frontiers in Genetics

ORIGINAL RESEARCH
published: 20 December 2021
doi: 10.3389/fgene.2021.809001

SNAREs-SAP: SNARE Proteins Identification With PSSM Profiles

Zixiao Zhang[1], Yue Gong[1], Bo Gao[2], Hongfei Li[1], Wentao Gao[1], Yuming Zhao[1]* and Benzhi Dong[1]**

[1]College of Information and Computer Engineering, Northeast Forestry University, Harbin, China, [2]Department of Radiology, The Second Affiliated Hospital, Harbin Medical University, Harbin, China

Soluble N-ethylmaleimide sensitive factor activating protein receptor (SNARE) proteins are a large family of transmembrane proteins located in organelles and vesicles. The important roles of SNARE proteins include initiating the vesicle fusion process and activating and fusing proteins as they undergo exocytosis activity, and SNARE proteins are also vital for the transport regulation of membrane proteins and non-regulatory vesicles. Therefore, there is great significance in establishing a method to efficiently identify SNARE proteins. However, the identification accuracy of the existing methods such as SNARE CNN is not satisfied. In our study, we developed a method based on a support vector machine (SVM) that can effectively recognize SNARE proteins. We used the position-specific scoring matrix (PSSM) method to extract features of SNARE protein sequences, used the support vector machine recursive elimination correlation bias reduction (SVM-RFE-CBR) algorithm to rank the importance of features, and then screened out the optimal subset of feature data based on the sorted results. We input the feature data into the model when building the model, used 10-fold crossing validation for training, and tested model performance by using an independent dataset. In independent tests, the ability of our method to identify SNARE proteins achieved a sensitivity of 68%, specificity of 94%, accuracy of 92%, area under the curve (AUC) of 84%, and Matthew's correlation coefficient (MCC) of 0.48. The results of the experiment show that the common evaluation indicators of our method are excellent, indicating that our method performs better than other existing classification methods in identifying SNARE proteins.

Keywords: SNARE proteins, position-specific scoring matrix, machine learning, support vector machine, SVM-RFE-CBR

OPEN ACCESS

Edited by:
Quan Zou,
University of Electronic Science and Technology of China, China

Reviewed by:
Yi Xiong,
Shanghai Jiao Tong University, China
Liran Juan,
Harbin Institute of Technology, China

****Correspondence:***
Zixiao Zhang
zixiao_zhang@nefu.edu.cn
Benzhi Dong
nefudbz@nefu.edu.cn
Yuming Zhao
zym@nefu.edu.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *04 November 2021*
Accepted: *15 November 2021*
Published: *20 December 2021*

Citation:
Zhang Z, Gong Y, Gao B, Li H, Gao W, Zhao Y and Dong B (2021) SNAREs-SAP: SNARE Proteins Identification With PSSM Profiles. Front. Genet. 12:809001. doi: 10.3389/fgene.2021.809001

1 INTRODUCTION

N-ethylmaleimide sensitive factor (NSF) (Whiteheart et al., 2001) protein and soluble NSF attachment proteins (SNAPS) (Whiteheart et al., 1993) are two essential factors for protein transport between membranes (Hohl et al., 1998) (Hanson et al., 1997). They were first discovered as essential proteins for protein transport from donor to receptor subcellular structures during the processes of Golgi modification and secretion. The discovery of these two proteins led to the discovery of multiple receptor proteins on transport vesicles and plasma membranes and snap receptors, which are collectively called soluble N-ethylmaleimide-sensitive factor activating protein receptor (SNARE) proteins (Ungar and Hughson, 2003; Zhao et al., 2019). According to the SNARE theory, exocytosis and secretory processes are completed by precise coordination between SNARE proteins. The specificity of membrane fusion is based on the specific

binding of SNARE protein members. At the molecular level, when the transport vesicle is close to the target membrane, syntaxin1A/B on the target membrane receives a signal to recognize, approach and combine with SNAP25, which is also located on the target membrane. At the same time, VAMP2 (q-snare) on the transport vesicle also recognizes (Kweon et al., 2003), draws close to and binds to form a 7S R-Q-SNARE complex, which guides the attachment and fusion of the transport vesicle and the target membrane, leading to the secretion of substances in the transport vesicle into the new subcellular structure or out of the cell through exocytosis, completing the intracellular transport and extracellular exocytosis and secretion processes.

The binding sites of SNARE proteins are specific, which is the reason for the specificity and precision of exocytosis and secretion in different organisms and organs (Fasshauer et al., 1998; Yin et al., 2021). SNARE theory convincingly explains the key role of synapses in the process of nerve impulse transmission at the molecular level (Chen and Scheller, 2001). Its new insights in the fields of molecular neurobiology and endocrinology have made research on SNARE proteins a hot spot in the basic life sciences worldwide. Such findings greatly enrich understanding of the regulation of intracellular information transmission, substance transport and exocytosis and secretion at the molecular level and improve knowledge of the interaction between proteins and the plasma membrane (Liu et al., 2019a; Wang et al., 2020a; Xu et al., 2021).

Due to the important roles of SNARE proteins in cell biology, research on SNARE proteins is also developing, and a variety of technologies are used to study SNARE proteins (Wang et al., 2020b; Yin et al., 2020), including the establishment of a SNARE protein database, the retrieval and classification of SNARE proteins, bioinformatics technology that was used to predict the role of SNARE proteins, and construction of a neural network model to recognize SNARE proteins.

With the development of computational biology, the application of machine learning to bioinformatics continues to be deep and widespread (Jiang et al., 2013; Tao et al., 2020; Zhao et al., 2021). Machine learning is complex and cross disciplinary across multiple fields (Cheng, 2020). Machine learning obtains new knowledge through learning from pre-existing knowledge and can continuously advance itself based on large quantities of this pre-existing knowledge and skills. Research on machine learning includes the study of computer algorithms, using data and previous techniques to improve the performance of computer algorithms. Machine learning also has significant implications for the development of artificial intelligence, through which computers continuously progress along a path of constant intelligence. A typical way to predict proteins is to transform each protein sequence into a numerical eigenvector used to represent the protein sequence, training a classification model based on the eigenvectors of the training samples and the labels. After feature construction, the classifiers that make predictions about proteins include covariant discriminant (CD) (Chou, 2000), support vector machine (SVM) (Hua and Sun, 2001), K-nearest neighbor (KNN) (Shen and Chou, 2006), deep learning and ensemble classifiers (Shen and Chou, 2006).

In this study, based on SVM classifier (Liu et al., 2010), we constructed a model to recognize SNARE proteins. We use position-specific scoring matrix (PSSM) profiles of protein sequences to extract features (Kumar et al., 2008), process the feature data by the min-max normalization method, build a model based on SVM, train the model with 10-fold cross validation and measure the performance of the model on an independent dataset.

2 MATERIALS AND METHODS

We developed a method to recognize SNARE proteins based on PSSM (Chou and Shen, 2007; Liu et al., 2019b; Hong et al., 2020a; Hong et al., 2020b) profiles and SVM. Method steps include data collection, data processing, feature extraction, feature selection, model training, and model performance evaluation. The overall flow of our designed method is summarized in **Figure 1**, and each section in the figure is described in detail in the following sections. We carried out experiments through the above process, constantly adjusted in our experiment, and finally constructed an excellent method to identify SNARE proteins. The following is a detailed description of the method.

2.1 Feature Extraction

It is very important to select good feature information for protein recognition (Zuo et al., 2017; Zheng et al., 2019; Tang et al., 2020a; Guo et al., 2020; Zhang et al., 2021). We chose the method based on PSSM profiles to extract the feature information of protein sequence data. We use the National Center for Biotechnology Information basic local alignment search tool (NCBI-BLAST) and select a non-redundant (NR) protein sequence database as a comparison dataset. We use the prepared SNARE protein FASTA sequence files to generate PSSM profiles. Each amino acid of the original sequence in the PSSM profiles consists of a vector of 20 values. Then, we transform the original PSSM files into PSSM profiles with 400 dimensions. Finally, 400-dimensional data are extracted as the feature data of each protein sequence for the experiment.

2.2 Data Processing

The feature data in the datasets are seriously unbalanced, especially the ratio of positive samples to negative samples in the independent dataset, which varies tremendously. The model would exhibit the problem of poor generalization, and the applicability would be low, so it is unable to effectively identify SNARE proteins. Therefore, we need to choose the appropriate method to deal with the data. In this study, the data processing methods we chose included Z-score standardization, min-max normalization and L2 regularization.

Normalization: Data can be changed to [0, one] ranges using the normalization method. Normalization, as an effective way to simplify calculation and scale down data values, can change the absolute values of data in the dataset into a relationship of some relative value. After normalization, the data can be calculated conveniently and quickly. This is mainly for the convenience of data processing, mapping the data to the range of 0–1, which will be convenient and fast to use. The method is defined as:

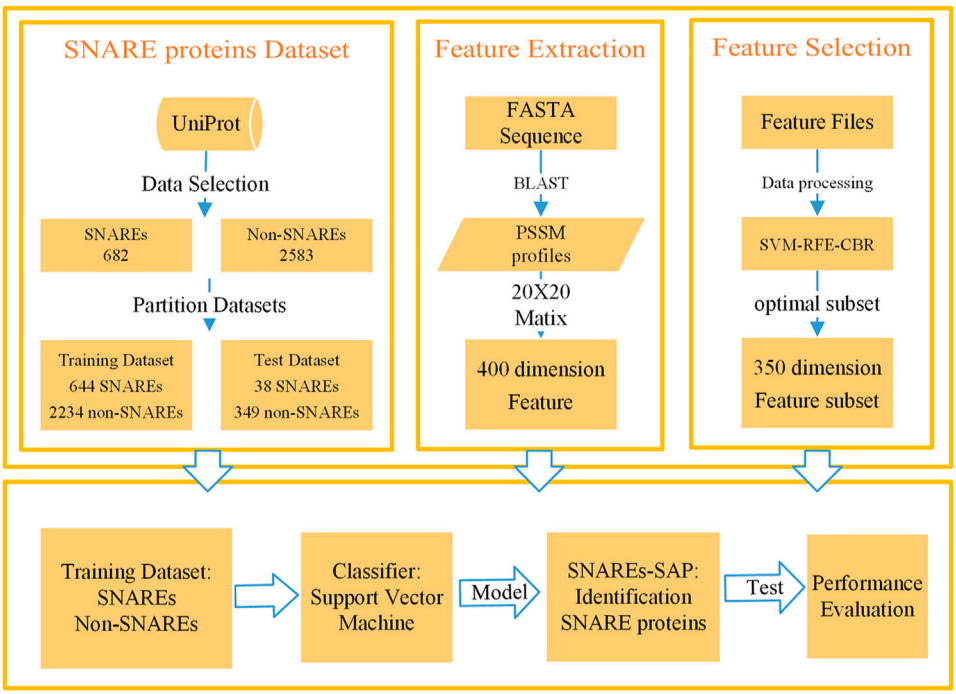

FIGURE 1 | Flow chart of SNARE proteins recognition based on PSSM profiles matrix and SVM.

$$x^{*} = \frac{x - \min}{\max - \min}, \# \quad (1)$$

The distribution of original data can be changed by normalization, and then the weights of each feature dimension can be balanced by varying the feature dimension, such as converting the distribution of data from planar to circular. Normalization can remove the influence of dimensionality on the experimental results by reducing the difference in dimensionality. After normalization, the data of different variables can be compared. Although the maximum and minimum values of the resulting data in the normalization process are affected by outliers in the dataset, and the resulting data are less robust, normalization does improve the accuracy of iterations in the operational data process as well as the efficiency of data convergence.

2.3 Feature Selection

Feature selection refers to sorting features by suitable techniques and algorithms and filtering out the better characterized subset of features based on the sorted results; this is a common technique in bioinformatics (Cheng et al., 2018; Zhu et al., 2019; Zhao et al., 2020a; Zhao et al., 2020b; Shao and Liu, 2021; Yu et al., 2021). After feature selection, the optimal feature subset selected from existing features is used to build the model, which can improve the performance of the model. Feature selection is a very important part of building models for pattern recognition and is a high priority in data processing (Wei et al., 2018; Xue et al., 2018; Li et al., 2020a; Yang et al., 2020a; Su et al., 2020; Wei et al., 2020; Yu et al., 2020; Zhang et al., 2020; Zheng et al., 2020; Wang et al., 2021a; Shang et al., 2021; Shao et al., 2021). Selecting the effective features from the original feature dataset and removing the redundant features can reduce the dimensionality of the feature data, and using more effective feature data can improve the performance of the model. Our original feature is based on PSSM to extract 400 dimensional features. In these original feature spaces, there will be irrelevant, noisy, and redundant features. Suitable feature selection methods with excellent performance are required for accurate screening of redundant features. In our experiment, we finally chose the SVM-RFE-CBR (Yan and Zhang, 2015) algorithm to screen features after comparing multiple feature selection methods. The algorithm ranks the importance of features and selects the optimal subset of features based on the sorted results.

SVM-RFE-CBR is an improved algorithm based on support vector machine recursive feature elimination (SVM-RFE), which introduces the strategy of correlation deviation reduction (CBR) into the process of feature elimination. SVM-RFE estimates feature importance based on the coefficient of the SVM model, and it is a powerful feature selection algorithm. There are linear and nonlinear versions. The SVM-RFE-CBR method adds the correlation reduction strategy (CBR) to the SVM-RFE algorithm to reduce the potential deviation of the algorithm, and the result of feature selection is improved by the integrated CBR strategy. SVM-RFE uses the sequential backward selection algorithm in SVM, which is based on the principle of maximum interval. During the model training process, SVM-RFE sort features based on the score of every feature, deletes the feature with the lowest score, puts the remaining feature data into the next round of training of the model, and finally outputs the feature sort result to

a table. The optimal feature subset can be selected according to the results of sorting. SVM is an excellent machine learning classification algorithm. The feature sort result derived from the SVM model has better performance, and it is also more convenient for subsequent experiments.

2.4 Support Vector Machine

SVM is currently a commonly used classifier in machine learning that classifies data by supervised learning (Cheng et al., 2019a; Cheng et al., 2019b). SVM is commonly used in data dichotomization. In addition, SVM can classify nonlinearly by using the kernel function (Ding et al., 2020a; Liu et al., 2020a; Yang et al., 2020b). SVM was developed from the generalized portrait algorithm in pattern recognition. The basic idea of SVM is to construct a model that separates the dataset accurately according to the geometric interval of the hyperplane with the maximum separation of samples. SVM can map the features of a dataset to points in space and draw a line to distinguish these points effectively. SVM uses a hinge loss function to computationally predict the presence of empirical risk, and a regularization term is added to ensure its robustness and correct rate. The process of SVM: Suppose the training set is $\{(x_i, y_i)\}_{i=1}^{N}$, $x_i \in R^D$,$y_i \in \{+1, -1\}$, x_i is the *i*th sample, N is the sample size, and D is the number of sample features. SVM finding the optimal classification hyperplane.$\omega \cdot x + b = 0$ The optimization problems that SVM needs to solve are:

$$\begin{aligned} &\min \frac{1}{2}||\omega||^2 + C\sum_{i=1}^{N}\varepsilon_i.\# \\ &\text{s.t. } y_i(\omega \cdot x_i + b) \geq 1 - \varepsilon_i,\ i = 1, 2, \cdots, N \\ &\varepsilon_i \geq 0,\ i = 1, 2, \cdots, N \end{aligned} \tag{2}$$

Transforming the original problem into the dual problem:

$$\min \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j(x_i \cdot x_j) - \sum_{i=1}^{N}\alpha_i . \tag{3}$$

$$\text{s.t. } \sum_{i=1}^{N} y_i\alpha_i = 0 .\# \tag{4}$$

$0 \leq \alpha_i \leq C,\ i = 1, 2, \cdots, N$ α_i is a Lagrangian

Finally, the solution of ω is:

$$\omega = \sum_{i=1}^{N}\alpha_i y_i x_i.\# \tag{5}$$

When we use SVM to solve nonlinear problems, we need to choose the appropriate kernel function (Yang et al., 2021a) (Ding et al., 2020b) and then map the data to the high-dimensional space to solve the linearly inseparable problem of the data in the original space.

In the experiment, the Python version of a library for support vector machine (LIBSVM) was selected to build an SVM model and identify SNARE proteins. The selection of different kernel functions using LIBSVM as well as the settings of kernel parameters are described as follows: The kernel function (Ding et al., 2020c) of SVM includes the linear kernel function (LKF), polynomial kernel function (PKF), radial basis function (RBF), and sigmoid kernel function (SKF). Formulas corresponding to four kernel functions are as follows:

Linear kernel function defined as:

$$K(x_i, x_j) = x_i^T x_j.\# \tag{6}$$

Polynomial kernel function:

$$K(x_i, x_j) = (\nu x_i^T x_j + r)^d,\ \nu > 0.\# \tag{7}$$

Radial basis functions:

$$K(x_i, x_j) = \exp(-\nu||x_i - x_j||^2),\ \nu > 0.\# \tag{8}$$

Sigmoid kernel function:

$$K(x_i, x_j) = \tanh(\nu x_i^T + r).\# \tag{9}$$

ν, r, and d in formulas are parameters of kernel function.

Parameters are different in different kernel functions. ν in the formula represents the parameter gamma in the kernel function, the default of which is 1/K (K is the number of classes), and g is used to set it in the LIBSVM.

r in the formula represents the parameter r in the kernel function, the default of which is 0, and r is used to set it in the LIBSVM. d in the formula represents the parameter d in the kernel function; it is used to set the highest number of times in the polynomial kernel function, and its default value is 3.

SVM is a very powerful model that allows the decision boundary to be very complex and performs well on both low-dimensional data and high-dimensional data. SVM has been widely used in bioinformatics, binding protein prediction, protein methylation site prediction and so on. We use the LIBSVM of Scikit-learn library integration in Python to train and build the model. In our experimental process, we optimize the parameters according to the results and finally build the model with the best performance.

3 RESULTS AND DISCUSSION

3.1 Dataset

Our research is devoted to constructing a method to recognize SNARE proteins. To establish a model to effectively distinguish SNARE proteins and non-SNARE proteins, we collected a SNARE protein dataset and a non-SNARE protein dataset for our prediction model. The dataset we use has been used by Le, N.Q.K. and V.-N. Nguyen (Le and Nguyen, 2019) previously. The data come from the UniProt database, which is the most informative and resource-free protein database. We collect all SNARE proteins from the UniProt database according to the keyword SNARE. To avoid the homology of the SNARE protein sequence data that we collect, we use BLAST to address the redundancy of the SNARE protein sequence and eliminate the redundant sequence. Finally, 682 SNARE protein sequences are obtained as a positive sample dataset. At the same time, we select vesicular transport proteins as negative samples to establish a non-SNARE protein dataset. We divide the two datasets into a cross-validation dataset and an independent test dataset, and the size and details of the datasets are summarized in **Table 1**.

TABLE 1 | Summary of SNARE protein and non-SNARE protein datasets.

Dataset	SNARE	Non-SANRE	Total
Original dataset	682	2,583	3,265
Train dataset	644	2,234	2,878
Test dataset	38	349	387

Table 1 shows that SNARE proteins and non-SNARE proteins correspond to two datasets: a training dataset and an independent test dataset, both of which include positive samples and negative samples. We use the cross-validation method to train the model with the training dataset, evaluate the performance of the model developed in this study, and optimize the model by adjusting the parameters according to the results of the training dataset. The independent test dataset is used to test and measure the predictive ability of the prediction model we developed.

3.2 Performance Measurements

Our research aims to establish a model to predict whether an amino acid sequence is a SNARE protein. Therefore, we need to use universally acknowledged evaluation indicators to measure the performance of the model. When training the model, we choose 10-fold cross validation as the training model after various considerations and take the average value of the crossing validation results as the result of model training. We optimize the parameters of SVM, select the best parameters to build the model, and evaluate the performance of the model through an independent test dataset to avoid systematic deviation in the process of cross validation. This study adopts some standard evaluation indicators that are widely used in bioinformatics research (Shen et al., 2019a; Shen et al., 2019b; Ao et al., 2020; Li et al., 2020b; Liu et al., 2020b; Tang et al., 2020b; Yin et al., 2020; Chen et al., 2021). The standard evaluation indicators include sensitivity (Sn), specificity (Sp), accuracy (Acc), area under the curve (AUC), Mathew's correlation coefficient (MCC), and F-score (Zhai et al., 2020; Wang et al., 2021b; Yang et al., 2021b). The calculation formulas are as follows (TP means true positive values, FP means false positive values, TN means true negative values, FN means false negative values):

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}},\ 0 \le Sn \le 1.\# \tag{10}$$

$$\text{Speccificity} = \frac{\text{TN}}{\text{TN} + \text{FP}},\ 0 \le Sp \le 1.\# \tag{11}$$

$$\text{Accurarcy} = \frac{\text{TN} + \text{TP}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}},\ 0 \le \text{Acc} \le 1.\# \tag{12}$$

$$\text{MCC} = \frac{\text{TP*TN} - \text{FP*FN}}{\sqrt{(\text{TP} + \text{FN})(\text{TN} + \text{FN})(\text{TP} + \text{FP})(\text{TN} + \text{FP})}},\ 0 \le \text{MCC} \le 1.\# \tag{13}$$

$$\text{F} - \text{score} = \frac{2\text{*TP}}{2\text{TP} + \text{FN} + \text{FP}},\ 0 \le \text{F} - \text{score} \le 1.\# \tag{14}$$

In machine learning research, receiver operating characteristic (ROC) curves are usually used to test the prediction performance of the model. AUC is a floating-point number from 0 to one of ROC. The AUC value can reflect the quality of the model. The greater the value, the better the performance of the model. ROC curves and AUCs are commonly used to compare the performance of different models as machine learning performance indicators, which is very reliable. MCC is often used to measure imbalanced data sets, which is one of the most important indicators to measure the performance of two kinds of classification in machine learning. We use Python's processing library to process data.

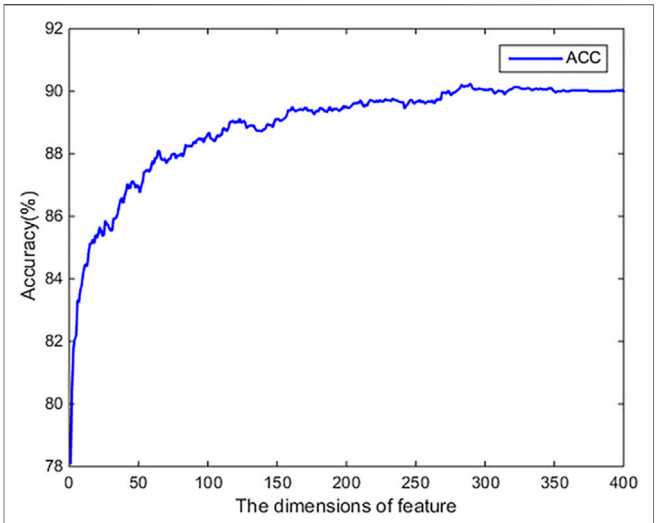

FIGURE 2 | The results of dimension reduction by using SVM-RFE-CBR algorithm.

TABLE 2 | Comparison of prediction results between SVM-RFE-CBR dimension reduction and original dimension.

Feature-dimension	Sn	Sp	Acc	AUC	MCC	F-score
350	**0.68**	**0.94**	**0.92**	**0.84**	**0.48**	**0.5**
400	0.68	0.94	0.91	0.83	0.48	0.5

Comparison of prediction results between SVM-RFE-CBR dimension reduction and original dimension. The bold values mean maximum value in the column.

3.3 Performance Comparison With Different Feature Dimensions

We use the SVM-RFE-CBR algorithm to evaluate the original 400-dimensional feature data. We use MATLAB to implement the SVM-REF-CBR algorithm to sort the features. When sorting features, a performance comparison will be given. The evaluation results are shown in **Figure 2**. From **Figure 2**, it can be found that the ACC achieved highest value, when the top 350-dimensional feature is used in the experiment. Therefore, we choose 350-dimensional feature data for the experiment.

We use the optimal 350-dimensional feature dataset after sorting for the experiment. First, 350-dimensional feature data are selected from the original feature training dataset and test dataset files according to the index obtained by the SVM-RFE-CBR algorithm. Then, the training dataset is 10-fold cross

TABLE 3 | The result of performance compares between SVM and other classification method.

	Sn	Sp	Acc	MCC
KNN	**0.870**	0.906	0.898	**0.73**
Random Forest	0.620	0.962	0.900	0.70
Naïve Bayes	0.853	0.595	0.624	0.28
SVM	0.650	**0.970**	**0.900**	0.70

The result of performances compares between SVM and other classification method. The bold values mean maximum value in the column.

validated, and the model is optimized. After many experiments, the optimal parameters of SVM are obtained. When we choose the radial basis function, penalty coefficient (C) = "11", gamma = "0.1", the model achieves the optimal performance. At the same time, we also use the original 400-dimensional feature data for the experiment and choose the optimal parameterization in the experiment. The comparison of experimental results in different dimensions is shown in **Table 2**.

The experimental results show that both Acc and MCC are improved after feature dimensionality reduction, which eliminates the redundant part of the original feature and improves the performance of the model.

3.4 Comparison of Different Classifier Performance on Dataset

With the development of computers, machine learning has been widely used in bioinformatics (Tang et al., 2019; Wang et al., 2020c; Fu et al., 2020; Cai et al., 2021; Wang et al., 2021c; Jin et al., 2021), and there are many classification models, including the linear classifier, SVM, naive byes, K-nearest neighbor (KNN), decision tree (DT), and ensemble model (random forest/GDBT, etc.). To obtain the most effective classifier method to identify SNARE proteins, we use various machine learning classifiers to construct a model of SNARE protein recognition, including random forest, KNN and naive Bayes.

We compare the experimental results of multiple machine learning classifier training models with the performance measurement results. The performance result of different classifier shown in **Table 3**.

As we can observe from **Table 3**, the results of SVM on training dataset are better than another classifier.

In particular, Sp = 0.970, Acc = 0.900. SVM shows higher performance. Meanwhile, we compare the ROC curves of different classifier method. The result shown in **Figure 3**. As we can observe from **Figure 3**, The ROC curve of SVM is obviously better than the other three classifiers.

3.5 Comparison of Different SNARE Protein Identification Methods

We compare the experimental results of SNARE CNN with the performance measurement results of our research method. The independent test results of using different methods to identify SNARE proteins are shown in **Figure 4**. **Figure 4A** shows the result of performance compares between our classification

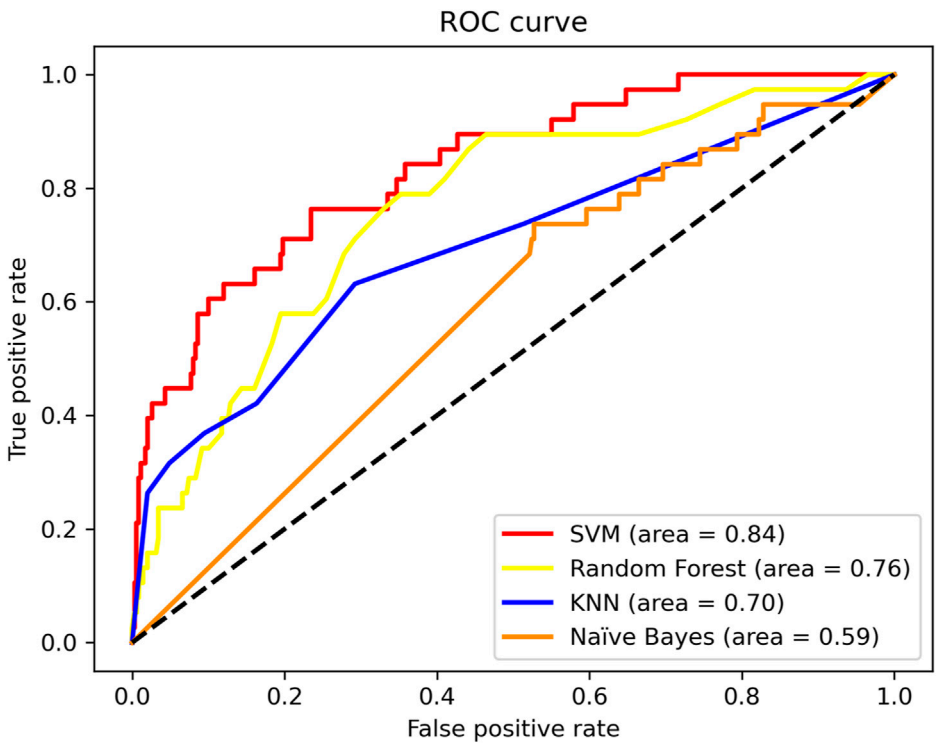

FIGURE 3 | ROC curves of different classifier methods.

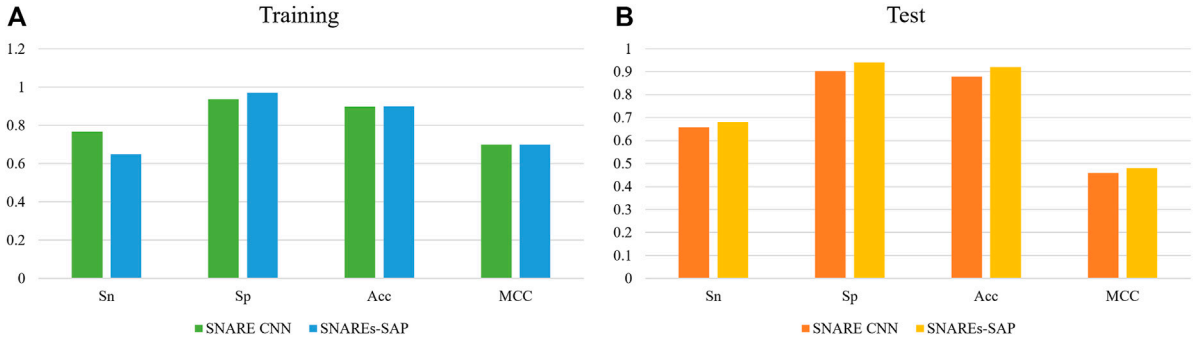

FIGURE 4 | (A) The result of performance compares between our classification method and other classification method on training datasets **(B)** The result of performance compares between our classification method and other classification method on test datasets.

method and other classification method on training datasets. **Figure 4B** shows the result of performance compares between our classification method and other classification method on test datasets.

The results show that our method gives good results in both training and independent test datasets. To compare the performance measurements of our method for identifying SNARE proteins with other methods more accurately, we compare the results of different methods on independent test datasets.As we can observe from **Figure 4B**, the independent test results of our method are better than SANRE CNN. Sn = 0.68, Sp = 0.940, Acc = 0.92 and MCC = 0.48, and all these indicators reach the highest values using our method. As shown above, our method shows higher performance. These results clearly demonstrate the superiority of our method over the existing methods, especially when using an independent dataset test. This means that our method can better recognize SNARE proteins.

4 DISCUSSION

Because of the importance of SNARE proteins and the vital significance of SNARE proteins in vesicular transport, there is an urgent need for classification methods to identify SNARE proteins. Extracting meaningful features and selecting an appropriate machine learning algorithm can greatly increase the model performance of protein prediction. We propose a method based on PSSM profiles to extract features and SVM to construct a model to identify SNARE proteins. We normalize the feature data and use the SVM-RFE-CBR algorithm to reduce the dimensions of feature. Then, we use a 10-fold crossing validation training model and use an independent dataset to test the performance of the model (Li et al., 2017; Li et al., 2020c). The accuracy, specificity, sensitivity, AUC, MCC and other performance indicators of our method have excellent experimental results. All results show that our model has better performance than other machine learning methods and advanced neural networks. Our method can effectively identify SNARE proteins. Taken together, the method proposed in our study is of great significance for the study of SNARE proteins and may also contribute to the prediction of protein function. Future works may include investigation of more kinds of proteins.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories: https://github.com/First-Leaner/Identify-proteins. The names of the repository/repositories and accession number(s) can also be found in the article/Supplementary Material.

AUTHOR CONTRIBUTIONS

ZZ and BD conceived and designed the project. ZZ, HL, and YG conducted the experiments and analyzed the data. ZZ and BG wrote the manuscript. BD, WG, and YZ revised the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by National Natural Science Foundation of China (No. 62172129).

REFERENCES

Ao, C., Zhou, W., Gao, L., Dong, B., and Yu, L. (2020). Prediction of Antioxidant Proteins Using Hybrid Feature Representation Method and Random forest. *Genomics* 112 (6), 4666–4674. doi:10.1016/j.ygeno.2020.08.016

Cai, L., Ren, X., Fu, X., Peng, L., Gao, M., and Zeng, X. (2021). iEnhancer-XG: Interpretable Sequence-Based Enhancers and Their Strength Predictor. *Bioinformatics* 37, 1060–1067. doi:10.1093/bioinformatics/btaa914

Chen, Y. A., and Scheller, R. H. (2001). SNARE-Mediated Membrane Fusion. *Nat. Rev. Mol. Cel. Biol.* 2 (2), 98–106. doi:10.1038/35052017

Chen, Y., Ma, T., Yang, X., Wang, J., Song, B., and Zeng, X. (2021). MUFFIN: Multi-Scale Feature Fusion for Drug-Drug Interaction Prediction. *Bioinformatics* 37, 2651–2658. doi:10.1093/bioinformatics/btab169

Cheng, L., Hu, Y., Sun, J., Zhou, M., and Jiang, Q. (2018). DincRNA: a Comprehensive Web-Based Bioinformatics Toolkit for Exploring Disease Associations and ncRNA Function. *Bioinformatics* 34 (11), 1953–1956. doi:10.1093/bioinformatics/bty002

Cheng, L., Yang, H., Zhao, H., Pei, X., Shi, H., Sun, J., et al. (2019). MetSigDis: a Manually Curated Resource for the Metabolic Signatures of Diseases. *Brief Bioinform.* 20 (1), 203–209. doi:10.1093/bib/bbx103

Cheng, L., Zhao, H., Wang, P., Zhou, W., Luo, M., Li, T., et al. (2019). Computational Methods for Identifying Similar Diseases. *Mol. Ther. - Nucleic Acids* 18, 590–604. doi:10.1016/j.omtn.2019.09.019

Cheng, L. (2020). Omics Data and Artificial Intelligence: New Challenges for Gene Therapy. *Cgt* 20 (1), 1. doi:10.2174/1566523220012006041500 41

Chou, K.-C., and Shen, H.-B. (2007). MemType-2L: A Web Server for Predicting Membrane Proteins and Their Types by Incorporating Evolution Information through Pse-PSSM. *Biochem. Biophys. Res. Commun.* 360 (2), 339–345. doi:10.1016/j.bbrc.2007.06.027

Chou, K.-C. (2000). Prediction of Protein Subcellular Locations by Incorporating Quasi-Sequence-Order Effect. *Biochem. Biophys. Res. Commun.* 278 (2), 477–483. doi:10.1006/bbrc.2000.3815

Ding, Y., Tang, J., and Guo, F. (2020). Identification of Drug-Target Interactions via Fuzzy Bipartite Local Model. *Neural Comput. Applic* 32 (14), 10303–10319. doi:10.1007/s00521-019-04569-z

Ding, Y., Tang, J., and Guo, F. (2020). Human Protein Subcellular Localization Identification via Fuzzy Model on Kernelized Neighborhood Representation. *Appl. Soft Comput.* 96, 106596. doi:10.1016/j.asoc.2020.106596

Ding, Y., Tang, J., and Guo, F. (2020). Identification of Drug-Target Interactions via Dual Laplacian Regularized Least Squares with Multiple Kernel Fusion. *Knowledge-Based Syst.* 204, 106254. doi:10.1016/j.knosys.2020.106254

Fasshauer, D., Sutton, R. B., Brunger, A. T., and Jahn, R. (1998). Conserved Structural Features of the Synaptic Fusion Complex: SNARE Proteins Reclassified as Q- and R-SNAREs. *Proc. Natl. Acad. Sci.* 95 (26), 15781–15786. doi:10.1073/pnas.95.26.15781

Fu, X., Cai, L., Zeng, X., and Zou, Q. (2020). StackCPPred: a Stacking and Pairwise Energy Content-Based Prediction of Cell-Penetrating Peptides and Their Uptake Efficiency. *Bioinformatics* 36 (10), 3028–3034. doi:10.1093/bioinformatics/btaa131

Guo, Z., Wang, P., Liu, Z., and Zhao, Y. (2020). Discrimination of Thermophilic Proteins and Non-thermophilic Proteins Using Feature Dimension Reduction. *Front. Bioeng. Biotechnol.* 8, 584807. doi:10.3389/fbioe.2020.584807

Hanson, P. I., Heuser, J. E., and Jahn, R. (1997). Neurotransmitter Release - Four Years of SNARE Complexes. *Curr. Opin. Neurobiol.* 7 (3), 310–315. doi:10.1016/s0959-4388(97)80057-8

Hohl, T. M., Parlati, F., Wimmer, C., Rothman, J. E., Söllner, T. H., and Engelhardt, H. (1998). Arrangement of Subunits in 20 S Particles Consisting of NSF, SNAPs, and SNARE Complexes. *Mol. Cel.* 2 (5), 539–548. doi:10.1016/s1097-2765(00)80153-7

Hong, J., Luo, Y., Zhang, Y., Ying, J., Xue, W., Xie, T., et al. (2020). Protein Functional Annotation of Simultaneously Improved Stability, Accuracy and False Discovery Rate Achieved by a Sequence-Based Deep Learning. *Brief Bioinform.* 21 (4), 1437–1447. doi:10.1093/bib/bbz081

Hong, J., Luo, Y., Mou, M., Fu, J., Zhang, Y., Xue, W., et al. (2020). Convolutional Neural Network-Based Annotation of Bacterial Type IV Secretion System Effectors with Enhanced Accuracy and Reduced False Discovery. *Brief Bioinform.* 21 (5), 1825–1836. doi:10.1093/bib/bbz120

Hua, S., and Sun, Z. (2001). Support Vector Machine Approach for Protein Subcellular Localization Prediction. *Bioinformatics* 17 (8), 721–728. doi:10.1093/bioinformatics/17.8.721

Jiang, Q., Wang, G., Jin, S., Li, Y., and Wang, Y. (2013). Predicting Human microRNA-Disease Associations Based on Support Vector Machine. *Ijdmb* 8 (3), 282–293. doi:10.1504/ijdmb.2013.056078

Jin, S., Zeng, X., Xia, F., Huang, W., and Liu, X. (2021). Application of Deep Learning Methods in Biological Networks. *Brief. Bioinform.* 22 (2), 1902–1917. doi:10.1093/bib/bbaa043

Kumar, M., Gromiha, M. M., and Raghava, G. P. S. (2008). Prediction of RNA Binding Sites in a Protein Using SVM and PSSM Profile. *Proteins* 71 (1), 189–194. doi:10.1002/prot.21677

Kweon, D.-H., Kim, C. S., and Shin, Y.-K. (2003). Regulation of Neuronal SNARE Assembly by the Membrane. *Nat. Struct. Mol. Biol.* 10 (6), 440–447. doi:10.1038/nsb928

Le, N. Q. K., and Nguyen, V.-N. (2019). SNARE-CNN: a 2D Convolutional Neural Network Architecture to Identify SNARE Proteins from High-Throughput Sequencing Data. *PeerJ Comp. Sci.* 5, e177. doi:10.7717/peerj-cs.177

Li, B., Tang, J., Yang, Q., Li, S., Cui, X., Li, Y., et al. (2017). NOREVA: Normalization and Evaluation of MS-based Metabolomics Data. *Nucleic Acids Res.* 45 (W1), W162–W170. doi:10.1093/nar/gkx449

Li, F., Zhou, Y., Zhang, X., Tang, J., Yang, Q., Zhang, Y., et al. (2020). SSizer: Determining the Sample Sufficiency for Comparative Biological Study. *J. Mol. Biol.* 432 (11), 3411–3421. doi:10.1016/j.jmb.2020.01.027

Li, J., Pu, Y., Tang, J., Zou, Q., and Guo, F. (2020). DeepATT: a Hybrid Category Attention Neural Network for Identifying Functional Effects of DNA Sequences. *Brief Bioinform.* 22 (3), bbaa159. doi:10.1093/bib/bbaa159

Li, J., Pu, Y., Tang, J., Zou, Q., and Guo, F. (2020). DeepAVP: a Dual-Channel Deep Neural Network for Identifying Variable-Length Antiviral Peptides. *IEEE J. Biomed. Health Inform.* 24 (10), 3012–3019. doi:10.1109/jbhi.2020.2977091

Liu, T., Zheng, X., and Wang, J. (2010). Prediction of Protein Structural Class for Low-Similarity Sequences Using Support Vector Machine and PSI-BLAST Profile. *Biochimie* 92 (10), 1330–1334. doi:10.1016/j.biochi.2010.06.013

Liu, D., Li, G., and Zuo, Y. (2019). Function Determinants of TET Proteins: the Arrangements of Sequence Motifs with Specific Codes. *Brief Bioinform.* 20 (5), 1826–1835. doi:10.1093/bib/bby053

Liu, B., Gao, X., and Zhang, H. (2019). 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 Res.* 47 (20), e127. doi:10.1093/nar/gkz740

Liu, B., Li, C.-C., and Yan, K. (2020). DeepSVM-fold: Protein Fold Recognition by Combining Support Vector Machines and Pairwise Sequence Similarity Scores Generated by Deep Learning Networks. *Brief. Bioinform.* 21 (5), 1733–1741. doi:10.1093/bib/bbz098

Liu, B., Zhu, Y., and Yan, K. (2020). Fold-LTR-TCP: Protein Fold Recognition Based on Triadic Closure Principle. *Brief. Bioinform.* 21 (6), 2185–2193. doi:10.1093/bib/bbz139

Shang, Y., Gao, L., Zou, Q., and Yu, L. (2021). Prediction of Drug-Target Interactions Based on Multi-Layer Network Representation Learning. *Neurocomputing* 434, 80–89. doi:10.1016/j.neucom.2020.12.068

Shao, J., and Liu, B. (2021). ProtFold-DFG: Protein Fold Recognition by Combining Directed Fusion Graph and PageRank Algorithm. *Brief. Bioinform.* 22, bbaa192. doi:10.1093/bib/bbaa192

Shao, J., Yan, K., and Liu, B. (2021). FoldRec-C2C: Protein Fold Recognition by Combining Cluster-To-Cluster Model and Protein Similarity Network. *Brief. Bioinform.* 22 (3), bbaa144. doi:10.1093/bib/bbaa144

Shen, H.-B., and Chou, K.-C. (2006). Ensemble Classifier for Protein Fold Pattern Recognition. *Bioinformatics* 22 (14), 1717–1722. doi:10.1093/bioinformatics/btl170

Shen, Y., Tang, J., and Guo, F. (2019). Identification of Protein Subcellular Localization via Integrating Evolutionary and Physicochemical Information into Chou's General PseAAC. *J. Theor. Biol.* 462, 230–239. doi:10.1016/j.jtbi.2018.11.012

Shen, Y., Ding, Y., Tang, J., Zou, Q., and Guo, F. (2019). Critical Evaluation of Web-Based Prediction Tools for Human Protein Subcellular Localization. *Brief. Bioinform.* 21 (5), 1628–1640. doi:10.1093/bib/bbz106

Su, R., Liu, X., and Wei, L. (2020). MinE-RFE: Determine the Optimal Subset from RFE by Minimizing the Subset-Accuracy-Defined Energy. *Brief. Bioinform.* 21 (2), 687–698. doi:10.1093/bib/bbz021

Tang, J., Fu, J., Wang, Y., Luo, Y., Yang, Q., Li, B., et al. (2019). Simultaneous Improvement in the Precision, Accuracy, and Robustness of Label-free Proteome Quantification by Optimizing Data Manipulation Chains*. *Mol. Cell Proteomics* 18 (8), 1683–1699. doi:10.1074/mcp.ra118.001169

Tang, Y.-J., Pang, Y.-H., and Liu, B. (2020). IDP-Seq2Seq: Identification of Intrinsically Disordered Regions Based on Sequence to Sequence Learning. *Bioinformaitcs* 36 (21), 5177–5186. doi:10.1093/bioinformatics/btaa667

Tang, J., Fu, J., Wang, Y., Li, B., Li, Y., Yang, Q., et al. (2020). ANPELA: Analysis and Performance Assessment of the Label-free Quantification Workflow for Metaproteomic Studies. *Brief Bioinform.* 21 (2), 621–636. doi:10.1093/bib/bby127

Tao, Z., Li, Y., Teng, Z., and Zhao, Y. (2020). A Method for Identifying Vesicle Transport Proteins Based on LibSVM and MRMD. *Comput. Math. Methods Med.* 2020, 8926750. doi:10.1155/2020/8926750

Ungar, D., and Hughson, F. M. (2003). SNARE Protein Structure and Function. *Annu. Rev. Cel Dev. Biol.* 19 (1), 493–517. doi:10.1146/annurev.cellbio.19.110701.155609

Wang, Z., Liu, D., Xu, B., Tian, R., and Zuo, Y. (2020). Modular Arrangements of Sequence Motifs Determine the Functional Diversity of KDM Proteins. *Brief Bioinform.* 22 (3), bbaa215. doi:10.1093/bib/bbaa215

Wang, H., Ding, Y., Tang, J., and Guo, F. (2020). Identification of Membrane Protein Types via Multivariate Information Fusion with Hilbert-Schmidt Independence Criterion. *Neurocomputing* 383, 257–269. doi:10.1016/j.neucom.2019.11.103

Wang, Y., Zhang, S., Li, F., Zhou, Y., Zhang, Y., Wang, Z., et al. (2020). Therapeutic Target Database 2020: Enriched Resource for Facilitating Research and Early Development of Targeted Therapeutics. *Nucleic Acids Res.* 48 (D1), D1031–D1041. doi:10.1093/nar/gkz981

Wang, H., Jijun, T., Ding, Y., and Guo, F. (2021). Exploring Associations of Non-coding RNAs in Human Diseases via Three-Matrix Factorization with Hypergraph-Regular Terms on center Kernel Alignment. *Brief. Bioinform.* 22 (5), bbaa409. doi:10.1093/bib/bbaa409

Wang, X., Yang, Y., Liu, J., and Wang, G. (2021). The Stacking Strategy-Based Hybrid Framework for Identifying Non-coding RNAs. *Brief Bioinform.* 22 (5), bbab023. doi:10.1093/bib/bbab023

Wang, D., Zhang, Z., Jiang, Y., Mao, Z., Wang, D., Lin, H., et al. (2021). DM3Loc: Multi-Label mRNA Subcellular Localization Prediction and Analysis Based on Multi-Head Self-Attention Mechanism. *Nucleic Acids Res.* 49 (8), e46. doi:10.1093/nar/gkab016

Wei, L., Chen, H., and Su, R. (2018). M6APred-EL: A Sequence-Based Predictor for Identifying N6-Methyladenosine Sites Using Ensemble Learning. *Mol. Ther. - Nucleic Acids* 12, 635–644. doi:10.1016/j.omtn.2018.07.004

Wei, L., Hu, J., Li, F., Song, J., Su, R., Zou, Q., et al. (2020). Comparative Analysis and Prediction of Quorum-sensing Peptides Using Feature Representation Learning and Machine Learning Algorithms. *Brief. Bioinform.* 21 (1), 106–119. doi:10.1093/bib/bby107

Whiteheart, S. W., Griff, I. C., Brunner, M., Clary, D. O., Mayer, T., Buhrow, S. A., et al. (1993). SNAP Family of NSF Attachment Proteins Includes a Brain-specific Isoform. *Nature* 362 (6418), 353–355. doi:10.1038/362353a0

Whiteheart, S. W., Schraw, T., and Matveeva, E. A. (2001). N-ethylmaleimide Sensitive Factor (NSF) Structure and Function. *Int. Rev. Cytol.* 207, 71–112. doi:10.1016/s0074-7696(01)07003-6

Xu, B., Liu, D., Wang, Z., Tian, R., and Zuo, Y. (2021). Multi-substrate Selectivity Based on Key Loops and Non-homologous Domains: New Insight into ALKBH Family. *Cell. Mol. Life Sci.* 78 (1), 129–141. doi:10.1007/s00018-020-03594-9

Xue, W., Yang, F., Wang, P., Zheng, G., Chen, Y., Yao, X., et al. (2018). 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 Chem. Neurosci.* 9 (5), 1128–1140. doi:10.1021/acschemneuro.7b00490

Yan, K., and Zhang, D. (2015). Feature Selection and Analysis on Correlated Gas Sensor Data with Recursive Feature Elimination. *Sens. Actuators B: Chem.* 212, 353–363. doi:10.1016/j.snb.2015.02.025

Yang, Q., Li, B., Tang, J., Cui, X., Wang, Y., Li, X., et al. (2020). Consistent Gene Signature of Schizophrenia Identified by a Novel Feature Selection Strategy from Comprehensive Sets of Transcriptomic Data. *Brief Bioinform.* 21 (3), 1058–1068. doi:10.1093/bib/bbz049

Yang, Q., Wang, Y., Zhang, Y., Li, F., Xia, W., Zhou, Y., et al. (2020). NOREVA: Enhanced Normalization and Evaluation of Time-Course and Multi-Class Metabolomic Data. *Nucleic Acids Res.* 48 (W1), W436–W448. doi:10.1093/nar/gkaa258

Yang, C., Ding, Y., Meng, Q., Tang, J., and Guo, F. (2021a). Granular Multiple Kernel Learning for Identifying RNA-Binding Protein Residues via Integrating Sequence and Structure Information. *Neural Comput. Appl.* 33, 11387–11399. doi:10.1007/s00521-020-05573-4

Yang, H., Luo, Y., Ren, X., Wu, M., He, X., Peng, B., et al. (2021). Risk Prediction of Diabetes: Big Data Mining with Fusion of Multifarious Physical Examination Indicators. *Inf. Fusion* 75, 140–149. doi:10.1016/j.inffus.2021.02.015

Yin, J., Sun, W., Li, F., Hong, J., Li, X., Zhou, Y., et al. (2020). VARIDT 1.0: Variability of Drug Transporter Database. *Nucleic Acids Res.* 48 (D1), D1042–D1050. doi:10.1093/nar/gkz779

Yin, J., Li, F., Zhou, Y., Mou, M., Lu, Y., Chen, K., et al. (2021). INTEDE: Interactome of Drug-Metabolizing Enzymes. *Nucleic Acids Res.* 49 (D1), D1233–D1243. doi:10.1093/nar/gkaa755

Yu, L., Shi, Q., Wang, S., Zheng, L., and Gao, L. (2020). Exploring Drug Treatment Patterns Based on the Action of Drug and Multilayer Network Model. *Ijms* 21 (14), 5014. doi:10.3390/ijms21145014

Yu, L., Wang, M., Yang, Y., Xu, F., Zhang, X., Xie, F., et al. (2021). Predicting Therapeutic Drugs for Hepatocellular Carcinoma Based on Tissue-specific Pathways. *Plos Comput. Biol.* 17 (2), e1008696. doi:10.1371/journal.pcbi.1008696

Zhai, Y., Chen, Y., Teng, Z., and Zhao, Y. (2020). Identifying Antioxidant Proteins by Using Amino Acid Composition and Protein-Protein Interactions. *Front. Cel Dev. Biol.* 8, 591487. doi:10.3389/fcell.2020.591487

Zhang, J., Zhang, Z., Pu, L., Tang, J., and Guo, F. (2020). AIEpred: an Ensemble Predictive Model of Classifier Chain to Identify Anti-Inflammatory Peptides. *Ieee/acm Trans. Comput. Biol. Bioinform.* PP, 1. doi:10.1109/TCBB.2020.2968419

Zhang, D., Chen, H. D., Zulfiqar, H., Yuan, S. S., Huang, Q. L., Zhang, Z. Y., et al. (2021). iBLP: An XGBoost-Based Predictor for Identifying Bioluminescent Proteins. *Comput. Math. Methods Med.* 2021, 6664362. doi:10.1155/2021/6664362

Zhao, J., Sun, T., Wu, S., and Liu, Y. (2019). High Mobility Group Box 1: An Immune-Regulatory Protein. *Cgt* 19 (2), 100–109. doi:10.2174/1566523219666190621111604

Zhao, T., Hu, Y., Peng, J., and Cheng, L. (2020). DeepLGP: a Novel Deep Learning Method for Prioritizing lncRNA Target Genes. *Bioinformatics* 36 (16), 4466–4472. doi:10.1093/bioinformatics/btaa428

Zhao, X., Jiao, Q., Li, H., Wu, Y., Wang, H., Huang, S., et al. (2020). ECFS-DEA: an Ensemble Classifier-Based Feature Selection for Differential Expression Analysis on Expression Profiles. *BMC Bioinform.* 21 (1), 43. doi:10.1186/s12859-020-3388-y

Zhao, X., Wang, H., Li, H., Wu, Y., and Wang, G. (2021). Identifying Plant Pentatricopeptide Repeat Proteins Using a Variable Selection Method. *Front. Plant Sci.* 12, 506681. doi:10.3389/fpls.2021.506681

Zheng, L., Huang, S., Mu, N., Zhang, H., Zhang, J., Chang, Y., et al. (2019). RAACBook: a Web Server of Reduced Amino Acid Alphabet for Sequence-Dependent Inference by Using Chou's Five-step Rule. *Database (Oxford)* 2019, baz131. doi:10.1093/database/baz131

Zheng, L., Liu, D., Yang, W., Yang, L., Zuo, Y., et al. (2020). RaacLogo: a New Sequence Logo Generator by Using Reduced Amino Acid Clusters. *Brief Bioinform.* 22 (3), bbaa096. doi:10.1093/bib/bbaa096

Zhu, X.-J., Feng, C.-Q., Lai, H.-Y., Chen, W., and Hao, L. (2019). Predicting Protein Structural Classes for Low-Similarity Sequences by Evaluating Different Features. *Knowledge-Based Syst.* 163, 787–793. doi:10.1016/j.knosys.2018.10.007

Zuo, Y., Li, Y., Chen, Y., Li, G., Yan, Z., and Yang, L. (2017). PseKRAAC: a Flexible Web Server for Generating Pseudo K-Tuple Reduced Amino Acids Composition. *Bioinformatics* 33 (1), 122–124. doi:10.1093/bioinformatics/btw564

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.