ln p\left(t_j >_{d_i} t_k \mid \theta\right) + \lambda_\theta \|\theta\|^2 \\ &= \sum_{\left(d_i,t_j,t_k\right)\in H_i} - \ln \sigma\left(r_{i,j} - r_{i,k}\right) + \lambda_\theta\left(\left\|F^D\right\|^2 + \left\|F^T\right\|^2\right).\end{aligned} \tag{6}$$

3.2. Adversarial Bayesian Personalized Ranking. As mentioned, since the error between the observed value and the true value does exist, in order to enhance the robustness of the model, it is necessary to consider gene perturbations. It is unreasonable to add noise (such as changing the labels of training data) at the input layer. For example, modifying the training data (d_i, t_j, t_k) to (d_i, t_k, t_j) means that the non-interacting drug-target pair (d_i, t_k) is ranked higher than interacting drug-target pair (d_i, t_j). Obviously, the latent factors obtained by such training data are unreasonable. Therefore, it is necessary to add perturbations to the latent factors. For drug and target gene perturbations, we defined it as the perturbation factor that are added to Bayesian Personalized Ranking:

$$\max_{\Delta, \|\Delta\|^2 \leq \varepsilon} L_{\mathrm{BPR}}(H_i \mid \theta + \Delta), \tag{7}$$

where Δ is the gene perturbations on model parameters, ε controls the magnitude of adversarial perturbations, $\|\bullet\|^2$ denotes the L2 norm, and θ denotes the current model parameters (i.e., latent factors).

Δ can be optimal by adversarial perturbations Δ_{adv} as follows [28]:

$$\Delta_{\mathrm{adv}} = \varepsilon \frac{\Gamma}{\|\Gamma\|^2}, \Gamma = \frac{\partial L_{\mathrm{BPR}}(H_i \mid \theta + \Delta)}{\partial \Delta}. \tag{8}$$

Finally, we define the objective function of ABPR as follows:

$$L_{\mathrm{AdvB-DTI}}(H_i \mid \theta) = L_{\mathrm{BPR}}(H_i \mid \theta) + \lambda \Delta_{\mathrm{adv}}, \tag{9}$$

where λ controls the adversarial strength. The training process of AdvB-DTI can be expressed as playing a minimax game:

$$\min_{\theta} \max_{\Delta, \|\Delta\|^2 \leq \varepsilon} L_{\mathrm{BPR}}(H_i \mid \theta) + \lambda L_{\mathrm{BPR}}(H_i \mid \theta + \Delta), \tag{10}$$

where the learning algorithm for model parameter latent factor θ is the minimizing player, which is aimed at obtaining accuracy prediction results. And the perturbation factor Δ acts as the maximizing player, which is aimed at identifying the worst-case perturbations against the current model. Finally, by playing this minimax game, it is able to make the model robust and simulate the error.

3.3. Dual Similarity Regularization. In the process of latent factors training, when drugs or targets are similar, their latent distance should be small. Conversely, when drugs or targets are different, their latent distance should be large. In order to meet this requirement, dual similarity regularization was introduced into this process.

In order to effectively combine the features of drugs and targets with matrix factorization methods, a Gaussian function needs to be introduced. Through this function, the features of drugs and targets can effectively influence the training of latent factors. Zheng et al. made the point that this function is sensitive to the latent distance of similarity between different drugs or targets [38]. The similarity between drugs (or targets) is negatively related to their latent distance. The function is defined as

$$\mathrm{SimGaus}\left(S^D, F^D, d_i\right) = \sum_{j=1}^{m} \left[S^D(i, j) - e^{-\left\| F_i^D - F_j^D \right\|^2}\right]^2,$$
$$S^D(i, j) = S^D(j, i) = \mathrm{Sim}\left(d_i, d_j\right). \quad (11)$$

where S^D denotes drug-drug similarity matrix ($S^D \in \mathbb{R}^{m \times m}$), $\|\bullet\|^2$ denotes latent distance, and $\mathrm{Sim}(\bullet)$ is a similarity calculation method.

Similarly, we can obtain

$$\mathrm{SimGaus}\left(S^T, F^T, t_j\right) = \sum_{k=1}^{n} \left[S^T(j, k) - e^{-\left\| F_j^T - F_k^T \right\|^2}\right]^2,$$
$$S^T(j, k) = S^T(k, j) = \mathrm{Sim}\left(t_j, \mathrm{t}_k\right), \quad (12)$$

where S^T denotes target-target similarity matrix ($S^T \in \mathbb{R}^{n \times n}$).

Commonly used similarity calculation methods include cosine similarity, Tanimoto coefficient, structural similarity index, and Spearman's rank correlation coefficient.

Tanimoto coefficient is an extension of Intersection over Union. It can be used to measure the similarity of nonbinary features. It calculates the degree of correlation based on the magnitude of the feature vector. The closer the calculation result is to 1, the more similar the two vectors are. It is defined as

$$T(x, y) = \frac{xy}{\|x\|^2 + \|y\|^2 - xy}. \quad (13)$$

Cosine similarity is determined by the angle between two vectors. The smaller the angle is, the more similar the two vectors are. It is defined as

$$\cos(x, y) = \frac{xy}{\|x\|\|y\|}. \quad (14)$$

Structural similarity index is a common similarity calculation method used in computer vision to measure image quality [39]. It is defined as

$$\mathrm{SSIM}(x, y) = \frac{\left(2\mu_x\mu_y + c_1\right)\left(2\sigma_{xy} + c_2\right)}{\left(\mu_x^2 + \mu_y^2 + c_1\right)\left(\sigma_x^2 + \sigma_y^2 + c_2\right)}, \quad (15)$$

where μ is the mean, σ^2 is the variance, σ_{xy} is the covariance, and $c_1 = 0.001$ and $c_2 = 0.001$ are constants to avoid the denominator being 0. The closer the calculation result is to 1, the more similar the two vectors are. Since technologies originating from computer vision have been widely used in DTI prediction in recent years, we attempt to use these methods to calculate the similarity between drugs and targets. Originally, μ is used as an estimate of the image brightness, σ^2 is an estimate of the image contrast, and σ_{xy} is the measure of the similarity of the image structure. In our problem, μ is used as an estimate of the amount of change in gene expression, σ^2 is used as an estimate of the relative change in gene expression, and σ_{xy} is used as an estimate of the change trend in gene expression.

Spearman's rank correlation coefficient is a similarity calculation method based on the ranking of feature data. It is defined as

$$\mathrm{sprm}(x, y) = 1 - \frac{6\sum_1^n g_i^2}{n\left(n^2 - 1\right)}, \quad (16)$$

where g_i is the difference in the ranks of x_i and y_i and the size of features is n. For example, if $x = (1, 0, 3)$ and $y = (1, 5, 2)$, then the rank of $x = (2, 1, 3)$ and $y = (1, 3, 2)$, thus $g = (1, -2, 1)$. Similarly, the closer the similarity value is to 1, the more similar the two vectors are.

Because the Gaussian function is a numerically "sensitive" function, which means it can increase the impact of similarity on latent factor training. Thus, it can extend the latent distance between drugs (or targets) to increase the scores of different $\left(r_{i,j} - r_{i,k}\right)$, which is to increase the penalty for wrong rankings and optimize the training latent factors.

We use stochastic gradient descent to optimize the final objective formula:

$$\min_{\theta} \max_{\Delta, \|\Delta\|^2 \le \varepsilon} \sum_{H_i \subseteq H, \left(d_i, t_j, t_k\right) \in H_i} L_{\mathrm{BPR}}(H_i \mid \theta) + \lambda_{\mathrm{adv}} \mathrm{L}_{\mathrm{BPR}}(H_i \mid \theta + \Delta)$$
$$+ \lambda_{\mathrm{sim}} \left[\mathrm{SimGaus}\left(S^D, F^D, d_i\right) + \mathrm{SimGaus}\left(S^T, F^T, t_j\right) + \mathrm{SimGaus}\left(S^T, F^T, t_k\right)\right], \quad (17)$$

where λ_{adv} and λ_{sim} are adversarial and similar hyperparameters, respectively.

4. Experiment and Analysis

The experiments are designed to answer the following three questions:

(i) How do different similarity calculation methods affect the prediction results of the model?

(ii) How do different numbers of latent factors, λ_{sim} and λ_{adv}, impact the model's performance?

(iii) Will our model (AdvB-DTI) outperform other prediction models?

4.1. Assessment Metrics. The assessment metrics used in the experiment are AUC [26], Top_k [40], and AUPR. AUC is defined as formula (18):

$$\mathrm{AUC} = \frac{1}{|D|} \sum_{d_i \epsilon D} \frac{\left|\left\{\left(d_i, t_j, t_k\right) \mid r_{i,j} > r_{i,k}, t_j \in T, t_k \in T, y_{i,j} = 1, y_{i,k} = 0\right\}\right|}{|H_i|}. \quad (18)$$

The set of interacting drug-target pairs is called the positive set, and the set of noninteracting drug-target pairs is called the negative set. One drug-target pair is randomly selected from the positive set and the negative set, respectively. AUC means the probability that the model correctly predicts that the score of the drug-target pair from the positive set is larger than that of the drug-target pair from the negative set. AUC can better reflect the overall performance of the model. The larger the value of AUC is, the better the performance of the model is.

Topk_i means for drug d_i, among the k top-ranked drug-target pairs, the proportion of targets that interact with d_i in all the targets that interact with d_i, which is defined as

$$Top\_k_i = \frac{\left|\left\{t_j \middle| \left|\left\{t_l \mid r_{i,j} \leq r_{i,l}, \forall t_l \in T, l \neq j\right\}\right| \leq k-1, \forall t_j \in T, y_{i,j}=1\right\}\right|}{\left|\left\{t_j \middle| \forall t_j \in T, y_{i,j}=1\right\}\right|}. \tag{19}$$

Top_k is the average of all Top_k_i $(1 \leq i \leq m)$. This assessment metric is equivalent to the recall rate. Top_k is defined as

$$\text{Top\_}k = \frac{1}{|D|} \sum_{d_i \epsilon D} \text{Top\_}k_i. \tag{20}$$

The meaning of prec_k_i is, for drug d_i, among the k top-ranked drug-target pairs, the proportion of targets that interact with d_i. Its definition is shown in

$$prec\_k_i = \frac{\left|\left\{t_j \middle| \left|\left\{t_l \mid r_{i,j} \leq r_{i,l}, \forall t_l \in T, l \neq j\right\}\right| \leq k-1, \forall t_j \in T, y_{i,j}=1\right\}\right|}{k}. \tag{21}$$

prec_k is the average of all prec_$k_i(1 \leq i \leq m)$. This assessment metric is equivalent to the precision rate. prec_k is defined as

$$prec\_k = \frac{1}{|D|} \sum_{d_i \epsilon D} prec\_k_i. \tag{22}$$

With different k values, drug d_i has different (Top_k_i, prec_k_i) pairs. Connecting all (Top_k_i, prec_k_i), we can obtain a curve. The area enclosed by the obtained curve and the coordinate axes is the AUPR$_i$ of d_i. AUPR$_i$ is also a comprehensive assessment metric, which is defined as

$$\text{AUPR}_i = \oiint_{\sigma \in \text{Top\_}k_i - \text{prec\_}k_i \text{ curve}}^{d_i} d\sigma. \tag{23}$$

AUPR calculates the average of all AUPR$_i(1 \leq i \leq m)$. The closer the value is to 1, the better the model performance. It is defined as

$$\text{AUPR} = \frac{1}{|D|} \sum_{d_i \epsilon D} \text{AUPR}_i \tag{24}$$

Table 2: The parameters and settings used in the experiments.

Hyperparameter	Setting
factor_size	[5, 10, 15, 20, 25, 30, 40, 50, 60]
λ_{sim}	[0,0.3,0.5,0.9,1.25]
λ_{adv}	[0,0.3,0.5,0.9]
ε	0.1
λ_θ	0.1
learning rate	0.03

4.2. Results and Analysis. We adopted 5-fold nested cross-validation to evaluate the performance of the proposed method, which means that when analyzing the impact of hyperparameters, we only utilized the training set. For fair comparison, we tuned the parameters of each method so that they could achieve the best performance in comparison. The hyperparameters used in the experiments and their values are listed in Table 2.

Matrix factorization methods demonstrated their power and versatility in bioinformatics, for example, in the prediction of disease subtype alignment [41], drug repositioning [42], and protease target prediction [37]. Thus, we treat a state-of-the-art method which predicts DTI via DNN [8] as baseline and compare it with other state-of-the-art matrix factorization methods [28, 30].

4.2.1. Comparative Experiment of Different Similarity Calculation Methods. Table 3 lists the results of comparative experiments of different similarity calculation methods performed independently in the seven cell lines. Four different methods were used for comparison.

From Table 3, it can be found that the prediction results of Tanimoto coefficient are better than those of the other three methods in seven cell lines. The performance based on Spearman's rank correlation coefficient is second to that of the Tanimoto coefficient in this experiment, and they are very close. The traditional cosine similarity calculation method was unstable in the experiment, and AUC is under 90% in cell lines A549 and HEPG2. The prediction performance of structural similarity index is similar to that of Spearman's rank correlation coefficient. Except cosine similarity, three similarity calculation methods all consider the value of the features in calculating the similarity. Cosine similarity only considers the angle between vectors. If two feature vectors have the same direction, they are considered similar regardless of value of the features. From the results of cosine similarity, it can be inferred that ignoring feature values may cause poor prediction performance. Therefore, based on the above results, Tanimoto coefficient is more suitable to the prediction problem.

4.2.2. Impact of Different Settings of Hyperparameters. Figure 3 reflects the relationship between the number of latent factors and the result of Top_10. For example, when factor_size = 5, Top_10 ≈ 0.5. It means that ten top-ranked drug-target pairs of a particular d_i predicted by the model

Table 3: The impact of different similarity calculation methods on prediction performance in seven cell lines.

Cell line		Tanimoto	cos	SSIM	sprm
A375	AUC	0.9202	0.9088	0.9037	0.9119
	AUPR	0.9437	0.9160	0.9389	0.9436
A549	AUC	0.9347	0.8944	0.9247	0.9192
	AUPR	0.9477	0.9109	0.9425	0.9367
HA1E	AUC	0.9249	0.9174	0.9082	0.9035
	AUPR	0.9450	0.9401	0.9380	0.9389
HCC515	AUC	0.9163	0.9018	0.9045	0.9045
	AUPR	0.9403	0.9332	0.9377	0.9305
HEPG2	AUC	0.9259	0.8828	0.9144	0.9124
	AUPR	0.9303	0.9161	0.9249	0.9279
PC3	AUC	0.9306	0.9090	0.9116	0.9228
	AUPR	0.9581	0.9471	0.9459	0.9536
VCAP	AUC	0.9466	0.9102	0.9349	0.9349
	AUPR	0.9645	0.9558	0.9453	0.9543

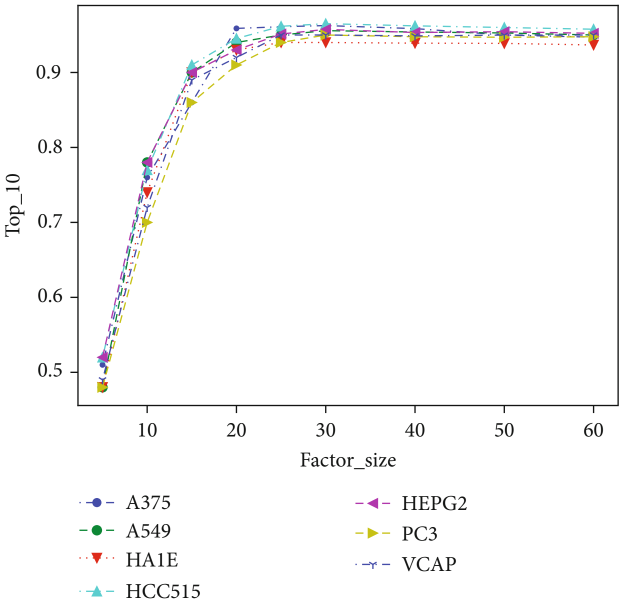


Figure 3: Impact of different numbers of latent factors on Top_10. Top_10 increases with factor_size and tends to be stable after factor_size is greater than 25.

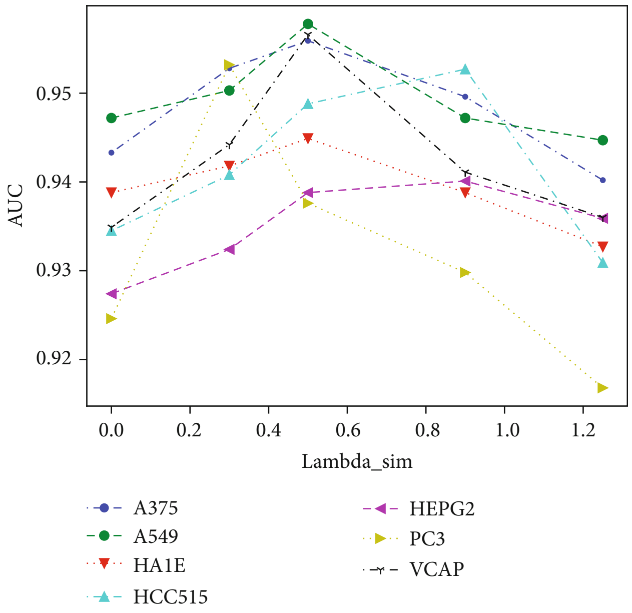


Figure 4: Impact of λ_{sim} on AUC. AUC increases with λ_{sim} but decreases when λ_{sim} is greater than a critical value.

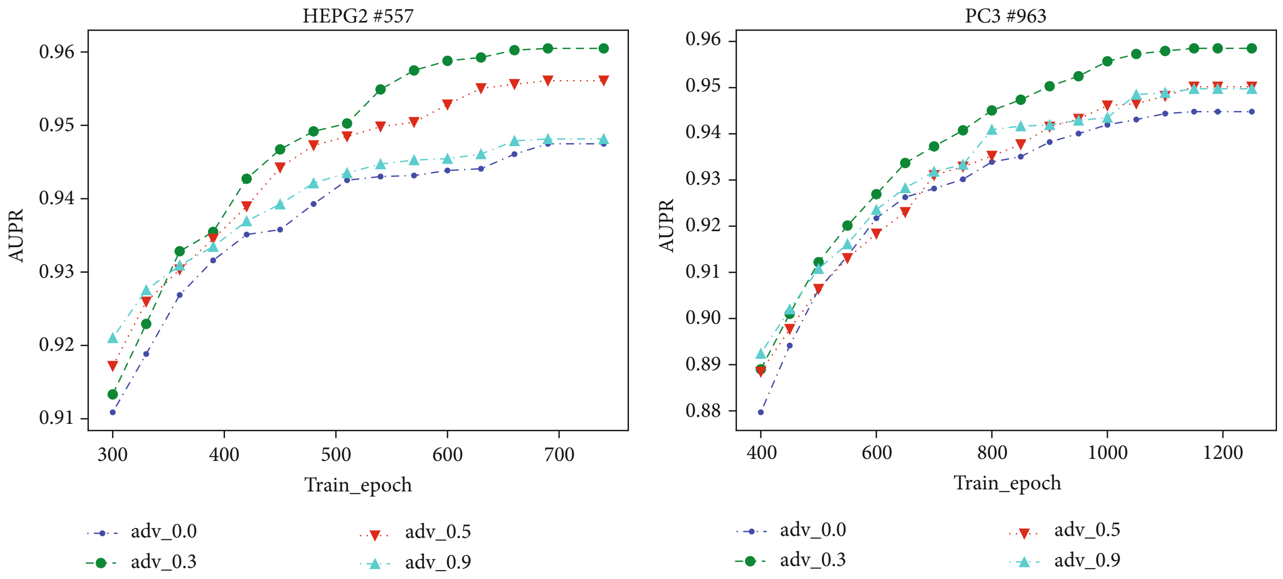
contain about half of all interacting drug-target pairs of this drug (i.e., the recall rate is about 0.5). The meaning of latent factors is to map high-dimensional feature vectors to low-dimensional latent space and capture the implicit features of gene expression. The larger the size of the low-dimensional latent space, the more sufficient the feature information of the original high-dimensional drug and target expression can be that can be extracted. That is why the value of Top_10 significantly rises with the increase of the latent factor size. As shown in Figure 3, when the size of the latent factor increases to a critical size (e.g., factor_size > 25), the feature information is almost completely extracted, and the performance of AdvB-DTI becomes stable.

Figure 4 shows the impact of λ_{sim} on the values of AUC. When dual similarity regularization was not used (i.e., $\lambda_{\text{sim}} = 0$), the values of AUC are lower than those using this method, which indicates that the method can improve the prediction performance.

Firstly, how does dual similarity regularization improve the training of latent factors? $r_{i,j}$ is the score to rank. The ranking interval between different drug-target pairs is calculated by the difference of different scores. If λ_{sim} is set to a larger value, the latent distance between the drug and the target will also become large, and the same thing happens to different scores. Therefore, making the interval between different drug-target pairs increase will aggravate the penalty for the model when ranking errors occur during the training process. Thus, dual similarity regularization improves the training of latent factors.

Secondly, how to select a proper value for λ_{sim}? The difference in $r_{i,j}$ between different drug-target pairs increases with λ_{sim}. Thus, the interval between different rankings increases. In cell lines with fewer positive samples, the model parameter θ will not be too large and increasing λ_{sim} can effectively improve the prediction performance. However, in cell lines with more positive samples, increasing λ_{sim} means that θ needs to increase beyond the limit of its regular term $\|\theta\|^2$, so the model will be underfitting and the value of AUC decreases, as shown in Figure 4. AUC increases with λ_{sim} but decreases when λ_{sim} is greater than a critical value.

Therefore, in a cell line with fewer positive samples, a larger λ_{sim} will improve the prediction performance; however, in a cell line with more positive samples, a smaller λ_{sim} is suitable.



(a)

(b)

Figure 5: Impact of λ_{adv} on AUPR. For cell lines HEPG2 and PC3, the best performance of AUPR is achieved when $\lambda_{\text{adv}} = 0.3$.

In HEPG2 cell line, the number of positive samples is the smallest among the 7 cell lines. In PC3 cell lines, the number of positive samples is the largest among 7 cell lines. Therefore, in this experiment, we select these two cell lines as representatives to study the impact of λ_{adv} on prediction performance. In Figures 5(a) and 5(b), the curve of $\lambda_{\text{adv}} = 0$ represents that ABPR was not used in the model, and the other curves represent that ABPR was used in the model. In the early stages of training, the values of AUPR by using ABPR are better than those by not using ABPR. This is because when using ABPR, the parameters of the model could change within a certain range without changing the past prediction results, that is, learning new knowledge without forgetting the knowledge learned in the past. Thus, the prediction performance of the model can be effectively and quickly improved in the early stages of model training. Using ABPR as far as possible, the better performance will be obtained in the early stage of training.

Because of using Dual Similarity Regularization, the difference of scores of different drug-target pairs will increase; that is, the model parameters can withstand a certain range of perturbations to improve the model prediction performance. However, when the value of λ_{adv} exceeds a certain range, due to the constraints of the regular terms of the model parameters, they cannot resist excessive perturbations, which leads to the model being underfitted. Therefore, if λ_{adv} is given a large value, the model converges fast. The upper bound of model convergence depends on the ability of model parameters to resist the perturbations, which can be verified in the PC3 cell line. As shown in Figures 5(a) and 5(b), the larger λ_{adv} is, the lower the upper bound of model convergence. When $\lambda_{\text{adv}} = 0.3$, the model obtained the best prediction performance.

4.2.3. Comparison with Other Methods. AdvB-DTI was compared with other state-of-the-art methods, and the prediction performances are listed in Table 4. The comparison methods include DNN [8], GMF [30], NeuMF [30], and AMF [28].

Xie et al. used a DNN framework [8] for DTI prediction based on transcriptome data in the L1000 database gathered from drug perturbation and gene knockout trials. We used the same configurations for DNN training.

NeuMF [30] is a deep learning matrix factorization framework for recommendation task with implicit feedback. In this method, DNN's input layer is defined as a latent vector instead of drug and target features. It is an improvement of GMF and DNN. To compare with NeuMF and GMF fairly, our model uses the same number of latent factors as NeuMF and GMF.

AMF [28] is a state-of-the-art approach designed for item recommendation with users' implicit feedback. It introduces the concept of ABPR and improves the method of BPR [26].

The results of DNN are used as baseline in Table 4. Since the DTI data are too sparse that each drug only has interactions with few targets, and DNN needs sufficient data for training, the performance of DNN is not attractive. DNN utilizes the transcriptome data as drug and target's feature. However, the transcriptome data has much noise, which also limits its performance. As shown in Table 4, other state-of-the-art matrix factorization methods' performances are better than that of the baseline.

When comparing AdvB-DTI with other state-of-the-art matrix factorization methods (NeuMF, GMF, and AMF), we could observe that only utilizing the relationship of drug and target could not guarantee an ideal prediction performance and efficiently exploiting the similarity of drug-drug and target-target will has a positive impact on the performance.

Notice that the performance of AMF is only second to that of AdvB-DTI. It demonstrates that adding perturbations to latent factors could make model learn noise, rather than utilize noise data to train model like DNN. That is the reason

Table 4: Comparison between AdvB-DTI and other methods.

Cell line		DNN	GMF	NeuMF	AMF	AdvB-DTI
A375	AUC	0.8984	0.8733	0.9013	0.9253	0.9564
	AUPR	0.8673	0.8385	0.8805	0.9350	0.9635
A549	AUC	0.9134	0.8927	0.9071	0.9246	0.9554
	AUPR	0.8724	0.8495	0.8986	0.9319	0.9673
HA1E	AUC	0.8938	0.8874	0.9052	0.9074	0.9428
	AUPR	0.8518	0.8424	0.8837	0.9137	0.9602
HCC515	AUC	0.8735	0.8912	0.8899	0.9009	0.9571
	AUPR	0.8259	0.8429	0.8493	0.9177	0.9654
HEPG2	AUC	0.8901	0.8742	0.8835	0.8896	0.9464
	AUPR	0.8135	0.8135	0.8297	0.8951	0.9624
PC3	AUC	0.8957	0.8774	0.8725	0.9205	0.9560
	AUPR	0.8647	0.8631	0.8538	0.9309	0.9632
VCAP	AUC	0.8975	0.9033	0.8920	0.9095	0.9556
	AUPR	0.8426	0.8388	0.8749	0.9126	0.9622

Table 5: Comparison of AdvB-DTI and AMF based on NDCG in seven cell lines.

Cell line	AdvB-DTI	AMF
A375	0.9469	0.9149
A549	0.9413	0.9136
HA1E	0.9373	0.8813
HCC515	0.9455	0.8951
HEPG2	0.9566	0.8854
PC3	0.9517	0.9098
VCAP	0.9535	0.9041

that AMF could achieve a better performance than other models except AdvB-DTI.

NDCG is mainly used for evaluating ranking methods [43]. As our model is a ranking method, we compared AdvB-DTI with AMF, which has the best performance in Table 4 except AdvB-DTI, as shown in Table 5. It can be seen from the results that AdvB-DTI outperforms AMF and it is verified that AdvB-DTI can effectively deal with the class imbalance problem and the problem of data sparsity.

Finally, we compared the computing resource consumption of these methods. All the algorithms were written using Python programming language and operated on a computer (Ubuntu 16.04.4 LTS, Core i9-7900X CPU, 3.3 GHz, 128 GB memory space). The algorithms were executed by CPU. We conducted 10 experiments in the cell line of A549, and each experiment concurrently executed 10 training procedures with 5-fold cross-validation. The average results are shown in Table 6.

It can be found that DNN has the largest memory cost because of its many parameters. GMF is a traditional matrix decomposition framework with simple structure and few parameters, so its memory cost is minimum. NeuMF is the framework of matrix decomposition combined with neural network, so its memory cost is slightly higher than that of GMF. AdvB-DTI improves AMF and NeuMF improves GMF. Comparing the two groups of models based on Tables 4 and 6, it can be found that the convergence time of the model is related to its final prediction performance, and the improvement of model performance may lead to the increase of training time. In addition, the neural network-based methods, such as DNN and NeuMF, take up a lot of CPU resources.

Table 6: Resources consumed by AdvB-DTI and other methods in the cell line of A549.

Method	Time (m) ↓	Memory (MB) ↓	CPU (%) ↓
DNN	5	518	33.8
GMF	5	80	36.4
NeuMF	6	101	44.7
AMF	7	230	5.7
AdvB-DTI	12	180	5.3

In summary, AdvB-DTI efficiently utilizes the similarity of drug-drug and target-target and the relationship of drugs and targets to train latent factors for drugs and targets to improve DTI prediction performance.

5. System Analysis of AdvB-DTI

After the comparison with other methods, we utilize top 1% of all the prediction results to demonstrate the strength of our method to predict novel DTIs. In order to verify our model, all the known DTIs which have been utilized in our model are removed for discussion in this section and the following analysis is in A375.

5.1. Examination of Results. To validate whether our prediction results are in accord with current knowledge, we examined the predicted DTIs using other DTI database, including TTD [44], IUPHARBPS [45], Matador [46], STITCH [47], DGIdb [48], and CTD [49].

We used $r_{i,j}$ to rank all predicted DTIs and calculated pair counts that overlap between the predicted results and the interactions from other databases. Then, we counted the number of overlapping pairs in the sliding bins of 500 consecutive interactions (as shown in Figure 6). It suggests that our model can predict novel DTIs validated by known knowledge in other databases. Considering that DTIs in CTD database are curated from the published literature, these interactions are both direct (e.g., "chemical binds to protein") and indirect (e.g., "chemical results in increased phosphorylation of a protein" via intermediate events); it is reasonable that CTD database covers a wider variety of drug-target interactions than other DTI databases.

5.2. Enrichment Analysis. In this study, the DrugBank database is considered the gold standard. The drug-target interactions from the DrugBank database are the most accurate and strict drug-target interactions. Besides the DrugBank database, there are some other databases containing a large amount of drug-target interaction data. These drug-target interaction data are much larger than the gold standard we used. Therefore, we compare our prediction results with the

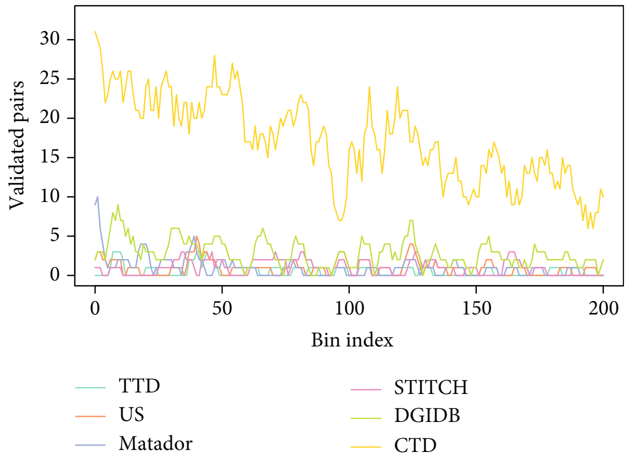


Figure 6: The overlap curves between predicted interactions and known DTIs.

drug-target interactions contained in these databases. Here, the drug-target interactions in the IUPHARBPS database, STITCH database, CTD database, TTD database, Matador database, and DGIdb database were used. If our prediction results appear in other databases, it indicates that our prediction results are consistent with prior knowledge.

In order to characterize and quantify the appearance of predicted drug-target relationships (and known drug-target interactions) in other databases, we used the enrichment score and P value.

We calculated enrichment score (ES) as follows:

$$ES = \frac{kN}{nm}, \tag{25}$$

where k is the number of predicted drug-target interactions that appear in the specified database (or the number of known drug-target interactions (i.e., drug-target interactions in our gold standard) that appear in the specified database); N is the number of all possible interactions between the drug set and the target set, that is, the drug-target interactions when the drug set and the target set are fully connected; n is the number of predicted drug-target interactions (or the number of known drug-target interactions in our gold standard); and m is the number of drug-target interactions in a specific database. And the interactions mentioned above only concern drugs and targets present in the gold standard.

Then, we used the hypergeometric distribution to calculate the P value as follows:

$$P(X \geq k) = \sum_{x=k}^{\infty} \frac{(m/x)(N - m/n - x)}{(N/n)}. \tag{26}$$

FDR correction is used to correct the P values for multitesting [50].

As shown in Table 7, the known drug-target interactions and the drug-target interactions predicted using AdvB-DTI are significantly enriched on other datasets except for the STITCH database. Obviously, the known drug-target interactions (drug-target interactions in our gold standard) have larger enrichment scores and smaller P value than predicted drug-target interactions.

Table 7: Enrichment of drug-target interactions on other datasets.

	ES	PES	EP-Value	PEP-Value
TTD	107.91	3.60	292.06	1.20
STITCH	12.32	0.52	16.72	0.04
DGIdb	70.37	2.43	∞	2.88
CTD	9.18	1.73	134.46	6.10
Matador	59.28	5.87	131.13	6.10
IUPHARBPS	99.74	3.84	856.72	2.33

ES: enrichment score of known drug-target interactions; PES: enrichment score of predicted drug-target interactions; EP-Value: enrichment P value (after -lg10) of known drug-target interactions; PEP-Value: enrichment P value (after -lg10) of predicted drug-target interactions.

The results indicate that the drug-target interactions predicted by AdvB-DTI can be verified on other DTI datasets and have a potential practical value.

5.3. Drug Treatment Property. Drug ATC (Anatomical Therapeutic Chemical) label, which reflects drugs' therapeutic, pharmacological and chemical properties, is an important label of drugs. By comparing the distribution of drug ATC label in the known drug-target interactions and that of drug ATC label in the predicted drug-target interactions, we can find out which type of drug is more likely to be predicted to be associated with targets.

The distribution of drug ATC label in the known drug-target interactions and that of drug ATC label in the predicted drug-target interactions are illustrated in Figures 7(a) and 7(b). The relative ratio between known and predicted DTIs for each ATC label is shown in Figure 7(c). If there are 25% of drugs with ATC label A in the gold standard and 50% of drugs with ATC label A in the prediction result, the relative ratio is $0.25/0.5 = 0.5$. The smaller the ratio, the more potential the drugs with that specific ATC label has to target proteins. So, the drugs with that specific ATC label should be studied further for broader use.

In Figure 7, the distributions of drug ATC labels for the gold standard and for the predictions (note that only the top 1% of all prediction results are taken) are almost the same. Notably, drugs with ATC label "B" (Blood and Blood Forming Organs) have a low relative ratio. In addition to A375, in most other cell lines, we also predicted more targets for drugs with ATC label "B". The result suggests that drugs with ATC label "B" have more potential to target proteins and should be studied further for broader use.

6. Case Study

To illustrate the reliability of the prediction results of AdvB-DTI, we studied several cases in this section. These examples are all from our prediction results.

Olomoucine (CID: 4592) is a cyclin-dependent kinase inhibitor. For Olomoucine, its predicted target is MAPK3 through AdvB-DTI.

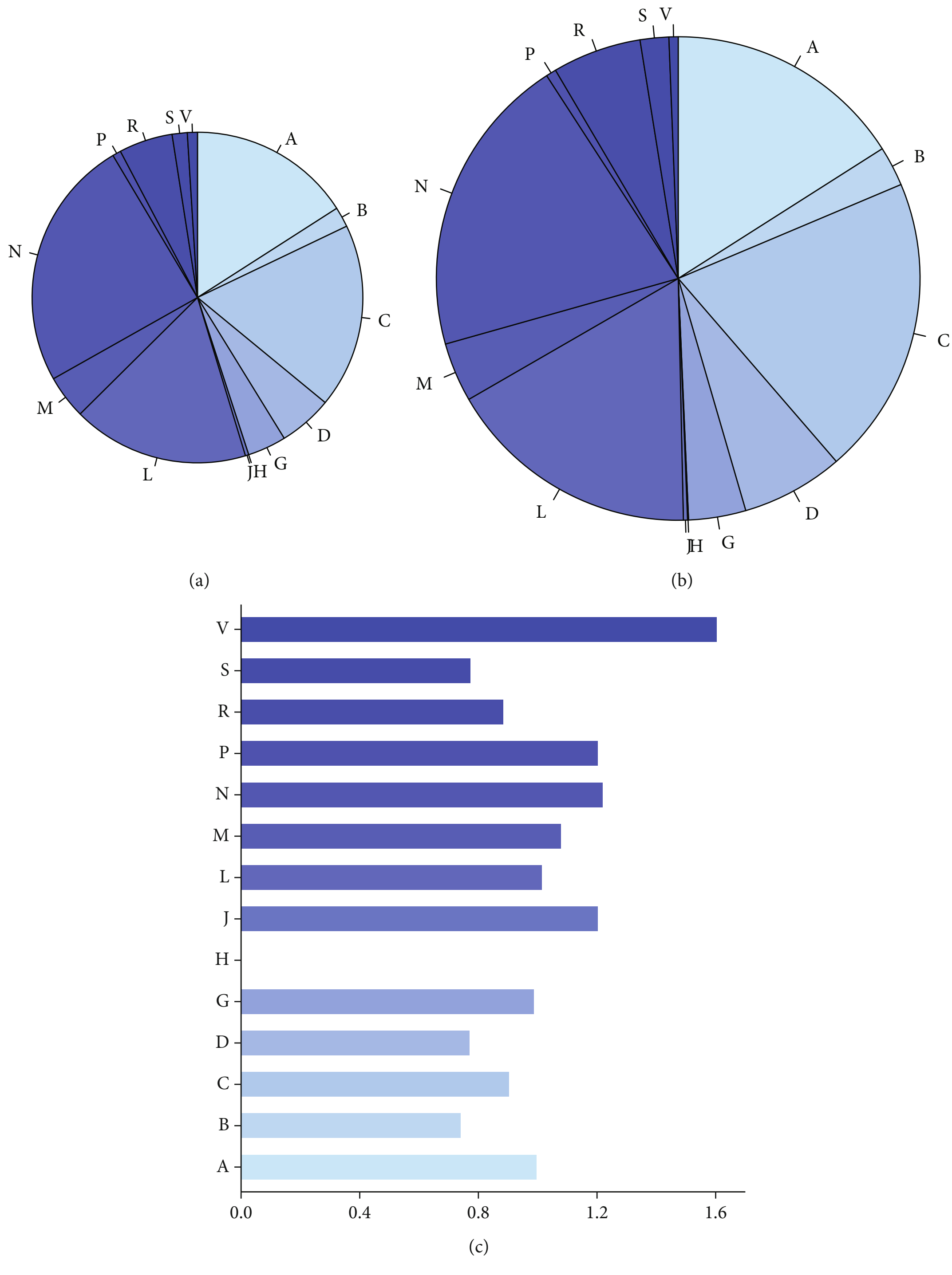


Figure 7: Distribution of ATC labels between DTIs in the known (a) and predicted (b) interactions. The relative ratio between known and predicted DTIs for each ATC label is shown in the right panel. ATC labels include the following: A—alimentary tract and metabolism; B—blood and blood-forming organs; C—cardiovascular system; D—dermatological; G—genitourinary system and sex hormones; H—systemic hormonal preparations, excluding sex hormones and insulins; J—anti-infectives for systemic use; L—antineoplastic and immunomodulating agents; M—musculoskeletal system; N—nervous system; P—antiparasitic products; R—respiratory system; S—sensory organs; and V—several others.

MAPK3 (Entrez ID: 5595) is a neighbor to the known target of Olomoucine (MAPK1, Entrez ID: 5594) in the protein-protein interaction (PPI) network. The PPI network, which contains 270,970 pairs of protein-protein interaction, is obtained from the BioGRID database [51]. By observing whether the edges (between two proteins) exist or not, we can judge whether drug known targets and predicted targets are neighbors in the PPI network. The closer two proteins are in the PPI network, the more likely they share the same functionality. Therefore, if the predicted targets are neighbors to the known targets of drugs, they might be targeted in the same way as known targets and the prediction results would be relatively reliable.

Indeed, recent research has shown that MAPK3 can be substantially inhibited by Olomoucine [52, 53]. This indicates that MAPK3 may be a novel target of Olomoucine.

Drug acetylsalicylic acid (commonly known or available as Aspirin, CID: 2244) is used for the treatment of pain and fever due to various causes. For acetylsalicylic acid, its predicted target is cyclin-dependent kinase-2 (CDK2) through AdvB-DTI.

CDK2 (Entrez ID: 1017) is a neighbor to two known targets of acetylsalicylic acid in the PPI network (Entrez IDs: 7157, 6256). Recent research has shown that CDK2 may be a novel target of acetylsalicylic acid [54]. This verifies our prediction.

CDK2 is a member of protein kinase family. It plays an important role in regulating various events of eukaryotic cell division cycle. Accumulated evidence indicated that overexpression of CDK2 should cause the abnormal regulation of cell-cycle, which would be directly associated with hyperproliferation in cancer cells [55]. Moreover, the examination of different kinds of human cancers, with defined molecular features, for their susceptibility to CDK2 inhibition has unveiled the scope in which CDK2 might represent a good therapeutic target [56–63].

Based on the above information, we speculate that acetylsalicylic acid, which is predicted to target CDK2, may have potential anticancer effects. Interestingly, the results of various studies have demonstrated that long-term use of acetylsalicylic acid may decrease the risk of various cancers, including colorectal, esophageal, breast, lung, prostate, liver, and skin cancer [64]. The predicted target CDK2 explains acetylsalicylic acid's anticancer effect to some extent.

Next example is the drug Panobinostat.

Panobinostat (CID: 6918837) is an oral deacetylase (DAC) inhibitor approved on February 23, 2015, by the FDA for the treatment of multiple myeloma. It acts as a nonselective histone deacetylase inhibitor (HDACi).

Histone deacetylase inhibitors (HDACis) are promising agents for cancer therapy. However, the mechanism(s) responsible for the efficacy of HDACi have not yet to be fully elucidated [65].

In this study, we predicted that Panobinostat's target is ATF3 through AdvB-DTI.

ATF3 (Entrez ID: 467) is a neighbor to six known targets of Panobinostat in the PPI network (Entrez IDs: 3065, 10013, 83933, 9759, 10014, 8841). As a proapoptotic factor, it plays a role in apoptosis and proliferation, two cellular processes critical for cancer progression [66–68]. And ATF3 has been postulated to be a tumor suppressor gene because it coordinates the expression of genes that may be linked to cancer [69].

Recent research has shown that ATF3 plays an important role in HDACi-induced apoptosis in multiple cell types [70]. HDACi can induce upregulation of ATF3 expression, thus eliciting the antitumor response [71].

Therefore, Panobinostat, as a HDACi, may treat myeloma by targeting ATF3.

Another interesting case is caffeine.

Caffeine (CID: 2519) is a widely consumed pharmacologically active product. It can be used for a variety of purposes, including the short-term treatment of apnea of prematurity in infants and pain relief and to avoid drowsiness [72].

For caffeine, its predicted targets include PTGS2 (Entrez ID: 5743) and PPARG (Entrez ID: 5468) through AdvB-DTI.

PTGS2 is one of two cyclooxygenases in humans. As a proinflammatory gene, it plays an important role in inflammation. Recent research has shown that caffeine treatment can reduce the expression of proinflammatory genes, including PTGS2 [73]. And caffeine can bind to PTGS2 acetaminophen complex with high energy, therefore modulating PTGS2 inhibition [74]. Furthermore, upregulation of PTGS2 is a critical oncogenic pathway in skin tumorigenesis. Han et al. verified that caffeine could block UVB-induced PTGS2 upregulation [75]. All these studies show that PTGS2 is a potential target for caffeine.

PPARG, another predicted target, is a ligand-activated transcription factor and important modulator for inflammation and lymphocyte homeostasis. There is also a study showing that PPARG were suppressed even with a low caffeine dose [76]. This suggests that PPARG is also a potential target for caffeine.

The above cases illustrate that our prediction results have a potential practical value and can provide clues to the analysis of the mechanism of action of certain drugs.

7. Conclusion

In this paper, we propose a DTI prediction framework named AdvB-DTI. Based on Bayesian Personalized Ranking, it uses the method of matrix factorization to predict DTIs. In order to solve the problem of existing DTI prediction methods based on matrix factorization, the proposed method combines the features of drugs and targets with the matrix factorization method. The advantage of this method over other similar methods is that BPR is combined with the perturbation factor and dual similarity regularization to make the model more robust and the training results more accurate. Experimental results verify that AdvB-DTI efficiently utilizes the similarity of drug-drug and target-target and the relationship of drugs and targets to train latent factors for drugs and targets to improve DTI prediction performance.

This study has the following positive impacts on the biomedical research.

Firstly, by integrating transcriptome data from drugs and genes, our model provides a practically useful and efficient tool for DTI prediction. The results of our study demonstrate that our method could discover reliable DTIs, thereby reducing the size of the search space for wet experiments and improving the drug discovery process.

Secondly, effective DTI prediction is achieved based on the transcriptome data. Our model used drug perturbation and gene knockout transcriptome data from the L1000 database of the LINCS project. Because the cost of experiments in LINCS project is relatively low, our prediction based on LINCS data not only ensures high accuracy but also has low cost.

Thirdly, our effective predictions verify that there is indeed a correlation between drug perturbation and the drug's target gene knockout at the transcriptional level. This correlation not only provides a basis for high-precision drug-target predictions but also provides a transcriptional perspective for the interpretation of drug mode of action. The correlation can also provide clues for future drug discovery.

Data Availability

Previously reported LINCS L1000 gene expression signature data were used to support this study and are available at DOI 10.1093/nar/gku476. This prior study (and dataset) is cited at relevant places within the text as a reference [31]. And previously reported DrugBank DTI data were used to support this study and are available at DOI 10.1093/nar/gkx1037. This prior study (and dataset) is cited at relevant places within the text as a reference [34].

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Authors' Contributions

Yihua Ye is responsible for the conceptualization, methodology, software, writing of the original draft, and investigation. Yuqi Wen is responsible for the resources, data curation, and writing of the original draft. Zhongnan Zhang did the formal analysis, writing of the review and editing, and supervision. Song He is involved in the investigation and writing of the review and editing. Xiaochen Bo is assigned in the validation and project administration. Yihua Ye and Yuqi Wen contributed equally to this work and should be considered as co-first authors.

References

[1] H. Zhou, M. Gao, and J. Skolnick, "Comprehensive prediction of drug-protein interactions and side effects for the human proteome," *Scientific Reports*, vol. 5, no. 1, p. 11090, 2015.

[2] Z. Yu, F. Huang, X. Zhao, W. Xiao, and W. Zhang, "Predicting drug–disease associations through layer attention graph convolutional network," *Briefings in Bioinformatics*, 2020.

[3] J. Vamathevan, D. Clark, P. Czodrowski et al., "Applications of machine learning in drug discovery and development," *Nature Reviews Drug Discovery*, vol. 18, no. 6, pp. 463–477, 2019.

[4] G. Schneider, "Automating drug discovery," *Nature Reviews Drug Discovery*, vol. 17, no. 2, pp. 97–113, 2018.

[5] R. Chen, X. Liu, S. Jin, J. Lin, and J. Liu, "Machine learning for drug-target interaction prediction," *Molecules*, vol. 23, no. 9, p. 2208, 2018.

[6] H. Öztürk, A. Özgür, and E. Ozkirimli, "DeepDTA: deep drug–target binding affinity prediction," *Bioinformatics*, vol. 34, no. 17, pp. i821–i829, 2018.

[7] W. Zhang, W. Lin, D. Zhang, S. Wang, J. Shi, and Y. Niu, "Recent advances in the machine learning-based drug-target interaction prediction," *Current Drug Metabolism*, vol. 20, no. 3, pp. 194–202, 2019.

[8] L. Xie, S. He, X. Song, X. Bo, and Z. Zhang, "Deep learning-based transcriptome data classification for drug-target interaction prediction," *BMC Genomics*, vol. 19, no. S7, p. 667, 2018.

[9] K. C. Chan and Z. H. You, "Large-scale prediction of drug-target interactions from deep representations," in *2016 International Joint Conference on Neural Networks (IJCNN)*, pp. 1236–1243, Vancouver, BC, Canada, July 2016.

[10] M. Hamanaka, K. Taneishi, H. Iwata et al., "CGBVS-DNN: prediction of compound-protein interactions based on deep learning," *Molecular Informatics*, vol. 36, 2016.

[11] K. Bleakley and Y. Yamanishi, "Supervised prediction of drug-target interactions using bipartite local models," *Bioinformatics*, vol. 25, no. 18, pp. 2397–2403, 2009.

[12] J. P. Mei, C. K. Kwoh, P. Yang, X. L. Li, and J. Zheng, "Drug-target interaction prediction by learning from local information and neighbors," *Bioinformatics*, vol. 29, no. 2, pp. 238–245, 2013.

[13] K. Buza and L. Peška, "Drug–target interaction prediction with bipartite local models and hubness-aware regression," *Neurocomputing*, vol. 260, pp. 284–293, 2017.

[14] T. van Laarhoven, S. B. Nabuurs, and E. Marchiori, "Gaussian interaction profile kernels for predicting drug-target interaction," *Bioinformatics*, vol. 27, no. 21, pp. 3036–3043, 2011.

[15] T. van Laarhoven and E. Marchiori, "Predicting drug-target interactions for new drug compounds using a weighted nearest neighbor profile," *PLoS One*, vol. 8, no. 6, article e66952, 2013.

[16] X. Chen, M.-X. Liu, and G.-Y. Yan, "Drug–target interaction prediction by random walk on the heterogeneous network," *Molecular BioSystems*, vol. 8, no. 7, pp. 1970–1978, 2012.

[17] Y. Luo, X. Zhao, J. Zhou et al., "A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information," *Nature Communications*, vol. 8, no. 1, p. 573, 2017.

[18] F. Wan, L. Hong, A. Xiao, T. Jiang, and J. Zeng, "NeoDTI: neural integration of neighbor information from a heterogeneous network for discovering new drug–target interactions," *Bioinformatics*, vol. 35, no. 1, pp. 104–111, 2019.

[19] M. A. Thafar, R. S. Olayan, H. Ashoor et al., "DTiGEMS+: drug–target interaction prediction using graph embedding, graph mining, and similarity-based techniques," *Journal of Cheminformatics*, vol. 12, no. 1, p. 44, 2020.

[20] T. Zhao, Y. Hu, L. R. Valsdottir, T. Zang, and J. Peng, "Identifying drug–target interactions based on graph convolutional network and deep neural network," *Briefings in Bioinformatics*, vol. bbaa044, 2020.

[21] S. Agarwal, D. Dugar, and S. Sengupta, "Ranking chemical structures for drug discovery: a new machine learning approach," *Journal of Chemical Information and Modeling*, vol. 50, no. 5, pp. 716–731, 2010.

[22] W. Zhang, L. Ji, Y. Chen et al., "When drug discovery meets web search: learning to rank for ligand-based virtual screening," *Journal of Cheminformatics*, vol. 7, no. 1, p. 5, 2015.

[23] Q. Yuan, J. Gao, D. Wu, S. Zhang, H. Mamitsuka, and S. Zhu, "DrugE-Rank: improving drug–target interaction prediction of new candidate drugs or targets by ensemble learning to rank," *Bioinformatics*, vol. 32, no. 12, pp. i18–i27, 2016.

[24] Z. Shi and J. Li, "Drug-target interaction prediction with weighted Bayesian ranking," in *ICBEB 2018: Proceedings of the 2nd International Conference on Biomedical Engineering*

and Bioinformatics, pp. 19–24, New York, NY, USA, September 2018.

[25] M. Bagherian, E. Sabeti, K. Wang, M. A. Sartor, Z. Nikolovska-Coleska, and K. Najarian, “Machine learning approaches and databases for prediction of drug–target interaction: a survey paper,” *Briefings in Bioinformatics*, vol. 22, no. 1, pp. 247–269, 2021.

[26] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, “BPR: Bayesian personalized ranking from implicit feedback,” in *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pp. 452–461, Arlington, Virginia, USA, June 2009.

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

[28] X. He, Z. He, X. Du, and T. S. Chua, “Adversarial personalized ranking for recommendation,” in *SIGIR '18: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 355–364, New York, NY, USA, June 2018.

[29] Y. Deng, X. Xu, Y. Qiu, J. Xia, W. Zhang, and S. Liu, “A multimodal deep learning framework for predicting drug–drug interaction events,” *Bioinformatics*, vol. 36, no. 15, pp. 4316–4322, 2020.

[30] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, “Neural collaborative filtering,” in *WWW '17: Proceedings of the 26th International Conference on World Wide Web*, pp. 173–182, Perth, WA, Australia, April, 2017.

[31] Q. Duan, C. Flynn, M. Niepel et al., “LINCS canvas browser: interactive web app to query, browse and interrogate LINCS L1000 gene expression signatures,” *Nucleic Acids Research*, vol. 42, no. W1, pp. W449–W460, 2014.

[32] N. A. Pabon, Y. Xia, S. K. Estabrooks et al., “Predicting protein targets for drug-like compounds using transcriptomics,” *PLoS Computational Biology*, vol. 14, no. 12, article e1006651, 2018.

[33] M. V. Chengalvala, V. M. Chennathukuzhi, D. S. Johnston, P. E. Stevis, and G. S. Kopf, “Gene expression profiling and its practice in drug development,” *Current Genomics*, vol. 8, no. 4, pp. 262–270, 2007.

[34] D. S. Wishart, Y. D. Feunang, A. C. Guo et al., “DrugBank 5.0: a major update to the DrugBank database for 2018,” *Nucleic Acids Research*, vol. 46, no. D1, pp. D1074–D1082, 2018.

[35] 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, no. 12, pp. 3399–3409, 2013.

[36] F. Huang, X. Yue, Z. Xiong, Z. Yu, S. Liu, and W. Zhang, “Tensor decomposition with relational constraints for predicting multiple types of microRNA-disease associations,” *Briefings in Bioinformatics*, vol. bbaa140, 2019.

[37] S. Marini, F. Vitali, S. Rampazzi, A. Demartini, and T. Akutsu, “Protease target prediction via matrix factorization,” *Bioinformatics*, vol. 35, no. 6, pp. 923–929, 2019.

[38] J. Zheng, J. Liu, C. Shi, F. Zhuang, J. Li, and B. Wu, “Dual similarity regularization for recommendation,” in *Advances in Knowledge Discovery and Data Mining. PAKDD 2016*, vol. 9652 of *Lecture Notes in Computer Science*, pp. 542–554, Auckland, New Zealand, April 2016.

[39] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612.

[40] R. Fagin, R. Kumar, and D. Sivakumar, “Comparing topklists,” *SIAM Journal on Discrete Mathematics*, vol. 17, no. 1, pp. 134–160, 2003.

[41] V. Gligorijević, N. Malod-Dognin, and N. Pržulj, “Fuse: multiple network alignment via data fusion,” *Bioinformatics*, vol. 32, no. 8, pp. 1195–1203, 2016.

[42] F. Vitali, L. D. Cohen, A. Demartini et al., “A network-based data integration approach to support drug repurposing and multi-target therapies in triple negative breast cancer,” *PLoS One*, vol. 11, no. 9, article e0162407, 2016.

[43] Y. Wang, L. Wang, Y. Li, D. He, W. Chen, and T. Y. Liu, “A theoretical analysis of NDCG ranking measures,” in *Proceedings of the 26th annual conference on learning theory (COLT 2013)*, vol. 8, p. 6, Princeton, NJ, USA, June 2013.

[44] Y. H. Li, C. Y. Yu, X. X. Li et al., “Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics,” *Nucleic Acids Research*, vol. 46, no. D1, pp. D1121–D1127, 2018.

[45] S. D. Harding, J. L. Sharman, E. Faccenda et al., “The IUPHAR/BPS guide to pharmacology in 2018: updates and expansion to encompass the new guide to immunopharmacology,” *Nucleic Acids Research*, vol. 46, no. D1, pp. D1091–D1106, 2018.

[46] S. Günther, M. Kuhn, M. Dunkel et al., “SuperTarget and Matador: resources for exploring drug-target relationships,” *Nucleic Acids Research*, vol. 36, no. Database issue, pp. 919–922, 2007.

[47] D. Szklarczyk, A. Santos, C. von Mering, L. J. Jensen, P. Bork, and M. Kuhn, “STITCH 5: augmenting protein–chemical interaction networks with tissue and affinity data,” *Nucleic Acids Research*, vol. 44, no. D1, pp. D380–D384, 2016.

[48] K. C. Cotto, A. H. Wagner, Y.-Y. Feng et al., “DGIdb 3.0: a redesign and expansion of the drug–gene interaction database,” *Nucleic Acids Research*, vol. 46, no. D1, pp. D1068–D1073, 2018.

[49] A. P. Davis, C. J. Grondin, R. J. Johnson et al., “The comparative toxicogenomics database: update 2019,” *Nucleic Acids Research*, vol. 47, no. D1, pp. D948–D954, 2019.

[50] Y. Benjamini and Y. Hochberg, “Controlling the false discovery rate: a practical and powerful approach to multiple testing,” *Journal of the Royal Statistical Society, Series B*, vol. 57, pp. 289–300, 1995.

[51] A. Chatr-Aryamontri, B. J. Breitkreutz, R. Oughtred et al., “The biogrid interaction database: 2015 update,” *Nucleic Acids Research*, vol. 43, no. Database issue, p. 470, 2017.

[52] G. Takan, O. K. Guldu, and E. I. Medine, “Radioiodination of cyclin dependent kinase inhibitor Olomoucine loaded Fe@ Au nanoparticle and evaluation of the therapeutic efficacy on cancerous cells,” *Radiochimica Acta*, vol. 105, no. 3, pp. 225–240, 2017.

[53] J. Vesely, L. Havlicek, M. Strnad et al., “Inhibition of cyclin-dependent kinases by purine analogues,” *European Journal of Biochemistry*, vol. 224, no. 2, pp. 771–786, 1994.

[54] R. Dachineni, G. Ai, D. R. Kumar, S. S. Sadhu, H. Tummala, and G. J. Bhat, “Cyclin A2 and CDK2 as novel targets of aspirin and salicylic acid: a potential role in cancer prevention,” *Molecular Cancer Research*, vol. 14, no. 3, pp. 241–252, 2016.

[55] T. Chohan, H. Qian, Y. Pan, and J. Z. Chen, “Cyclin-dependent kinase-2 as a target for cancer therapy: progress in the

development of CDK2 inhibitors as anti-cancer agents," *Current Medicinal Chemistry*, vol. 22, no. 2, pp. 237–263, 2014.

[56] S. Tadesse, A. T. Anshabo, N. Portman et al., "Targeting CDK2 in cancer: challenges and opportunities for therapy," *Drug Discovery Today*, vol. 25, no. 2, pp. 406–413, 2020.

[57] S. Tadesse, E. C. Caldon, W. Tilley, and S. Wang, "Cyclin-dependent kinase 2 inhibitors in cancer therapy: an update," *Journal of Medicinal Chemistry*, vol. 62, no. 9, pp. 4233–4251, 2019.

[58] L. Yang, D. Fang, H. Chen et al., "Cyclin-dependent kinase 2 is an ideal target for ovary tumors with elevated cyclin E1 expression," *Oncotarget*, vol. 6, no. 25, pp. 20801–20812, 2015.

[59] J. J. Molenaar, M. E. Ebus, D. Geerts et al., "Inactivation of CDK2 is synthetically lethal to MYCN over-expressing cancer cells," *Proceedings of the National Academy of Sciences*, vol. 106, no. 31, pp. 12968–12973, 2009.

[60] S. Hu, Y. Lu, B. Orr et al., "Specific CP110 phosphorylation sites mediate anaphase catastrophe after CDK2 inhibition: evidence for cooperation with USP33 knockdown," *Molecular Cancer Therapeutics*, vol. 14, no. 11, pp. 2576–2585, 2015.

[61] M. Takada, W. Zhang, A. Suzuki et al., "FBW7 loss promotes chromosomal instability and tumorigenesis via cyclin E1/CDK2–mediated phosphorylation of CENP-A," *Cancer Research*, vol. 77, no. 18, pp. 4881–4893, 2017.

[62] J. Wang, T. Yang, G. Xu et al., "Cyclin-dependent kinase 2 promotes tumor proliferation and induces radio resistance in glioblastoma," *Translational Oncology*, vol. 9, no. 6, pp. 548–556, 2016.

[63] A. Faber and T. C. Chiles, "Inhibition of cyclin-dependent kinase-2 induces apoptosis in human diffuse large B-cell lymphomas," *Cell Cycle*, vol. 6, no. 23, pp. 2982–2989, 2014.

[64] L. Alfonso, G. Ai, R. C. Spitale, and G. J. Bhat, "Molecular targets of aspirin and cancer prevention," *British Journal of Cancer*, vol. 111, no. 1, pp. 61–67, 2014.

[65] J. Liu, M. Edagawa, H. Goshima et al., "Role of ATF3 in synergistic cancer cell killing by a combination of HDAC inhibitors and agonistic anti-DR5 antibody through ER stress in human colon cancer cells," *Biochemical and Biophysical Research Communications*, vol. 445, no. 2, pp. 320–326, 2014.

[66] X. Yin, J. W. DeWille, and T. Hai, "A potential dichotomous role of ATF3, an adaptive-response gene, in cancer development," *Oncogene*, vol. 27, no. 15, pp. 2118–2127, 2008.

[67] F. G. Bottone, Y. Moon, J. S. Kim, B. Alston-Mills, M. Ishibashi, and T. E. Eling, "The anti-invasive activity of cyclooxygenase inhibitors is regulated by the transcription factor ATF3 (activating transcription factor 3)," *Molecular Cancer Therapeutics*, vol. 4, no. 5, pp. 693–703, 2005.

[68] D. Lu, C. D. Wolfgang, and T. Hai, "Activating Transcription Factor 3, a Stress-inducible Gene, Suppresses Ras- stimulated Tumorigenesis∗," *Journal of Biological Chemistry*, vol. 281, no. 15, pp. 10473–10481, 2006.

[69] T. W. Fawcett, J. L. Martindale, K. Z. Guyton, T. Hai, and N. J. Holbrook, "Complexes containing activating transcription factor (ATF)/cAMP-responsive-element-binding protein (CREB) interact with the CCAAT/enhancer-binding protein (C/EBP)–ATF composite site to regulate Gadd153 expression during the stress response," *Biochemical Journal*, vol. 339, no. 1, pp. 135–141, 1999.

[70] A. C. Chüeh, J. Tse, M. Dickinson et al., "ATF3 repression of BCL-XLDetermines apoptotic sensitivity to HDAC inhibitors across tumor types," *Clinical Cancer Research*, vol. 23, no. 18, pp. 5573–5584, 2017.

[71] C. St Germain, A. O'Brien, and J. Dimitroulakos, "Activating transcription factor 3 regulates in part the enhanced tumour cell cytotoxicity of the histone deacetylase inhibitor M344 and cisplatin in combination," *Cancer Cell International*, vol. 10, no. 1, p. 32, 2010.

[72] J. Evans, J. R. Richards, and A. S. Battisti, "Caffeine," in *In StatPearls*, StatPearls Publishing, 2020.

[73] J. H. Hwang, K. J. Kim, S. J. Ryu, and B. Y. Lee, "Caffeine prevents LPS-induced inflammatory responses in RAW264. 7 cells and zebrafish," *Chemico-Biological Interactions*, vol. 248, pp. 1–7, 2016.

[74] G. C. Krisnamurti and F. Fatchiyah, "Interaction of acetaminophen and caffeine towards cyclooxygenase-2 (COX-2) in inhibition of prostaglandin (PGH 2) synthesis," *JPhCS*, vol. 1146, no. 1, article 012004, 2019.

[75] W. Han, M. Ming, and Y. Y. He, "Caffeine promotes ultraviolet B-induced apoptosis in human keratinocytes without complete DNA repair," *Journal of Biological Chemistry*, vol. 286, no. 26, pp. 22825–22832, 2011.

[76] M. Iris, P. S. Tsou, and A. H. Sawalha, "Caffeine inhibits STAT1 signaling and downregulates inflammatory pathways involved in autoimmunity," *Clinical Immunology*, vol. 192, pp. 68–77, 2018.

Hindawi
BioMed Research International
Volume 2021, Article ID 6627650, 10 pages
https://doi.org/10.1155/2021/6627650

Research Article

Assessing Dry Weight of Hemodialysis Patients via Sparse Laplacian Regularized RVFL Neural Network with $L_{2,1}$-Norm

Xiaoyi Guo,[1] Wei Zhou,[1] Qun Lu,[2] Aiyan Du,[1] Yinghua Cai,[2] and Yijie Ding[3]

[1] *Hemodialysis Center, The Affiliated Wuxi People's Hospital of Nanjing Medical University, 214000 Wuxi, China*
[2] *Nursing Department, The Affiliated Wuxi People's Hospital of Nanjing Medical University, 214000 Wuxi, China*
[3] *School of Electronic and Information Engineering, Suzhou University of Science and Technology, 215009 Suzhou, China*

Correspondence should be addressed to Wei Zhou; 285403434@qq.com, Yinghua Cai; 179098331@qq.com, and Yijie Ding; wuxi_dyj@163.com

Received 1 January 2021; Revised 21 January 2021; Accepted 25 January 2021; Published 5 February 2021

Academic Editor: Quan Zou

Dry weight is the normal weight of hemodialysis patients after hemodialysis. If the amount of water in diabetes is too much (during hemodialysis), the patient will experience hypotension and shock symptoms. Therefore, the correct assessment of the patient's dry weight is clinically important. These methods all rely on professional instruments and technicians, which are time-consuming and labor-intensive. To avoid this limitation, we hope to use machine learning methods on patients. This study collected demographic and anthropometric data of 476 hemodialysis patients, including age, gender, blood pressure (BP), body mass index (BMI), years of dialysis (YD), and heart rate (HR). We propose a Sparse Laplacian regularized Random Vector Functional Link (SLapRVFL) neural network model on the basis of predecessors. When we evaluate the prediction performance of the model, we fully compare SLapRVFL with the Body Composition Monitor (BCM) instrument and other models. The Root Mean Square Error (RMSE) of SLapRVFL is 1.3136, which is better than other methods. The SLapRVFL neural network model could be a viable alternative of dry weight assessment.

1. Introduction

Fluid overload in patients with chronic renal failure is closely related to poor cardiovascular outcomes [1, 2]. Maintenance of hemodialysis (HD) is the main method for patients with renal failure [3]. However, the accurate assessment of body water volume is still a concern [4]. At present, dry weight has been used as an important indicator to assess the homeostasis of fluids in hemodialysis patients. Medical staff can use the patient's dry weight to estimate the amount of water needed for dialysis during hemodialysis. The conventional clinical-based dry weight assessment method is time-consuming and labor-intensive [1]. There are already some methods based on bioelectrical impedance analysis (BIA) [5] to determine dry weight, including body composition monitor (BCM) [6] and lung ultrasound (LUS). However, all the above methods require special instruments and professional technicians to complete. Medical staff can use some clinical data to build predictive models [7] to accurately assess dry weight. Currently, machine learning (ML) or deep learning has solved many common clinical problems in medicine, such as brain diseases [8–10], cancer analysis, and diabetes.

Some scholars have used artificial neural networks (ANN) to predict the total water volume of hemodialysis patients and have obtained better results than conventional clinical calculation equations [11]. In addition, deep learning methods are also emerging in clinical diagnosis, including pixel-based convolutional neural networks to diagnose skin cancer [12]. In the biological field, microbiology analysis [13], CircRNAs [14], microRNAs, and cancer association prediction [15–17], lncRNA-miRNA association prediction, O-GlcNAcylation site prediction [18], DNA methylation site [19–21], protein remote homology [22], function prediction of proteins [23–29], electron transport proteins [30], breast cancer [31], cell-specific replication [32], osteoporosis diagnoses [33], and drug complex network analysis [34–38].

In our previous research, a Multiple Kernel Support Vector Regression (MKSVR) [39] predictor was proposed to assess the dry weight and obtain good predictive performance. Inspired by the previous work and baseline Random Vector Functional Link (RVFL) network [40], we propose a new dry weight assessment model, called Sparse Laplacian regularized RVFL neural network with $L_{2,1}$-norm (SLapRVFL), which considers the topological relationship between samples and more sparse connections between the input layer and the hidden layer.

2. Materials and Methods

2.1. Materials. This work collects demographic and anthropometric data and bioimpedance spectroscopy (BIS) from historical data (2018-9 to 2019-9) from Wuxi people's hospital and the northern Jiangsu people's hospital. This study has been approved by the ethics committees of the hospitals (Nos. KYLLKS201813 and 2018KY-001). The collected patient data meet the following requirements: age greater than 18 years; ESRD for more than three months and maintenance hemodialysis [41]; no heart failure, no metal implants, no pregnancy, no disability, no infection, and no edema and other diseases; and hemodialysis treatment 3 times a week, 4 hours each time. Finally, we obtain a data set of 476 hemodialysis patients. DW is the normal body weight after clinical diabetes. DW is obtained by a clinician under strict clinical supervision using a clinical scoring system (using trial and error method) [42, 43].

We choose 7 features, including age, gender (binary feature), systolic blood pressure (SBP), diastolic blood pressure (DBP), body mass index (BMI), heart rate (HR), and years of dialysis (YD) to build our predictive model. Table 1 shows the information of the data set. BMI is measured before hemodialysis treatment.

Table 1: The information of data set.

Feature	Value	r^*
Age (years)	54.17 ± 14.22	-0.2341
Gender (males/females)	312/164	-0.4489
BMI	22.96 ± 2.95	0.9558
Systolic blood pressure (mmHg)	150.64 ± 29.36	-0.1739
Diastolic blood pressure (mmHg)	88.32 ± 19.56	-0.1249
Heart rate (times/min)	73.41 ± 8.92	0.1862
Years of dialysis (years)	5.97 ± 3.22	-0.1069

*Denotes that each feature correlated with dry weight using Pearson correlation coefficient (r).

2.2. Methods. The baseline RVFL was proposed for regression or classification. The schematic diagram of RVFL is shown in Figure 1. The basic information of the patient is put into the RVFL neural network model for processing, and the predicted dry weight is the output.

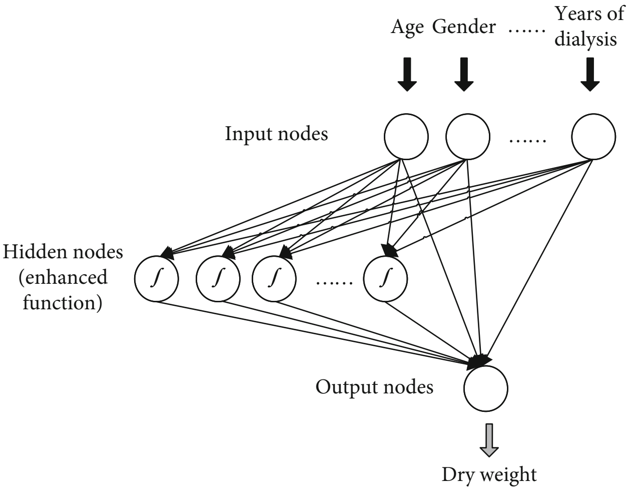


Figure 1: Schematic of our proposed method.

Suppose, there are N training samples with $\{x_i, y_i\}$, $i = 1, 2, \cdots, N$. The output value is $y_i \in R^{1\times c}$ and the input data is $x_i \in R^{1\times d}$. d denotes the dimension of x_i. As per Figure 1, RVFL randomly initializes all weights and deviations between the hidden layer and the input layer. These parameters are fixed during the training process and do not need to be tuned. There are connections between the output layer, input layer, and hidden layer. This part of the weight needs to be obtained by training RVFL. The output layer of RVFL is connected to both the input layer and the hidden layer, so as to ensure the nonlinear and linear relationships between the input and the output. The RVFL network with P hidden nodes are formulated as

$$H\beta = Y, \tag{1}$$

where β denotes the output weight matrix; H is the concatenated matrix, which combines the output of the hidden layer and the input layer; and Y denotes the label matrix. H and β can be represented as

$$H = [H_1\, H_2], \tag{2}$$

$$H_1 = \begin{bmatrix} x_{11} & \cdots & x_{1d} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{Nd} \end{bmatrix}_{N\times d}, \tag{3}$$

$$H_2 = \begin{bmatrix} G(a_1 x_1 + b_1) & \cdots & G(a_P x_1 + b_P) \\ \vdots & \ddots & \vdots \\ G(a_1 x_N + b_1) & \cdots & G(a_P x_N + b_P) \end{bmatrix}_{N\times P}, \tag{4}$$

$$\beta = \begin{bmatrix} \beta_1^T \\ \beta_2^T \\ \vdots \\ \beta_{d+P}^T \end{bmatrix}_{(d+P)\times C}. \tag{5}$$

```
Require: Training set $\{x_i, y_i\}, i = 1, 2, \cdots, N$, test set $\{x_j^{te}\}, j = 1, 2, \cdots, M$, the numbers of hidden layer nodes ($P$), the maximum number of iterations tmax, coefficients of $\lambda_1$ and $\lambda_2$;
Ensure: The predictive values of $\{y_j^{te}\}, j = 1, 2, \cdots, M$
(1) Randomly initializing all weights and deviations between the hidden layer and the input layer. Calculating the hidden layer output matrix $H$ (training set)and Laplacian matrix $L$ by Equations (2), (12), and (13);
(2) Set $t = 0$, estimate the initial $\beta^0$ using Equation (7);
Repeat
    (3) Update the diagonal matrix $G$ with
        $G^{t+1} = \begin{bmatrix} 1/2\|\beta_1^t\|_2 & & \\ & \ddots & \\ & & 1/2\|\beta_{d+P}^t\|_2 \end{bmatrix}_{(d+P)\times(d+P)}$,
    (4) Update $\beta$ via Equation (11d);
Untilt > tmax;
(5) Calculate the hidden layer output matrix $H^{te}$ (test set);
(6) Estimate $\{y_j^{te}\}, j = 1, 2, \cdots, M$ by $Y^{te} = H^{te}\beta$.
```

Algorithm 1. Algorithm of SLapRVFL

In Equation (4), a_j and b_j are the weights and bias of the hidden and input layers. C and P are numbers of output and hidden layer nodes. In general, the activation function is a Gaussian function: $g(x) = e^{-x^2}$. The activation function has a nonlinear approximation effect. To consider the potential linear relationship between the input data and the output value, RVFL adds a direct connection weight between the input layer and the output layer. Therefore, RVFL is a model that contains both linear and nonlinear approximations to improve prediction performance. For optimal β, the RVFL can be formulated as a regularized least-squares:

$$\beta^* = \arg\min \frac{1}{2}\|H\beta - Y\|_2^2 + \frac{\lambda}{2}\|\beta\|_2^2, \tag{6}$$

where λ is the parameter of regularization term. The solution of Equation (6) can be found by setting its gradient to 0:

$$\beta^* = \left(H^T H + \lambda I\right)^{-1} H^T Y, \tag{7}$$

where I denotes the identity matrix. However, the RVFL network did not consider the topological relationship between samples. For the output node, it must be connected to both the input and the hidden layer.

In order to further improve the robustness of RVFL, we propose Sparse Laplacian regularized RVFL neural network with $L_{2,1}$-norm (SLapRVFL). The objective function is

$$\beta^* = \arg\min \frac{1}{2}\|H\beta - Y\|_2^2 + \frac{\lambda_1}{2} Tr\left((H\beta)^T LH\beta\right) + \frac{\lambda_2}{2}\|\beta\|_{2,1}^2, \tag{8}$$

where $L \in R^{N\times N}$ denotes the Laplacian matrix. λ_1 and λ_2 are the coefficients of Laplacian regularization the and L_{21}-norm term, respectively. Laplacian regularization is used to indicate the potential manifold between samples. It can better describe the topological association between samples to improve the generalization ability of the model. Since the third term of $\|\beta\|_{2,1}^2$ is not diversified, we convert Equation (8) to

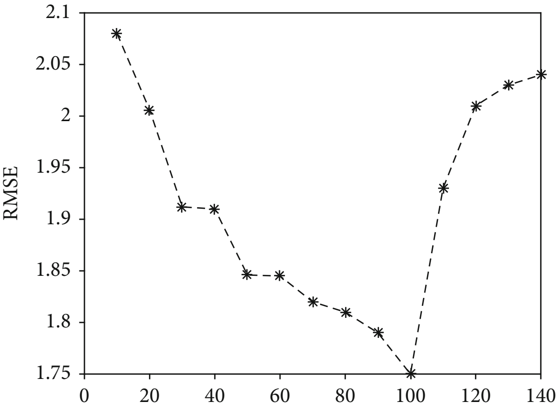


Figure 2: The RMSE under different numbers of hidden layer nodes (SLapRVFL network).

$$\beta^* = \arg\min \frac{1}{2}\|H\beta - Y\|_2^2 + \frac{\lambda_1}{2} Tr\left((H\beta)^T LH\beta\right) + \frac{\lambda_2}{2} Tr\left(\beta^T G\beta\right), \tag{9}$$

where $G \in R^{(d+P)\times(d+P)}$ denotes a diagonal matrix whose ith-diagonal element

$$G_{ii} = \frac{1}{2\|\beta_i\|_2}, \quad i = 1, 2, \cdots, (d + P). \tag{10}$$

We take the derivative of the formula Equation (10) as

$$H^T(H\beta - Y) + \lambda_1 H^T LH\beta + \lambda_2 G\beta = 0, \tag{11a}$$

$$H^T H\beta + \lambda_1 H^T LH\beta + \lambda_2 G\beta = H^T Y, \tag{11b}$$

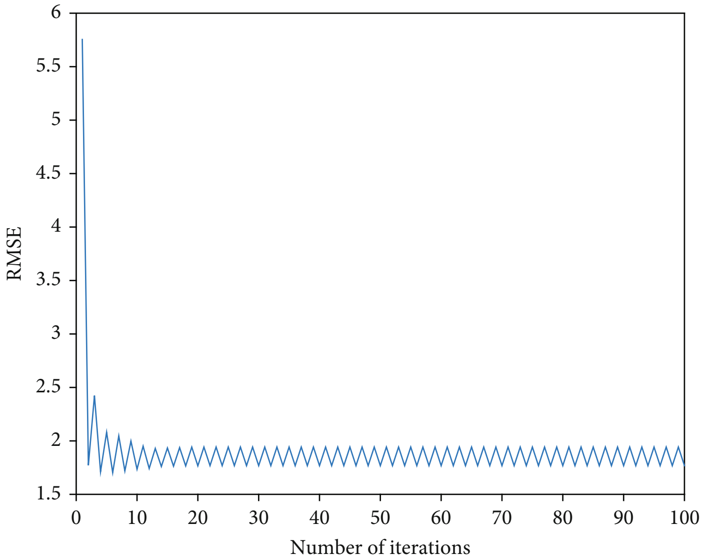


Figure 3: The RMSE of iterations on the training set.

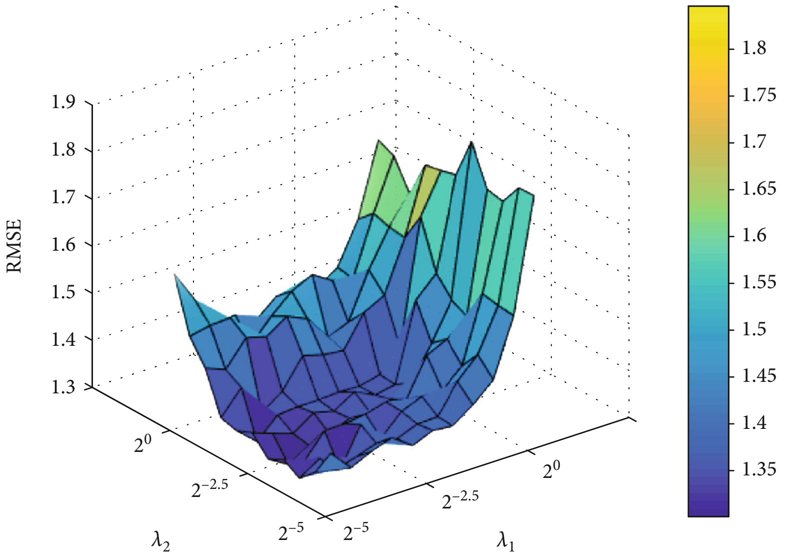


Figure 4: The RMSE under different λ_1 and λ_2.

$$\left(H^T H + \lambda_1 H^T L H + \lambda_2 G\right)\beta = H^T Y, \tag{11c}$$

$$\beta = \left(H^T H + \lambda_1 H^T L H + \lambda_2 G\right)^{-1} H^T Y. \tag{11d}$$

We use the baseline RVFL solution with Equation (7) as the initial β^0. In addition, the Laplacian matrix can be calculate as

$$L = D^{-1/2} \Delta D^{-1/2}, \tag{12a}$$

$$\Delta = D - S, \tag{12b}$$

where D is diagonal matrix, $D_{ii} = \sum_{j=1}^{N} S_{ij}$. Similarity matrix S is built by Radial Basis Function (RBF):

$$S_{ij} = \exp\left(-\gamma \left\| x_i - x_j \right\|^2\right). \tag{13}$$

The process of SLapRVFL is list in Algorithm 1.

3. Results

We test our model on the benchmark data set and obtain the optimal parameters of the predictor through cross-validation. The SLapRVFL network is compared to other machine learning-based models. In addition, the body composition monitor (BCM) device (Fresenius Medical Care, Baden Humboldt, Germany) is also compared with the SLapRVFL network.

3.1. Evaluation Measurements. The 10-fold cross-validation (10-CV) is employed to evaluate the robustness of methods. Root Mean Square Error (RMSE), R square, correlation coefficient (R), Bland–Altman analysis, and Empirical Cumulative Distribution Plot (ECDP) [44] are all used in our study. To evaluate the agreement of two different methods, the Bland–Altman analysis usually can obtain whether the two

Table 2: Comparison on existing methods via 10-fold cross-validation.

Method	R	R squared	RMSE	Empirical cumulative distribution plot		
				Highest value	Lowest value	Median value
BCM*	0.9473	0.9137	1.9694	3.2235	-6.2776	-0.9863
LR*	0.9403	0.9308	1.4335	4.2524	-4.4014	0.1418
ANN (BP)*	0.9398	0.9295	1.4794	7.3661	-4.7447	0.1324
MKRR*	0.9399	0.9289	1.5015	4.9227	-4.2604	0.1104
MKSVR*	0.9412	0.9321	1.3817	4.3962	-4.1273	0.0082
RVFL	0.9389	0.9300	1.3828	6.7004	-4.3557	0.0704
SLapRVFL (our method)	0.9632	0.9501	1.3136	3.1940	-3.5066	0.1014

*The results are from previous work on MKSVR [39].

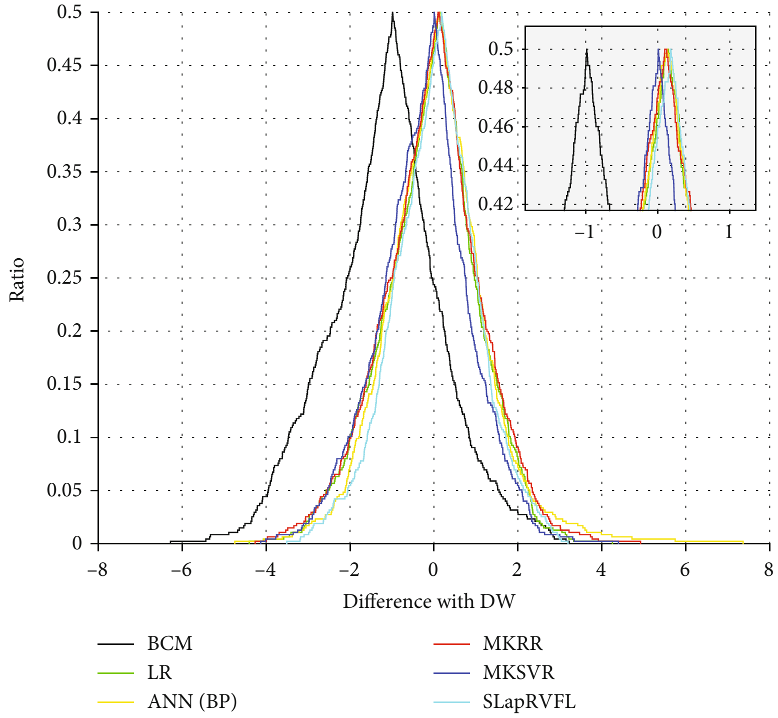
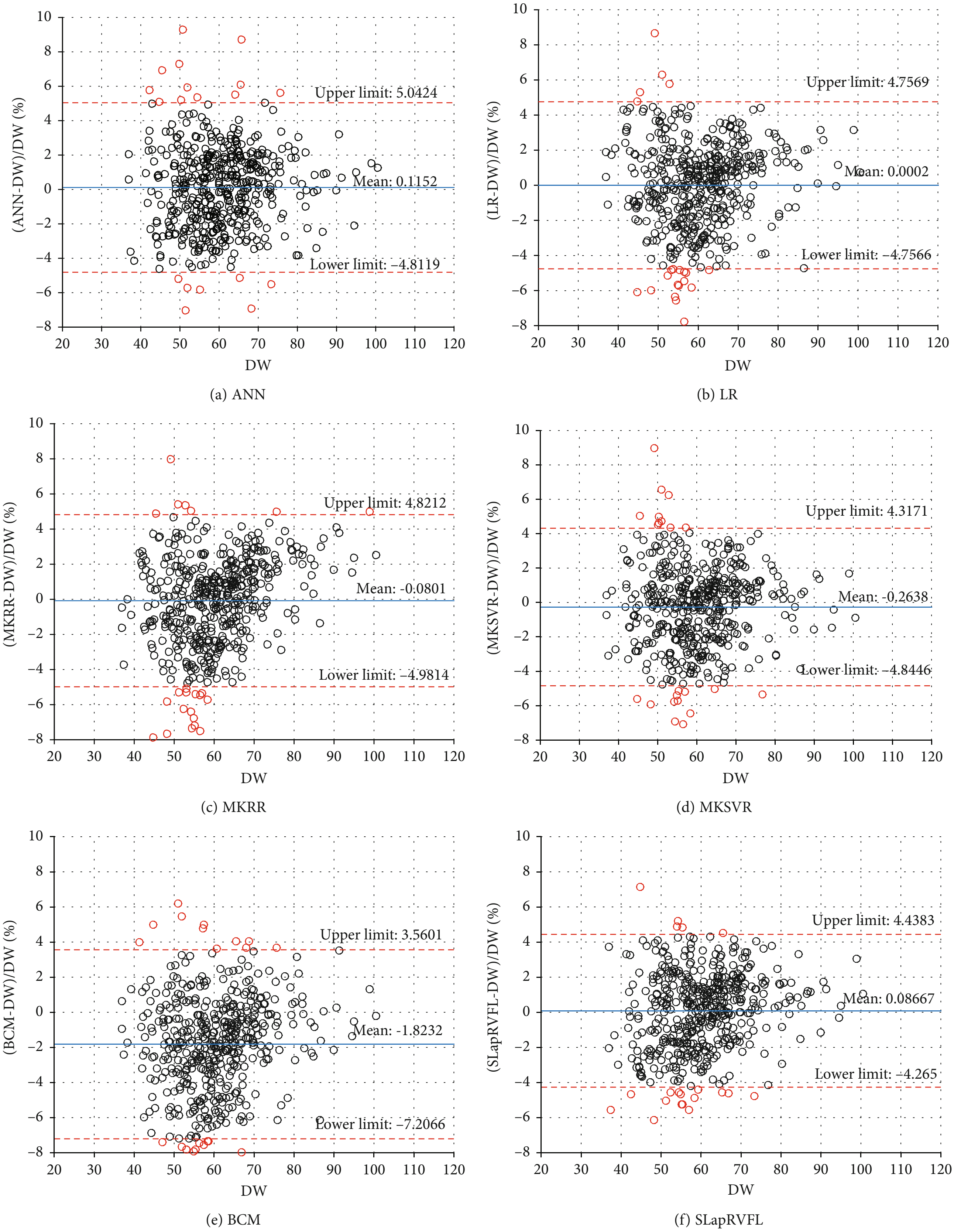


Figure 5: Folded empirical cumulative distribution plot between different methods.

Table 3: Bland–Altman plot analysis for different models.

Model	Differences with DW (%)			Limits of agreement (%)		
	Mean	SD	95% confidence interval	Lower limit	Upper limit	Number (ratio) of outside agreement interval
BCM*	-1.8232	2.7466	-2.0706 to -1.5759	-7.2066	3.5601	30/476 (6.30%)
LR*	0.0002	2.4269	-0.2184 to 0.2187	-4.7566	4.7569	21/476 (4.41%)
ANN (BP)*	0.1152	2.5139	-0.1112 to 0.3416	-4.8119	5.0424	22/476 (4.62%)
MKRR*	-0.0801	2.5007	-0.3053 to 0.1451	-4.9814	4.8212	23/476 (4.83%)
MKSVR*	-0.2638	2.3372	-0.4743 to -0.05329	-4.8446	4.3171	22/476 (4.62%)
SLapRVFL (our method)	0.0867	2.2202	-0.1133 to 0.2866	-4.2650	4.4383	20/476 (4.20%)

*The results are from previous work on MKSVR [39].



(a) ANN

(b) LR

(c) MKRR

(d) MKSVR

(e) BCM

(f) SLapRVFL

Figure 6: Bland–Altman plot analysis.

methods can be substituted for each other (equivalence). Evaluating the agreement of the two methods can answer the question, “Can these two methods replace each other?”

3.2. Selection of Optimal Parameters. To get the optimal parameters of the predictive method, we obtain them through a grid search method. The parameters that need to be determined include the numbers of hidden layer nodes P , maximum iterations, and coefficients of λ_1 and λ_2. For the numbers of hidden layer nodes P, we fix the iterations, λ_1 and λ_2. Setting the maximum number as 50, $\lambda_1 = 1$ and $\lambda_2 = 1$. The value of P is from 10 to 140 with step of 10. The results are shown in Figure 2. From 10 to 100, the more neurons in the hidden layer, the lower the RMSE. Since then, RMSE has gradually increased. So, we get the lower RMSE under $P = 100$.

Next, $P = 100$, $\lambda_1 = 1$, and $\lambda_2 = 1$. We gradually increase the number of iterations from 1 to 100 (shown in Figure 3). After the number of iterations reaches 10, the RMSE value drops to a minimum and slightly oscillates within a certain value. In our study, maximum number of iterations is 10.

Then, we use the better number of hidden layer nodes and iterations to search for the best λ_1 and λ_2. The search range of parameters is from 2^{-5} to 2^0 (with step of $2^{0.5}$). Figure 4 shows the results of different parameters. When λ_1 and λ_2 are 2^{-3} and $2^{-2.5}$, RMSE is the lowest.

3.3. Comparison to Other Predictive Models and BCM. To evaluate our model, SLapRVFL is compared with our previous work of Multiple Kernel Support Vector Regression (MKSVR) [39], Multikernel Ridge Regression (MKRR), Linear Regression (LR), Artificial Neural Network based on Back Propagation algorithm (ANN with BP), and BCM measuring instrument. Clinical dry weight is our reference standard (also the regression target value of the prediction model). The comparisons are listed in Table 2, which shows that SLapRVFL achieves best performance of RMSE (1.3136). Although the ECDP median value (peak) of MKSVR (0.0082) is more close to zero, Figure 5 shows that SLapRVFL has the least bias and much less tails than MKSVR (smaller width). The RMSE of BCM is 1.9694, which is larger than SLapRVFL.

3.4. Bland–Altman Analysis. Bland–Altman plot is a useful tool to evaluate the agreement between predictive methods and clinical DW. In Table 3 and Figure 6, SLapRVFL, MKSVR, LR, ANN (BP), MKRR, and BCM are analyzed via Bland-Altman difference plot. SLapRVFL achieves the smallest range of 95% confidence interval (-0.1133 to 0.2866) and standard deviation (2.2202). In addition, the number (ratio) of outside agreement interval for predictive models is all less than 24 (5%) predictive samples. These results of models are clinically acceptable. SLapRVFL achieves least number (20) of the outside agreement interval in Table 3. As shown in Figure 6, two red horizontal dotted lines (upper and lower) denote the upper and lower limits of the 95% agreement limit, respectively. The middle blue solid line is the average value of the difference (between measurement methods and clinical DW). While one measurement method and clinical method can be considered as a better agreement, they can be substituted for each other (equivalence). If 95% of the points of the data set are in the agreement range, the measurement method (predictive model) is clinically acceptable. The results of the evaluation show that SLapRVFL can help clinicians assess DW with low cost.

4. Discussion

Due to the limitations of clinical and BCM measurement (more time and cost), this study uses a machine learning method to assess the dry weight of hemodialysis patients. Based on the basic RVFL, we propose a sparse Laplace regularized RVFL network (SLapRVFL) model. SLapRVFL is compared not only with other machine learning methods (such as LR, MKRR, ANN with BP, and MKSVR) but also with BCM equipment (commonly used in hospitals). The RMSE and Bland–Altman analysis of the model are better than the BCM instrument. It is proven that the predictive model driven by data can provide reference for clinical dry weight assessment.

BCM requires the patient’s information on weight (before hemodialysis) and height. It is a portable, inexpensive, and noninvasive technology that has been used to measure DW [45, 46]. For the Bland–Altman analysis, SLapRVFL achieves the least number (20) of outside agreement interval. However, BCM has 30/476 (6.30%) points (ratio) of the outside agreement interval. Obviously, our method has better agreement with the clinical method.

5. Conclusions

To further improve the robustness of RVFL, we introduce sparse Laplacian regular term with $L_{2,1}$-norm. In the training process, the graph topology information and the sparse weight matrix (output) are employed to improve the robustness of the RVFL. In fact, our work provides a new idea for assessing patients’ dry weight. Not only that, in the fields of biology [47–57], pharmacy [58], and medicine [12, 59, 60], machine learning methods have helped solve many analysis tasks. In future research, we will consider collecting more samples, introducing more patient personal information, and building a predictor based on a deep learning model to more accurately assess the dry weight of hemodialysis patients.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Ethical Approval

This study had been approved by the ethics committee of the hospital (ethical approval Nos. KYLLKS201813 and 2018KY-001). The experimental protocol was established, according to the ethical guidelines of the Helsinki Declaration, and was approved by the Human Ethics Committee (Wuxi

People's Hospital Ethics Committee and Northern Jiangsu People's Hospital Ethics Committee).

Consent

Written informed consent for publication was obtained from all participants.

Conflicts of Interest

The authors declare that they have no conflict of interest.

Authors' Contributions

Xiaoyi Guo and Wei Zhou are joint first authors.

Acknowledgments

The authors give their thanks to the Hemodialysis Center of Wuxi People's Hospital and Northern Jiangsu People's Hospital for collecting the data in this study. This work is supported by a grant from the National Natural Science Foundation of China (NSFC 61902271, 61772362, and 61972280) and the Natural Science Research of Jiangsu Higher Education Institutions of China (19KJB520014).

References

[1] A. Grassmann, I. Uhlenbusch-Körwer, E. Bonnie-Schorn, and J. Vienken, "Composition and management of hemodialysis fluids," *Good Dialysis Practice*, vol. 2, pp. 13–25, 2000.

[2] P. Wabel, P. Chamney, U. Moissl, and T. Jirka, "Importance of whole-body bioimpedance spectroscopy for the management of fluid balance," *Blood Purification*, vol. 27, no. 1, pp. 75–80, 2009.

[3] G. Alexiadis, S. Panagoutsos, S. Roumeliotis et al., "Comparison of multiple fluid status assessment methods in patients on chronic hemodialysis," *International Urology and Nephrology*, vol. 49, no. 3, pp. 1–8, 2016.

[4] Y. Ohashi, K. Sakai, H. Hase, and N. Joki, "Dry weight targeting: the art and science of conventional hemodialysis," *Seminars in Dialysis*, vol. 31, 2018.

[5] H. Asmat, R. Iqbal, F. Sharif, A. Mahmood, A. Abbas, and W. Kashif, "Validation of bioelectrical impedance analysis for assessing dry weight of dialysis patients in Pakistan," *Saudi Journal of Kidney Diseases & Transplantation*, vol. 28, no. 2, p. 285, 2017.

[6] C. Jiang, S. Patel, A. Moses, M. V. DeVita, and M. F. Michelis, "Use of lung ultrasonography to determine the accuracy of clinically estimated dry weight in chronic hemodialysis patients," *International Urology and Nephrology*, vol. 49, no. 12, pp. 2223–2230, 2017.

[7] P. Susantitaphong, S. Laowaloet, K. Tiranathanagul et al., "Reliability of blood pressure parameters for dry weight estimation in hemodialysis patients," *Therapeutic Apheresis and Dialysis*, vol. 17, no. 1, pp. 9–15, 2013.

[8] G. Liu, Y. Hu, Z. Han, S. Jin, and Q. Jiang, "Genetic variant rs17185536 regulates SIM1 gene expression in human brain hypothalamus," *Proceedings of the National Academy of Sciences*, vol. 116, no. 9, pp. 3347-3348, 2019.

[9] G. Liu, S. Jin, Y. Hu, and Q. Jiang, "Disease status affects the association between rs4813620 and the expression of Alzheimer's disease susceptibility gene TRIB3," *Proceedings of the National Academy of Sciences*, vol. 115, no. 45, pp. E10519–E10520, 2018.

[10] X. A. Bi, Y. Liu, Y. Xie, X. Hu, and Q. Jiang, "Morbigenous brain region and gene detection with a genetically evolved random neural network cluster approach in late mild cognitive impairment," *Bioinformatics*, vol. 36, no. 8, pp. 2561–2568, 2020.

[11] J. S. Chiu, C. F. Chong, Y. F. Lin, C. C. Wu, Y. F. Wang, and Y. C. Li, "Applying an artificial neural network to predict total body water in hemodialysis patients," *American Journal of Nephrology*, vol. 25, no. 5, pp. 507–513, 2005.

[12] A. Esteva, B. Kuprel, R. A. Novoa et al., "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[13] K. Qu, F. Guo, X. Liu, Y. Lin, and Q. Zou, "Application of machine learning in microbiology," *Frontiers in Microbiology*, vol. 10, p. 827, 2019.

[14] Q. Zhao, Y. Yang, G. Ren, E. Ge, and C. Fan, "Integrating bipartite network projection and KATZ measure to identify novel CircRNA-disease associations," *IEEE Transactions on Nanobioscience*, vol. 18, no. 4, pp. 578–584, 2019.

[15] L. Jiang, Y. Xiao, Y. Ding, J. Tang, and F. Guo, "FKL-Spa-LapRLS: an accurate method for identifying human microRNA-disease association," *BMC Genomics*, vol. 19, no. S10, p. 911, 2018.

[16] X. Zeng, L. Liu, L. Lü, and Q. Zou, "Prediction of potential disease-associated microRNAs using structural perturbation method," *Bioinformatics*, vol. 34, no. 14, pp. 2425–2432, 2018.

[17] Y. Ding, L. Jiang, J. Tang, and F. Guo, "Identification of human microRNA-disease association via hypergraph embedded bipartite local model," *Computational Biology and Chemistry*, vol. 89, p. 107369, 2020.

[18] C. Jia, Y. Zuo, and Q. Zou, "O-GlcNAcPRED-II: an integrated classification algorithm for identifying O-GlcNAcylation sites based on fuzzy undersampling and a K-means PCA oversampling technique," *Bioinformatics*, vol. 34, no. 12, pp. 2029–2036, 2018.

[19] L. Wei, S. Luan, L. A. Nagai, R. Su, and Q. Zou, "Exploring sequence-based features for the improved prediction of DNA N4-methylcytosine sites in multiple species," *Bioinformatics*, vol. 35, no. 8, pp. 1326–1333, 2018.

[20] Q. Zou, P. Xing, L. Wei, and B. Liu, "Gene2vec: gene subsequence embedding for prediction of mammalian N6-methyladenosine sites from mRNA," *RNA*, vol. 25, no. 2, pp. 205–218, 2019.

[21] C. Dai, P. Feng, L. Cui, R. Su, W. Chen, and L. Wei, "Iterative feature representation algorithm to improve the predictive performance of N7-methylguanosine sites," *Briefings in Bioinformatics*, 2020.

[22] B. Liu, S. Jiang, and Q. Zou, "HITS-PR-HHblits: protein remote homology detection by combining PageRank and hyperlink-induced topic search," *Briefings in Bioinformatics*, vol. 21, no. 1, pp. 298–308, 2018.

[23] L. Wei, Y. Ding, R. Su, J. Tang, and Q. Zou, "Prediction of human protein subcellular localization using deep learning," *Journal of Parallel and Distributed Computing*, vol. 117, pp. 212–217, 2018.

[24] Y. Ding, J. Tang, and F. Guo, "Protein crystallization identification via fuzzy model on linear neighborhood representation,"

IEEE/ACM Transactions on Computational Biology and Bioinformatics, p. 1, 2019.

[25] Y. Wang, Y. Ding, J. Tang, Y. Dai, and F. Guo, "CrystalM: a multi-view fusion approach for protein crystallization prediction," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, p. 1, 2019.

[26] H. Wang, Y. Ding, J. Tang, and F. Guo, "Identification of membrane protein types via multivariate information fusion with Hilbert–Schmidt Independence criterion," *Neurocomputing*, vol. 383, pp. 257–269, 2019.

[27] Y. Shen, Y. Ding, J. Tang, Q. Zou, and F. Guo, "Critical evaluation of web-based prediction tools for human protein subcellular localization," *Briefings in Bioinformatics*, vol. 21, no. 5, pp. 1628–1640, 2019.

[28] Y. Ding, J. Tang, and F. Guo, "Human protein subcellular localization identification via fuzzy model on kernelized neighborhood representation," *Applied Soft Computing*, vol. 96, p. 106596, 2020.

[29] R. Su, L. He, T. Liu, X. Liu, and L. Wei, "Protein subcellular localization based on deep image features and criterion learning strategy," *Briefings in Bioinformatics*, 2020.

[30] X. Ru, L. Li, and Q. Zou, "Incorporating distance-based top-n-gram and random forest to identify electron transport proteins," *Journal of Proteome Research*, vol. 18, no. 7, pp. 2931–2939, 2019.

[31] J. Liu, R. Su, J. Zhang, and L. Wei, "Classification and gene selection of triple-negative breast cancer subtype embedding gene connectivity matrix in deep neural network," *Briefings in Bioinformatics*, 2021.

[32] L. Wei, W. He, A. Malik, R. Su, L. Cui, and B. Manavalan, "Computational prediction and interpretation of cell-specific replication origin sites from multiple eukaryotes by exploiting stacking framework," *Briefings in Bioinformatics*, 2020.

[33] R. Su, T. Liu, C. Sun, Q. Jin, R. Jennane, and L. Wei, "Fusing convolutional neural network features with hand-crafted features for osteoporosis diagnoses," *Neurocomputing*, vol. 385, pp. 300–309, 2020.

[34] Y. Ding, J. Tang, and F. Guo, "Identification of drug-side effect association via multiple information integration with centered kernel alignment," *Neurocomputing*, vol. 325, pp. 211–224, 2019.

[35] Y. Ding, J. Tang, and F. Guo, "Identification of drug-side effect association via semisupervised model and multiple kernel learning," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 6, pp. 2619–2632, 2019.

[36] Y. Ding, J. Tang, and F. Guo, "Identification of drug-target interactions via dual Laplacian regularized least squares with multiple kernel fusion," *Knowledge-Based Systems*, vol. 204, p. 106254, 2020.

[37] Y. J. Ding, T. Jijun, and F. Guo, "Identification of drug-target interactions via fuzzy bipartite local model," *Neural Computing and Applications*, vol. 32, no. 14, pp. 10303–10319, 2020.

[38] X. Guo, W. Zhou, Y. Yu, Y. Ding, J. Tang, and F. Guo, "A novel triple matrix factorization method for detecting drug-side effect association based on kernel target alignment," *BioMed Research International*, vol. 2020, 11 pages, 2020.

[39] X. Guo, W. Zhou, B. Shi et al., "An efficient multiple kernel support vector regression model for assessing dry weight of hemodialysis patients," *Current Bioinformatics*, vol. 15, 2020.

[40] Y.-H. Pao, G.-H. Park, and D. J. Sobajic, "Learning and generalization characteristics of the random vector functional-link net," *Neurocomputing*, vol. 6, no. 2, pp. 163–180, 1994.

[41] J. Passauer, H. Petrov, A. Schleser, J. Leicht, and K. Pucalka, "Evaluation of clinical dry weight assessment in haemodialysis patients using bioimpedance spectroscopy: a cross-sectional study," *Nephrology Dialysis Transplantation*, vol. 25, no. 2, pp. 545–551, 2009.

[42] M. Kraemer, C. Rode, and V. Wizemann, "Detection limit of methods to assess fluid status changes in dialysis patients," *Kidney International*, vol. 69, no. 9, pp. 1609–1620, 2006.

[43] Y. Jian, X. Li, X. Cheng et al., "Comparison of bioimpedance and clinical methods for dry weight prediction in maintenance hemodialysis patients," *Blood Purification*, vol. 37, no. 3, pp. 214–220, 2014.

[44] J. S. Krouwer and K. L. Monti, "A simple, graphical method to evaluate laboratory assays," *European Journal of Clinical Chemistry and Clinical Biochemistry*, vol. 33, no. 6, pp. 525–527, 1995.

[45] K. Cha, G. M. Chertow, J. Gonzalez, J. M. Lazarus, and D. W. Wilmore, "Multifrequency bioelectrical impedance estimates the distribution of body water," *Journal of Applied Physiology*, vol. 79, no. 4, pp. 1316–1319, 1995.

[46] L. T. Ho, R. F. Kushner, D. A. Schoeller, R. Gudivaka, and D. M. Spiegel, "Bioimpedance analysis of total body water in hemodialysis patients," *Kidney international*, vol. 46, no. 5, pp. 1438–1442, 1994.

[47] Y. Wang, F. Shi, L. Cao et al., "Morphological segmentation analysis and texture-based support vector machines classification on mice liver fibrosis microscopic images," *Current Bioinformatics*, vol. 14, no. 4, pp. 282–294, 2019.

[48] M. N. F. Fajila, "Gene subset selection for leukemia classification using microarray data," *Current Bioinformatics*, vol. 14, no. 4, pp. 353–358, 2019.

[49] L. Wei, M. Liao, Y. Gao, R. Ji, Z. He, and Q. Zou, "Improved and promising identification of human microRNAs by incorporating a high-quality negative set," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 11, no. 1, pp. 192–201, 2014.

[50] L. Wei, R. Su, B. Wang, X. Li, Q. Zou, and X. Gao, "Integration of deep feature representations and handcrafted features to improve the prediction of N^6-methyladenosine sites," *Neurocomputing*, vol. 324, pp. 3–9, 2019.

[51] L. Wei, S. Wan, J. Guo, and K. K. L. Wong, "A novel hierarchical selective ensemble classifier with bioinformatics application," *Artificial Intelligence in Medicine*, vol. 83, pp. 82–90, 2017.

[52] L. Wei, P. Xing, J. Zeng, J. X. Chen, R. Su, and F. Guo, "Improved prediction of protein-protein interactions using novel negative samples, features, and an ensemble classifier," *Artificial Intelligence in Medicine*, vol. 83, pp. 67–74, 2017.

[53] L. Wei, C. Zhou, H. Chen, J. Song, and R. Su, "ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides," *Bioinformatics*, vol. 34, no. 23, pp. 4007–4016, 2018.

[54] C. Yang, Y. Ding, Q. Meng, J. Tang, and F. Guo, "Granular multiple kernel learning for identifying RNA-binding protein residues via integrating sequence and structure information," *Neural Computing and Applications*, pp. 1–13, 2021.

[55] H. Wang, J. Tang, Y. Ding, and F. Guo, "Exploring associations of non-coding RNAs in human diseases via three-matrix factorization with hypergraph-regular terms on center kernel alignment," *Briefings in Bioinformatics*, 2021.

[56] H. Wang, Y. Ding, J. Tang, Q. Zou, and F. Guo, "Identify RNA-associated subcellular localizations based on multi-label

learning using Chou's 5-steps rule," *BMC Genomics*, vol. 22, no. 1, p. 56, 2021.

[57] Y. Zou, H. Wu, X. Guo et al., "MK-FSVM-SVDD: a multiple kernel-based fuzzy SVM model for predicting DNA-binding proteins via support vector data description," *Current Bioinformatics*, vol. 15, p. 1, 2020.

[58] J. Wang, H. Wang, X. Wang, and H. Chang, "Predicting drug-target interactions via FM-DNN learning," *Current Bioinformatics*, vol. 15, no. 1, pp. 68–76, 2020.

[59] Y. Xiao, J. Wu, Z. Lin, and X. Zhao, "A deep learning-based multi-model ensemble method for cancer prediction," *Computer Methods and Programs in Biomedicine*, vol. 153, pp. 1–9, 2017.

[60] Y. Huang, K. Yuan, M. Tang et al., "Melatonin inhibiting the survival of human gastric cancer cells under ER stress involving autophagy and Ras-Raf-MAPK signalling," *Journal of Cellular and Molecular Medicine*, pp. 1–13, 2020.