Frontiers in Bioscience Landmark

Front. Biosci. (Landmark Ed) **2023**; 28(12): 322
https://doi.org/10.31083/j.fbl2812322

Original Research

ProSE-Pero: Peroxisomal Protein Localization Identification Model Based on Self-Supervised Multi-Task Language Pre-Training Model

Jianan Sui[1,†], Jiazi Chen[2,†], Yuehui Chen[3,*], Naoki Iwamori[2,*], Jin Sun[4]

[1]School of Information Science and Engineering, University of Jinan, 250022 Jinan, Shandong, China
[2]Laboratory of Zoology, Graduate School of Bioresource and Bioenvironmental Sciences, Kyushu University, Fukuoka-shi, 819-0395 Fukuoka, Japan
[3]School of Artificial Intelligence Institute and Information Science and Engineering, University of Jinan, 250022 Jinan, Shandong, China
[4]School of Computer Science and Engineering, University of Electronic Science and Technology of China, 611731 Chengdu, Sichuan, China
*Correspondence: yhchen@ujn.edu.cn (Yuehui Chen); iwamori@agr.kyushu-u.ac.jp (Naoki Iwamori)
[†]These authors contributed equally.
Academic Editor: Haseeb Ahmad Khan
Submitted: 27 April 2023 Revised: 17 July 2023 Accepted: 24 July 2023 Published: 1 December 2023

Abstract

Background: Peroxisomes are membrane-bound organelles that contain one or more types of oxidative enzymes. Aberrant localization of peroxisomal proteins can contribute to the development of various diseases. To more accurately identify and locate peroxisomal proteins, we developed the ProSE-Pero model. **Methods**: We employed three methods based on deep representation learning models to extract the characteristics of peroxisomal proteins and compared their performance. Furthermore, we used the SVMSMOTE balanced dataset, SHAP interpretation model, variance analysis (ANOVA), and light gradient boosting machine (LightGBM) to select and compare the extracted features. We also constructed several traditional machine learning methods and four deep learning models to train and test our model on a dataset of 160 peroxisomal proteins using tenfold cross-validation. **Results**: Our proposed ProSE-Pero model achieves high performance with a specificity (Sp) of 93.37%, a sensitivity (Sn) of 82.41%, an accuracy (Acc) of 95.77%, a Matthews correlation coefficient (MCC) of 0.8241, an F1 score of 0.8996, and an area under the curve (AUC) of 0.9818. Additionally, we extended our method to identify plant vacuole proteins and achieved an accuracy of 91.90% on the independent test set, which is approximately 5% higher than the latest iPVP-DRLF model. **Conclusions**: Our model surpasses the existing In-Pero model in terms of peroxisomal protein localization and identification. Additionally, our study showcases the proficient performance of the pre-trained multitasking language model ProSE in extracting features from protein sequences. With its established validity and broad generalization, our model holds considerable potential for expanding its application to the localization and identification of proteins in other organelles, such as mitochondria and Golgi proteins, in future investigations.

Keywords: peroxisomal localization identification; SVMSMOTE; multitasking language model; feature selection; deep learning; vacuole proteins identification

1. Introduction

Organelle proteins are a diverse group of proteins that are either bound to or distributed throughout different regions of the organelle [1]. Their presence is essential for the organelle to carry out a range of life-sustaining activities. Each organelle protein has a specific biological function that contributes to the overall functionality of the organelle [2]. Accurate identification of organelle protein types is crucial for researchers to gain a deeper understanding of their roles and to develop effective treatment strategies for diseases. Moreover, precise knowledge of the spatial distribution of organelle proteins is essential for their functional characterization. This knowledge has far-reaching implications for advancing our understanding of cell biology and developing targeted therapeutic interventions.

Most studies on identifying the localization of organelle proteins rely on machine-learning approaches. For instance, Zhou *et al*. [3] introduced a novel method for predicting Golgi protein types, which integrates pseudo amino acid composition (PseAAC), dipeptide composition (DC), pseudo-position specific scoring matrix (PsePSSM), and encoding based on grouped weight (EBGW) to extract feature vectors. The authors employed the extreme gradient boosting (XGBoost) algorithm as a classifier and achieved an impressive overall prediction accuracy of 92.1% in the internal validation using the training set, surpassing the performance of existing state-of-the-art methods. However, when evaluating the model's generalization ability on an independent test set, the accuracy drops to 86.5%. This discrepancy suggests that further improvements are needed to enhance the method's performance. Lv *et al*. [4] developed a Golgi protein classifier called rfGPT, which employs 2-gap dipeptide and split amino acid composition as feature vectors. The authors utilized the SMOTE technique to balance the dataset and analysis of variance (ANOVA) as the feature selection method and then input the selected features into the random forest (RF) model. The independent test accuracy of rfGPT was found to be 90.6%. While rfGP presents itself as a practical tool that eliminates the need

for location-specific scoring matrices and their derived features, the lower accuracy observed on the independent test set suggests that further enhancements are required in the tool's feature fusion methodology. In another study, Yu *et al.* [5] proposed SubMito-XGBoost, an XGBoost-based method for predicting protein submitochondrial type, using two training datasets, M317 and M983. The SubMito-XGBoost method demonstrated high prediction accuracies of 97.7% and 98.9%, respectively, on these datasets while achieving a prediction accuracy of 94.8% on an independent test set, M495. While SubMito-XGBoost exhibits improvements in the accuracy of protein submitochondrial prediction to some extent, there remains significant potential for further enhancement in both prediction accuracy and algorithm efficiency. Numerous other studies have also investigated the identification of organelle proteins [6–8].

In this paper, we studied the localization identification of peroxisomal proteins. Peroxisomes, also known as microbodies, are important organelles surrounded by a monolayer of membranes containing one or more oxidases. Peroxisomes play an important role in regulating cellular immunity and cancers characterized by metabolic abnormalities [9]. These cancers include prostate cancer [10,11], bladder cancer [12], and so on. Human peroxisomal malfunction can result in certain diseases, such as Alzheimer's disease and X-linked adrenoleukodystrophy (X-ALD) [13]. At present, the treatment of these diseases mainly utilizes different chemical drugs, such as anti-inflammatory and neuroprotective therapy, but in most cases, these treatments cannot provide a permanent cure [14–17]. Therefore, it is very important to detect abnormalities and injuries in time. Accurate identification and localization of peroxisomal proteins play an important role and significance in the treatment of corresponding diseases. However, the problem of localization and recognition of peroxisomal proteins has received too little attention. At present, the localization and identification tool of peroxisomal proteins is only In-Pero, constructed by Anteghini *et al.* [18] in 2021. They utilized deep learning embedding methods UniRep [19] and SeqVec [20] to extract the characteristics of peroxisomal protein sequences and compared four different machine learning methods, namely logistic regression (LR), random forest (RF), support vector machine (SVM) and partial least squares discriminant analysis (PLS-DA). By combining five protein embedding methods, a cross-validation classification accuracy of 0.92 was ultimately achieved. This work became the first work on this topic and provided a complete method and benchmark.

In this work, we proposed the ProSE-Pero model, which utilized the deep learning method to locate and iden-

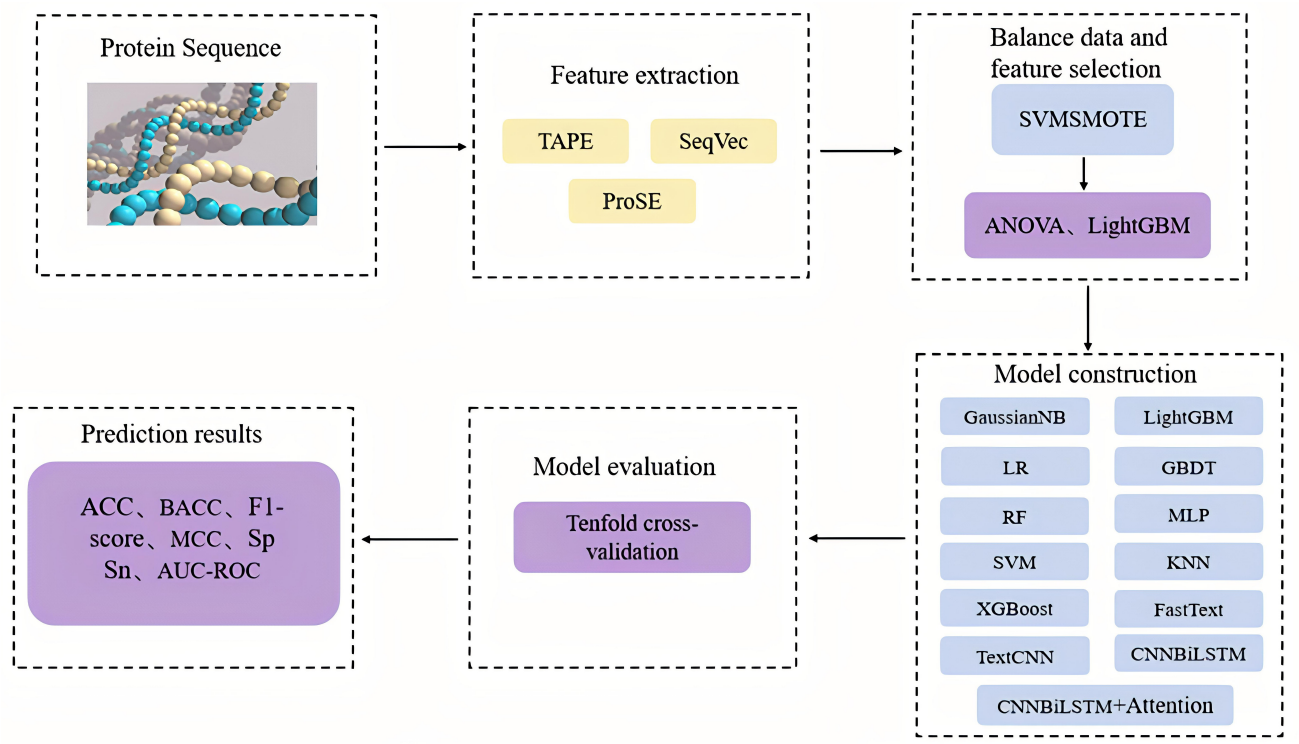

Fig. 1. Flow chart of the ProSE-Pero model. GaussianNB, Gaussian Naive Bayes; LR, Logistic Regression; RF, Random Forest; SVM, Support Vector Machine; LightGBM, Light Gradient Boosting; GBDT, Gradient Boosted Decision Trees; MLP, Multilayer Perceptron; KNN, K-Nearest Neighbors; Acc, Accuracy; BACC, Balanced Accuracy; Sn, Sensitivity; Sp, Specificity; MCC, Matthews correlation coefficient; TP, True Positive; FP, False Positive; TN, True Negative; FN, False Negative; ROC-AUC, Receiver Operating Characteristic - Area Under the Curve; PR-AUC, Precision-Recall Area Under the Curve.

tify peroxisomal proteins for the first time. We utilized three deep representation learning models to extract the features of peroxisome protein sequences. These three methods include SeqVec [20], which is based on the ELMO model, TAPE [21], which is based on the BERT model; and ProSE, which is based on a pre-trained multi-task language model [22]. In order to address the issue of imbalanced data, the SVMSMOTE technique was employed to balance the dataset. Furthermore, variance analysis using ANOVA [23] and a light gradient boosting machine (LightGBM) were utilized to select the most informative features from the extracted feature set. At the same time, these feature extraction and feature selection methods were compared. Finally, the selected features were applied to nine traditional machine learning methods and four deep learning methods. The overall flowchart of the ProSE-Pero model is shown in Fig. 1.

2. Materials and Methods

2.1 Datasets

2.1.1 Peroxisomal Datasets

The selection of appropriate datasets is a crucial step in the classification model and has a significant impact on the model's performance. In this study, we utilized the peroxisomal protein dataset, as constructed by Anteghini *et al.* [18] in 2021, which was obtained from the UniprotKB/SwissProt database (https://www.uniprot.org/) [24]. After filtering the data, CD-HIT [25] was applied for clustering with a sequence similarity threshold of 40%. The final dataset comprised 132 peroxisomal membrane protein sequences and 28 peroxisome matrix protein sequences, resulting in an imbalanced dataset with a ratio of approximately 5:1 between the two classes. This observation underscores the importance of addressing class imbalance when training classification models.

2.1.2 Vacuole Datasets

In the study of plant vacuole protein identification, we used the data set collected by Yadav *et al.* [26] to train and test the model. Both PVPs and non-PVPs are from the UniProtKB/SwissProt database [24]. They utilized CD-HIT software to remove redundant samples by setting the sequence identity threshold to 60%. A total of 274 positive and 274 negative samples were initially obtained. Subsequently, a sequence identity threshold of 40% was applied, resulting in the screening of 200 out of the 274 PVPs as positive samples for the training set, while the remaining PVPs were assigned as positive samples for the test set. Similarly, the same number of 40% identical negative samples were collected to construct balanced training and independent test datasets, respectively, as shown in Fig. 2.

2.2 Feature Extraction

In previous models, feature extraction is mainly based on component features, location features, physical and chemical properties, etc. In recent years, with the continuous maturity and development of deep learning methods, deep learning has begun to be applied to sequence-based protein characterization tasks [27–32]. Natural language processing (NLP) has received more and more attention in the field of protein sequence analysis in bioinformatics [33]. To obtain a vector representation of a protein sequence, the sequence is treated as a sentence, where an amino acid or k-mers is treated as a word [34,35].

In this work, we utilized SeqVec, ProSE, and TAPE, three feature extraction methods based on NLP pre-training models; we utilized the idea of transfer learning. And we will introduce these three feature extraction methods.

2.2.1 SeqVec

This feature extraction method utilizes the deep bidirectional model ELMo, commonly used in natural language processing (NLP), to represent protein sequences as continuous vectors known as embeddings. ELMo effectively captures the biophysical properties of protein sequences by leveraging unlabeled large-scale data. It employs a probability distribution model to generate embeddings that incorporate evolutionary information. The trained model captures important biophysical properties from the unlabeled database (UniRef50) and transfers this knowledge to individual protein sequences by predicting relevant sequence characteristics [20].

2.2.2 ProSE

The feature extraction method uses three learning tasks to simultaneously train a three-layer bidirectional LSTM with skip connections: (a) Masked language modeling task; (b) Contact prediction between residues in protein structure; (c) Structural similarity prediction. Training protein language models by self-supervised learning of large amounts of natural sequence data and structural supervision of smaller sequence sets [22]. The authors believed that prior knowledge of protein function and structure could be encoded into the learned representation through supervised training of structural similarity tasks.

2.2.3 TAPE

With the continuous development of protein representation learning in machine learning research, the author introduced a task to evaluate protein embedding (TAPE). The author selected supervised tasks based on three areas of protein biology where self-supervised learning can lead to improvements (structural prediction, remote identification, protein engineering). In this paper, we chose the BERT-based TAPE model.

Each organelle protein sequence is first converted to an integer sequence according to the following function:

$$f(m_j) = i \tag{1}$$

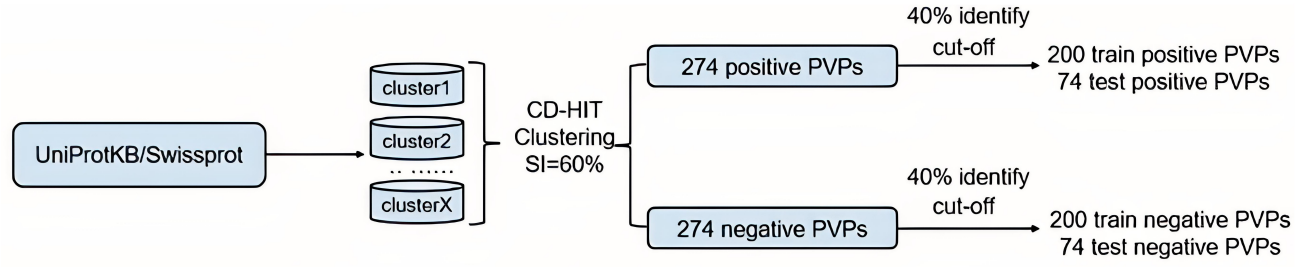

Fig. 2. Flow chart of vacuole datasets construction. PVPs, plant vacuole proteins.

$$i = 1, 2 \ldots\ldots, 20, \textit{if } 20 \textit{ canonical amino acid} \quad (2)$$

where m_j is the j th amino acid of the sequence, The integer sequence f $(\mathrm{m_j})$, j = 1,2,3,4, ……L (length of protein sequence) was embedded into 1024-long feature vectors via the SeqVec method, 6165-long feature vectors via the ProSE method, 768-long feature vectors via the TAPE method.

2.3 Feature Selection

Since the extracted features may have redundant information to make the prediction results inaccurate, it may also lead to overfitting problems. We employed the SHAP interpretation model visualization technique to identify the feature dimension that strongly influences prediction results. Subsequently, we used ANOVA [23] and LightGBM to select the relevant features within this dimension and compared their performance by incorporating them into the classifier. The better feature selection method is selected from the two and utilized as the feature extraction method of the model.

2.4 Balanced Dataset

Since we utilized the peroxisomal protein data set constructed by Anteghini *et al.* [18] in 2021, there are 132 membrane protein sequences and 28 matrix proteins, and the ratio of the two is about 5:1. There is an imbalance in the data set; and unbalanced data sets will affect the performance of the model. The SMOTE algorithm is a method for random oversampling of samples, and it is also a common method for processing unbalanced data. In this work, we utilized the SVMSMOTE algorithm, which focused on adding a few points along the decision boundary [36].

2.5 Classification Model

In the construction and selection of classification models, we first constructed nine traditional machine learning models, including Gaussian naive Bayes (GaussianNB), LR, RF, SVM, LightGBM, gradient tree boosting (GBDT), multilayer perceptron (MLP), k-nearest neighbor (KNN), and XGBoost. These models were implemented through the scikit-learn [37], and we fine-tuned their hyperparameters through grid search to achieve the best possible performance. In this study, we fed feature vectors of peroxisomal proteins into different algorithms and compared their performance to select the most effective one.

In the past studies of protein identification and localization, most of them utilized traditional machine learning methods as classification models, and there was almost no use of deep learning methods as classification models, but we believe that deep learning methods as classification models will also achieve good results, not worse than traditional machine learning methods. So we tried to construct a deep learning model; first, because the ordered amino acids of a protein can be seen as words in a sentence, we see the protein as a 'language', so we can model it using neural structures developed for natural language. Therefore we constructed TextCNN [38] and FastText [39] models. In the construction of TextCNN, the model is composed of two superimposed CNN layers, a maximum pool layer, and two linear layers. We embed each protein sequence into a matrix X of 1 $\times$ M dimension, M is the dimension of feature extraction, the batch size is 8, the step distance of the first convolution layer is 2, the convolution kernel size is 3, the CNN layer is followed by a maximum pool layer, the step distance is 3, the same as the second convolution layer, the maximum pool layer. Since the first two have achieved good results, we combined CNN and BiLSTM [40] to construct a CNNBiLSTM model [41], which uses a convolutional layer, a maximum pool layer, BiLSTM, and two fully connected layers. The convolutional layer has a step distance of 2, the convolution kernel size is 3, and the CNN layer is followed by a maximum pool layer with a step distance of 3. The size of the hidden layer of the BiLSTM is 100, the number of cycles is set to 1, and the dropout random inactivation is 0.5. At the same time, we also tried to add an attention mechanism to quantify the degree to which each part of the protein sequence is focused, as shown in Fig. 3. Each depth model uses the Adam optimizer, the learning rate is 0.001, the activation function is the Softmax function, and the loss function is the cross entropy loss function.

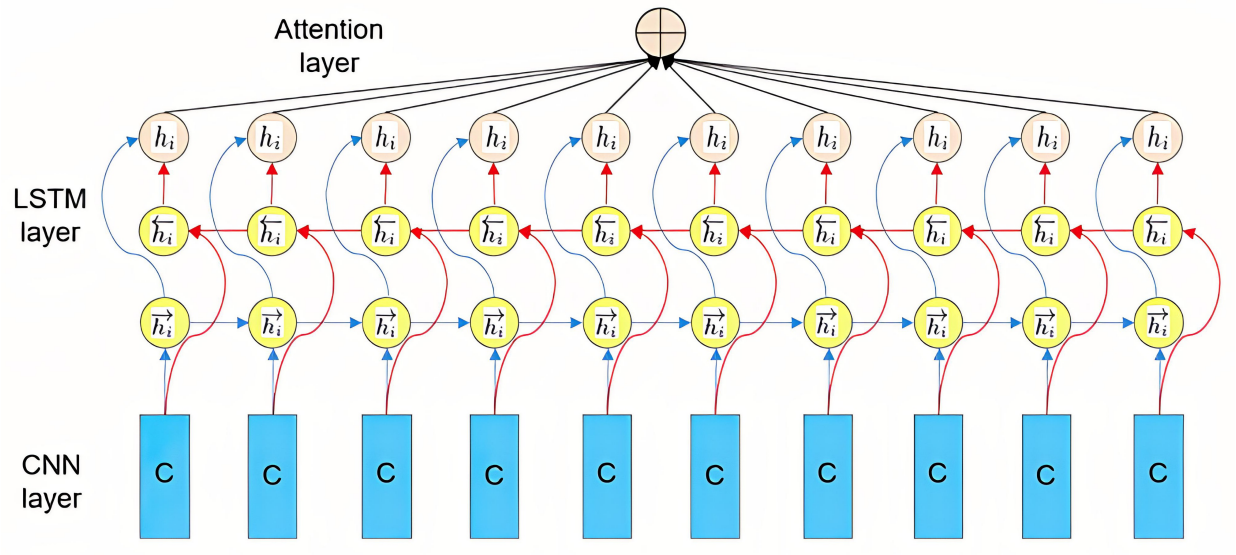

Fig. 3. CNNBiLSTM + Attention model. CNN, Convolutional Neural Network; LSTM, Long Short-Term Memory.

2.6 Evaluation Metrics and Methods

Accuracy (Acc), sensitivity (Sn), specificity (Sp), Matthews correlation coefficient (MCC), and F1-score were used to evaluate the performance of the prediction system [42–48]. The calculation method is as follows:

$$Sp = \frac{TN}{TN + FP} \tag{3}$$

$$Sn = \frac{TP}{TP + FN} \tag{4}$$

$$Acc = \frac{TP + TN}{TP + FN + TN + FP} \tag{5}$$

$$F1 = \frac{2 \times TP}{2 \times TP + FN + FP} \tag{6}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FN) \times (TN + FP)}} \tag{7}$$

For a binary classification problem, the actual prediction will have only two values, 0 and 1. True class (TP) if the instance is positive and is predicted to be positive, false positive class (FP) if the instance is negative and is predicted to be positive, and negative class if the instance is negative and is predicted to be negative. Sn, Sp are the proportion of correct predictions in positive and negative samples, respectively. The F1 score reflects the robustness of the model. The higher the score, the more robust the model is. Acc reflects the overall accuracy of the predictor. When the data set is unbalanced, Acc cannot really assess the quality of the classification results. In this case, it can be evaluated by MCC. The horizontal axis of the receiver operating characteristic (ROC) curve is generally the ratio of false positive rate(FPR), i.e., the ratio of negative class samples being judged as positive class samples, and the vertical axis is the ratio of true positive rate(TPR), i.e., the ratio of positive class samples being judged as positive class samples. In addition, we also draw the PR curve. The vertical axis of the curve is precision, and the horizontal axis is recall. In this paper, area under the curve (AUC) defaults to ROC-AUC. ROC-AUC represents the area under the ROC curve, and the higher the value, the better the model. Like ROC-AUC, we can calculate the area under the PR curve to describe the performance of the model. We can think of PRAUC as the average precision calculated for each Recall threshold. In this study, we utilized the PyCharm software, specifically version 2020.3.2, developed by JetBrains, to write the model code. The software originates from Prague, Czech Republic.

3. Results

3.1 Experiment on Peroxisome Protein Dataset

3.1.1 Performance of Features Extracted by Different Methods on Different Classification Models after Balancing the Dataset

In this study, three feature extraction methods, namely SeqVec [18] based on the ELMO model, TAPE [19] based on the BERT model, and ProSE [20] based on a pre-trained multi-task language model, were employed to extract features from peroxisomal protein sequences. To address class imbalance, the SVMSMOTE algorithm was utilized to balance the dataset. Subsequently, the extracted features were inputted into nine traditional machine learning models, in-

Table 1. SeqVec + SVMSMOTE.

Model	Acc	F1-score	Sp	Sn	MCC	AUC
GaussianNB	0.6332	0.6639	0.7959	0.5636	0.3291	0.7679
LR	0.7339	0.8216	0.3262	0.9462	0.3734	0.8652
RF	0.8797	0.9140	0.6931	0.9748	0.7099	0.8829
SVM	0.8922	0.9223	0.7759	0.9461	0.7383	0.9200
LightGBM	0.8930	0.9228	0.7736	0.9486	0.7385	0.9153
GBDT	0.8934	0.9225	0.7744	0.9465	0.7372	0.9106
MLP	0.8695	0.9025	0.8334	0.8808	0.6990	0.9047
KNN	0.7739	0.8079	0.8292	0.7482	0.5612	0.8666
XGBoost	0.7789	0.8187	0.7852	0.7750	0.5476	0.9036

LR, logistic regression; RF, random forest; SVM, support vector machine; GBDT, gradient tree boosting; MLP, multilayer perceptron; KNN, k-nearest neighbor; Acc, accuracy; Sp, specificity; MCC, Matthews correlation coefficient; AUC, area under the curve; LightGBM, light gradient boosting machine.

Table 2. TAPE + SVMSMOTE.

Model	Acc	F1-score	Sp	Sn	MCC	AUC
GaussianNB	0.8933	0.8889	0.8844	0.8978	0.7954	0.9316
LR	0.9390	0.9356	0.9838	0.8948	0.8844	0.9619
RF	0.9427	0.9360	0.9775	0.9061	0.8927	0.9917
SVM	0.9447	0.9392	0.9852	0.9035	0.8961	0.9783
LightGBM	0.9428	0.9378	0.9819	0.9022	0.8914	0.9823
GBDT	0.9409	0.9355	0.9793	0.9014	0.8878	0.9832
MLP	0.9275	0.9308	0.9383	0.9108	0.8590	0.9238
KNN	0.9125	0.9043	0.9656	0.8543	0.8350	0.9345
XGBoost	0.9048	0.8940	0.9598	0.8463	0.8178	0.9551

cluding GaussianNB, LR, RF, SVM, LightGBM, GBDT, MLP, KNN, and XGBoost. The experimental results, as summarized in Table 1, Table 2, and Table 3, revealed that the ProSE feature extraction method outperformed the other two methods across all nine traditional machine learning models. Notably, LightGBM achieved the highest performance on the tenfold cross-validation, with an accuracy of 95.22%, F1-score of 0.9510, specificity of 96.63%, sensitivity of 93.71%, MCC of 0.9072, and AUC of 0.9901. Moreover, the TAPE method in combination with SVM achieved an accuracy of 94.47%, F1-score of 0.9392, specificity of 98.52%, sensitivity of 90.35%, MCC of 0.8961, and AUC of 0.9783 on the tenfold cross-validation. Finally, the SeqVec method combined with the GBDT model demonstrated the best performance on the tenfold cross-validation, yielding an accuracy of 89.34%, F1-score of 0.9225, specificity of 77.44%, sensitivity of 94.65%, MCC of 0.7372, and AUC of 0.9106.

3.1.2 Performance of Features Extracted by Different Methods on Different Classification Models after Feature Selection

In the next step, we conducted experiments on the features extracted by the ProSE method and utilized the SHAP interpretation model to plot all instances. In this way, we can see that the size of the feature's impact on the prediction is shown in Fig. 4. Each row in the figure represents a feature and the abscissa is the Shap value. The ranking of features is based on the average absolute value of Shap, which can be seen as an arrangement diagram of feature importance. The 4614th dimension feature shown in the figure is the most important feature of the model and has the greatest impact on the results. The features of the first N that have the greatest impact on the model are generally obtained by the mean of the absolute values of each feature (abs $\rightarrow$ mean ()). The absolute value is used to solve the problem of positive and negative cancellation, and the size of the correlation is more concerned; as shown in Fig. 5, it can be seen from the figure that the first 4614 dimensional features have the greatest effect on the model. Combining the results of the first two graphs, we selected the features of the first 4616.

It is evident that the 4614-dimensional features have a significant impact on the prediction results. Therefore, the features extracted using the ProSE method were subjected to feature selection using ANOVA and LightGBM, resulting in a feature dimension of 4614. These selected features were then fed into nine traditional machine learning models and four deep learning models, and the results are presented in Table 4 and Table 5. The ANOVA and LightGBM feature selection methods exhibit varying performances across different models. Notably, for the FastText model, the

Table 3. ProSE + SVMSMOTE.

Model	Acc	F1-score	Sp	Sn	MCC	AUC
GaussianNB	0.8937	0.8921	0.8901	0.8923	0.7952	0.8941
LR	0.9473	0.9463	0.9656	0.9230	0.8974	0.9770
RF	0.9434	0.9435	0.9531	0.9331	0.8894	0.9893
SVM	0.9472	0.9460	0.9653	0.9288	0.8973	0.9869
LightGBM	0.9522	0.9510	0.9663	0.9371	0.9072	0.9901
GBDT	0.9443	0.9437	0.9525	0.9336	0.8909	0.9795
MLP	0.9434	0.9420	0.9502	0.9293	0.8909	0.9799
KNN	0.9415	0.9382	0.9670	0.9103	0.8870	0.9787
XGBoost	0.9344	0.9309	0.9621	0.9018	0.8726	0.9858

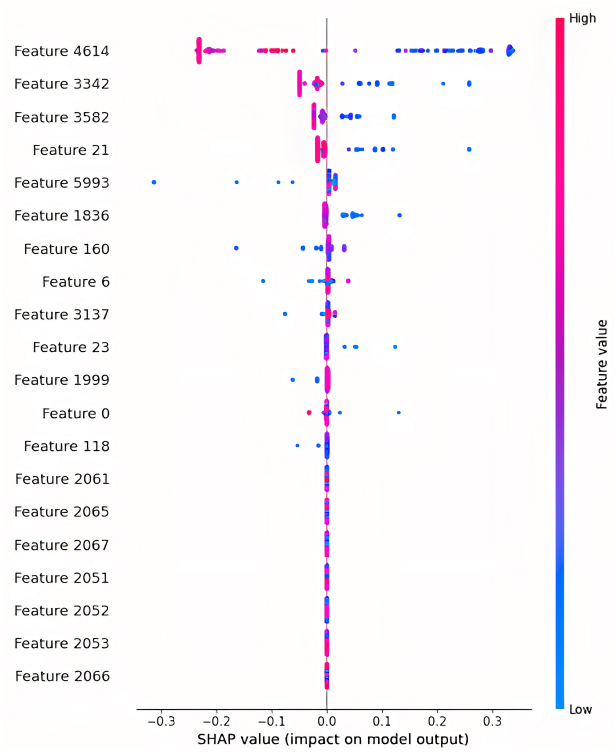

Fig. 4. Size of influence of features on prediction.

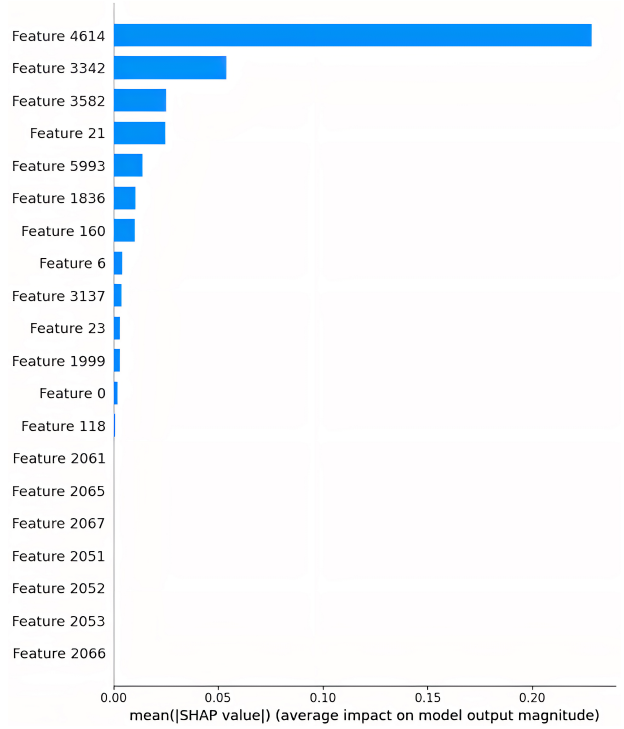

Fig. 5. Size of influence of mean prediction of feature absolute values.

ANOVA feature selection method demonstrates the best performance on the tenfold cross-validation, achieving an accuracy of 95.77%, F1-score of 0.8996, specificity of 93.37%, sensitivity of 82.41%, MCC of 0.8241, AUC of 0.9818, and PRAUC of 0.9880. Conversely, for the traditional machine learning model LightGBM, the ANOVA feature selection method yields the best results on the tenfold cross-validation, with an accuracy of 95.22%, F1-score of 0.9514, specificity of 96.75%, sensitivity of 93.39%, MCC of 0.9068, AUC of 0.9925, and PRAUC of 0.9924. Furthermore, the LightGBM method demonstrates the best performance when combined with the MLP model in the traditional machine learning setting, achieving an accuracy of 95.46%, F1-score of 0.9524, specificity of 97.47%, sensitivity of 92.93%, MCC of 0.9119, AUC of 0.9759, and PRAUC of 0.9811 on the tenfold cross-validation.

At the same time, we also draw the ROC and PR curves of the deep learning model after ANONA and LightGBM feature selection methods, as shown in Fig. 6, Fig. 7, Fig. 8, and Fig. 9.

To evaluate the impact of the SVMSMOTE method on data set balancing, we conducted experiments without incorporating SVMSMOTE-generated features into the FastText model. The results, as depicted in Fig. 10, clearly demonstrate significant improvements in model performance after applying the SVMSMOTE method. This observation confirms the crucial role of data set balancing in enhancing model indicators.

3.1.3 Comparison with Previous Models

Finally, our ProSE-Pero model was compared with the In-Pero model developed by Anteghini *et al.* [18] in 2021,

Table 4. ProSE + SVMSMOTE + ANOVA.

Model	Acc	F1-score	Sp	Sn	MCC	AUC	PRAUC
GaussianNB	0.8900	0.8896	0.8824	0.8923	0.7882	0.8948	0.9184
LR	0.9509	0.9494	0.9747	0.9238	0.9050	0.9758	0.9691
RF	0.9396	0.9388	0.9531	0.9234	0.8801	0.9923	0.9539
SVM	0.9472	0.9458	0.9684	0.9239	0.8966	0.9910	0.9730
LightGBM	0.9522	0.9514	0.9675	0.9339	0.9068	0.9925	0.9924
GBDT	0.9489	0.9487	0.9619	0.9343	0.9003	0.9868	0.9296
MLP	0.9433	0.9428	0.9565	0.9230	0.8900	0.9871	0.9560
KNN	0.9415	0.9386	0.9701	0.9071	0.8866	0.9849	0.9671
XGBoost	0.9345	0.9318	0.9620	0.9029	0.8724	0.9899	0.9671
FastText	0.9577	0.8996	0.9337	0.8841	0.8241	0.9818	0.9880
TextCNN	0.9430	0.8798	0.9314	0.8610	0.7971	0.9617	0.9745
CNNBiLSTM	0.9450	0.8921	0.9260	0.8763	0.8108	0.9715	0.9790
CNNBiLSTM + Attention	0.9434	0.8787	0.9186	0.8638	0.7928	0.9733	0.9792

ANOVA, analysis of variance.

Table 5. ProSE + SVMSMOTE + LightGBM.

Model	Acc	F1-score	Sp	Sn	MCC	AUC	PRAUC
GaussianNB	0.8708	0.8693	0.8741	0.8619	0.7480	0.8817	0.9210
LR	0.9472	0.9447	0.9747	0.9147	0.8975	0.9741	0.7500
RF	0.9511	0.9502	0.9622	0.9398	0.9051	0.9939	0.9883
SVM	0.9529	0.9512	0.9734	0.9321	0.9087	0.9842	0.9886
LightGBM	0.9535	0.9524	0.9687	0.9371	0.9098	0.9884	0.9923
GBDT	0.9481	0.9473	0.9623	0.9323	0.8991	0.9822	0.9689
MLP	0.9546	0.9524	0.9747	0.9293	0.9119	0.9759	0.9811
KNN	0.9471	0.9434	0.9792	0.9103	0.8976	0.9766	0.9994
XGBoost	0.9396	0.9371	0.9673	0.9078	0.8825	0.9844	0.9994
FastText	0.9507	0.8899	0.9325	0.8749	0.8159	0.9833	0.9884
TextCNN	0.9477	0.8883	0.9177	0.8778	0.8032	0.9549	0.9555
CNNBiLSTM	0.9438	0.8907	0.9156	0.8803	0.8023	0.9708	0.9813
CNNBiLSTM + Attention	0.9493	0.8926	0.8926	0.8905	0.8078	0.9681	0.9763

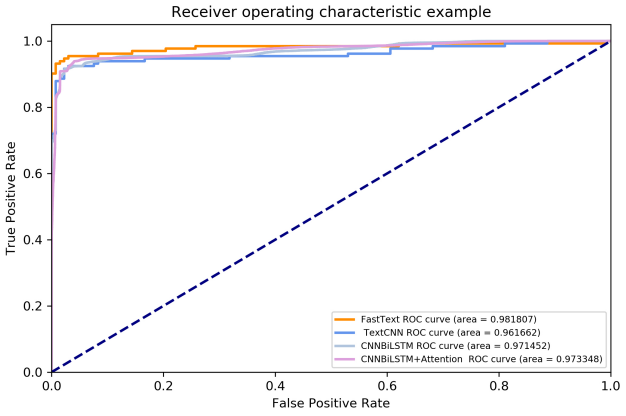

Fig. 6. ROC curve for deep learning models based on ProSE + SVMSMOTE + ANOVA.

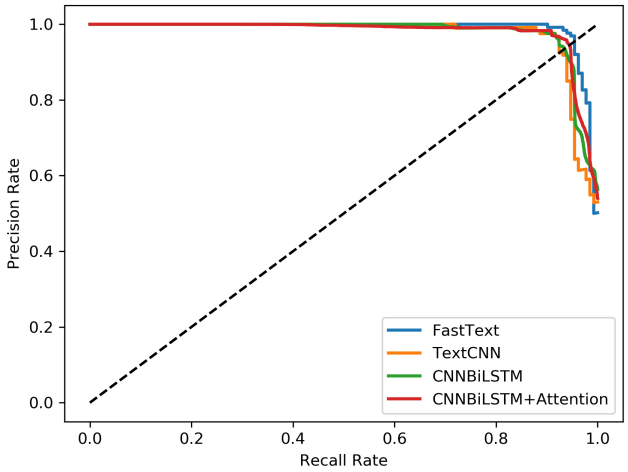

Fig. 7. PR curve for deep learning models based on ProSE + SVMSMOTE + ANOVA.

as depicted in Fig. 11. The comparison clearly illustrates that our proposed ProSE-Pero model achieves an approximately 4% higher accuracy than the In-Pero model. This

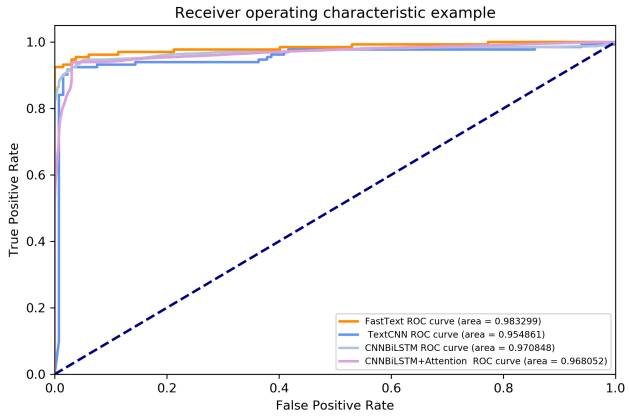

Fig. 8. ROC curve for deep learning models based on ProSE + SVMSMOTE + LightGBM.

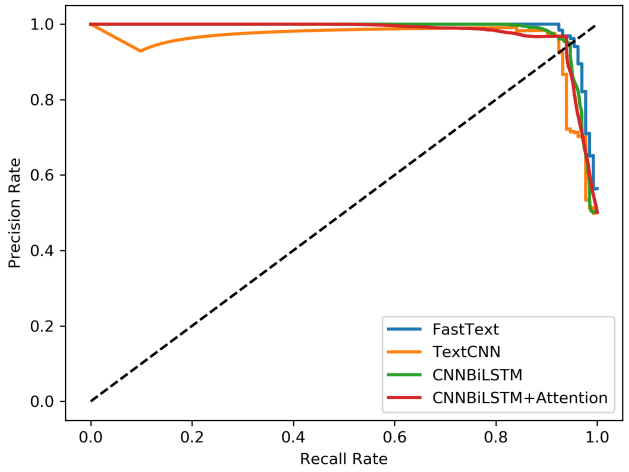

Fig. 9. PR curve for deep learning models based on ProSE + SVMSMOTE + LightGBM.

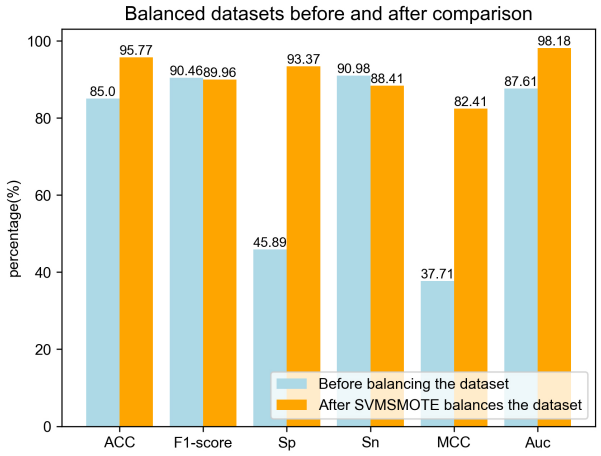

Fig. 10. Performance of FastText model before and after balancing the dataset.

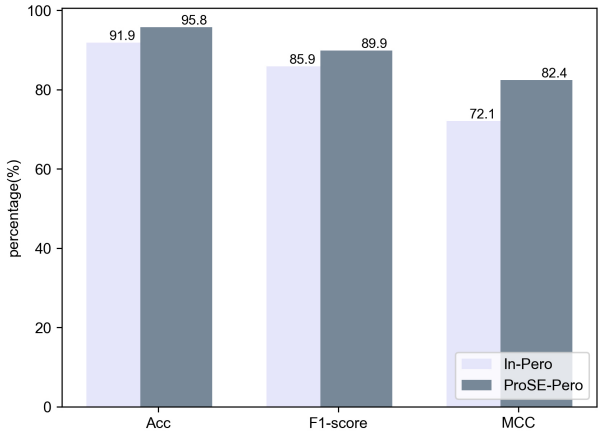

Fig. 11. Performance of In-Pero and ProSE-Pero on the peroxisomal protein dataset.

notable improvement underscores the effectiveness of our model. The detailed parameters of our ProSE-Pero model are provided in Table 6.

3.2 Extending ProSE-Pero to Predict Vacuole Proteins

3.2.1 Comparison of the Performance of Different Classification Models on the Plant Vacuole Protein Independent Data Set

Vacuoles are unique organelles in plant cells and play a key role in plant growth and development. Vacuoles have cell functions such as degradation, autolysis, and regulation. The basis for studying the maintenance mechanism of vacuole biogenesis is to understand the biochemical and physiological functions of vacuole proteins [49–51]. Accurate identification of vacuolar proteins plays an important role in understanding their biological properties. But now, there are few tools for identifying vacuolar proteins [52–54].

In order to verify the generalization performance of our model and find an effective way to identify plant vacuole proteins. We extended our method to the identification of vacuole proteins and utilized the ProSE method based on the self-supervised multi-task language pre-training model to extract the features of vacuole protein sequences. By using the SHAP interpretable model and ANOVA method to select the extracted features, we can see the size of the influence of the features on the prediction, as shown in Fig. 12, and select 606-dimensional data.

Subsequently, we conducted a comparative analysis of the performance of nine traditional machine learning models and the deep learning model FastText on the independent test set. As shown in Table 7, FastText exhibited superior performance, achieving an accuracy of 91.90%, F1-score of 0.9122, specificity of 86.64%, sensitivity of 97.05%, MCC of 0.8379, and AUC of 0.9626 on the independent dataset. Notably, among the nine traditional machine learning models, LightGBM demonstrated the highest accuracy of 89.19%, along with an F1-score of 0.9000, specificity of 81.08%, sensitivity of 97.30%, MCC of 0.7943, and AUC of 0.9573 on the independent dataset.

Table 6. The parameters of Our ProSE-Pero model.

Parameter Name	Parameter value
Hidden Units	256
Learn Rate	0.001
Dropout Rate	0.5
Activation Function	Relu
Optimizer	Adam

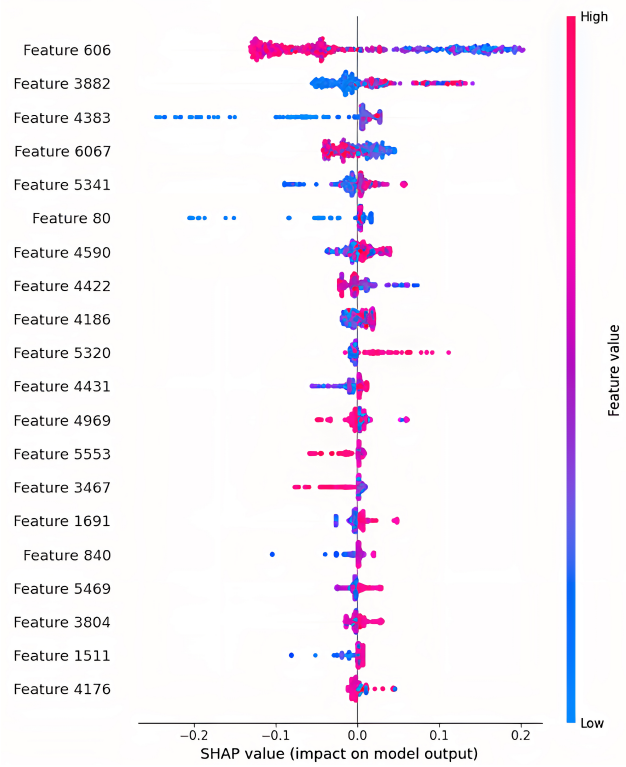

Fig. 12. Size of influence of features on prediction.

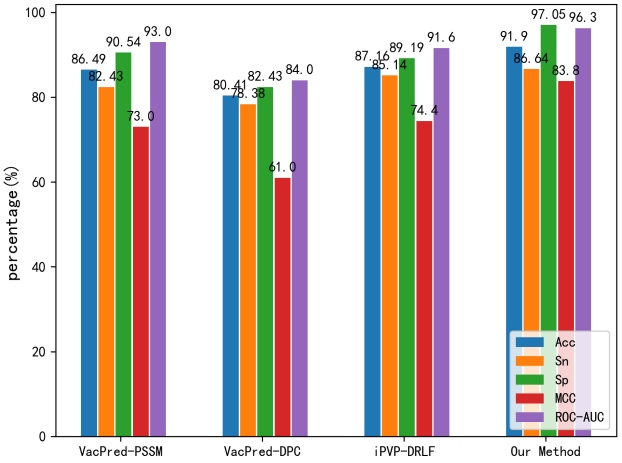

Fig. 13. Performance of our method with previous models on the plant vacuole protein independent data set.

3.2.2 Comparison with Previous Models

Finally, we compared our method with the previous vacuole protein identification model. As shown in Fig. 13, it can be seen that our method is superior to the iPVP-DRLF model [48] and the previous model in Acc Sp, Sn, MCC, and AUC, which are about 5%, 2%, 8%, 0.1 and 0.05 higher than the PVP-DRLF model respectively.

4. Discussion

The experimental results of our study have demonstrated the effectiveness of our approach, which utilizes the multi-training task pre-training model ProSE, in extracting peroxisomal and plant vacuole proteins. These findings hold significant biomedical implications as they provide insights into the understanding of protein localization and function within specific organelles. Moreover, the success of our approach opens up avenues for its application in extracting features of proteins from other organelles.

The accurate identification and localization of organelle proteins play a crucial role in unraveling the biological functions of organelles. For instance, dysregulation of the Golgi apparatus has been implicated in various genetic and neurodegenerative disorders, including diabetes [55], cancer [56], Alzheimer's disease [57], and Parkinson's disease [58]. Although current therapeutic strategies primarily rely on pharmacological interventions such as anti-inflammatory and neuroprotective treatments, they often fall short of providing a definitive cure [3]. To gain deeper insights into Golgi dysfunction, timely detection of abnormalities and damage is of utmost importance. Hence, precise identification of Golgi-resident protein types holds significant potential in advancing our understanding of the roles played by Golgi proteins in the aforementioned pathologies. Mitochondria, essential organelles in eukaryotic cells, play critical roles in various physiological processes, including cell differentiation, cellular signaling, apoptosis, and growth [5]. Impaired mitochondrial function disrupts energy metabolism and ultimately leads to cell death [59]. Aberrant identification and localization of submitochondrial proteins can lead to detrimental interactions, thereby contributing to the onset and progression of various disorders, including Parkinson's disease [60], multifactorial diseases [61], and type II diabetes [62], among others. Therefore, investigating the subcellular localization of mitochondrial proteins holds significant importance in elucidating the molecular mechanisms underlying these diseases, facilitating their diagnosis, and fostering the development of novel therapeutic interventions. Vacuoles, being the largest organelle in plants, play a pivotal role in diverse cellular functions such as the storage of inorganic ions and metabolites, protein degradation, detoxification, and the regulation of cytoplasmic ionic homeostasis [63]. These vital functions contribute to the overall cellular integrity and homeostasis in plants. Accurate identification of plant vacuole proteins and subsequent exploration of their biochemical properties and physiological functions serve as fundamental steps toward understanding

Table 7. Comparison of different models.

Model	Acc (%)	F1-score	Sp (%)	Sn (%)	MCC	AUC
GaussianNB	79.73	0.7887	83.78	75.68	0.5966	0.8456
LR	83.78	0.8421	81.08	86.49	0.6767	0.9255
RF	85.81	0.8758	83.78	90.54	0.7449	0.9385
SVM	87.16	0.8645	81.08	90.54	0.7194	0.9368
LightGBM	89.19	0.9000	81.08	97.30	0.7943	0.9573
GBDT	86.49	0.8734	79.73	93.24	0.7365	0.9556
MLP	89.19	0.8987	82.43	95.95	0.7910	0.9272
KNN	88.51	0.8903	83.78	93.24	0.7737	0.9546
XGBoost	83.11	0.8344	81.08	85.14	0.6627	0.9262
FastText	91.90	0.9122	86.64	97.05	0.8379	0.9626

the mechanisms underlying vacuole biogenesis and maintenance [53]. In this study, we have demonstrated the validity and broad generalizability of our proposed ProSE-Pero model. The ProSE-Pero model presented in this study holds significant potential for its application in accurately identifying and precisely localizing the organelle proteins mentioned above, including submitochondrial proteins and Golgi proteins. This model offers promising prospects for future studies in this field, allowing for an improved understanding of the roles and functions of these organelle proteins in various cellular processes.

However, it is important to acknowledge the limitations of our research. Currently, our focus is primarily on the identification of organelle proteins, and our methods may not be directly applicable to other protein prediction tasks. Further research and refinement are needed to expand the scope of our methods to encompass other protein-related analyses, such as protein function prediction, protein folding studies, solubility prediction, and drug design.

By addressing these limitations and advancing our methods, we aim to contribute to the broader field of proteomics and facilitate advancements in protein analysis and prediction. Ultimately, our research holds the potential to enable more accurate and comprehensive investigations into protein structure, function, and their roles in biological processes, ultimately benefiting biomedical research and applications.

5. Conclusions

Through this study, we discovered that the ProSE method, which is based on a self-supervised multi-task language pre-training model, is highly effective in identifying peroxisomal protein localization. In addition to traditional machine learning methods, we also utilized deep learning methods such as FastText, TextCNN, CNNBiLSTM, and CNNBiLSTM with an attention mechanism. Our deep learning methods achieved accuracy rates of over 94% in peroxisomal protein localization and identification, yielding impressive results. After balancing the dataset with SVMSMOTE and comparing feature selection methods such as ANOVA and LGBM, our approach achieved 95.77% in Acc, 0.8996 in F1-score, 93.37% in Sp, 82.41% in Sn, 0.8241 in MCC, and 0.9818 in AUC on the FastText model using tenfold cross-validation. These results represent a 4% improvement over the In-Pero model proposed by Anteghini *et al.* [18] in 2021, placing our approach at the forefront of peroxisome protein localization and identification research. This study highlights the importance of balancing imbalanced datasets and utilizing feature selection methods to enhance model performance. Moreover, in comparison with the In-Pero model that combines the SeqVec method and UniRep method, our approach only uses ProSE as the feature extraction method, demonstrating the superior performance of the ProSE method in peroxisomal protein localization and identification.

Furthermore, our approach has also been extended to identify vacuolar proteins in plant organelles. Notably, our method achieved remarkable results on the independent test set using the FastText model, with an accuracy of 91.90%, F1-score of 0.9122, specificity of 86.64%, sensitivity of 97.05%, MCC of 0.8379, and AUC of 0.9626, which is approximately four percentage points higher than the iPVP-DRLF model ACC proposed by Jiao *et al.* [54] in 2022. Moreover, the method we utilize in the ProSE-Pero model has demonstrated excellent effectiveness and generalization, as evidenced by the leading level of performance achieved on the independent test set for tonoplast proteins.

The above results show that the ProSE method based on a self-supervised multi-task language pre-training model has a good effect on extracting the features of organelle protein sequences. It also shows the superiority of enriching the model with biological prior knowledge and integrating protein structure knowledge into coding. At the same time, we believe that our method can be extended to other organelle protein localization and recognition, such as mitochondria and Golgi proteins. In the future, we will put it into practice and expand it on the basis of this work.

Availability of Data and Materials

The pre-trained ELMo-based SeqVec model and a description on how to implement the embeddings can be found here: https://github.com/Rostlab/SeqVec. The

ProSE model and a description on how to implement the embeddings can be found here: https://github.com/tbepler/prose. The ProSE-Pero model and datasets can be found here: https://github.com/SJNNNN/ProSE-Pero.

Author Contributions

JSui and JC conducted the experimental procedures, analyzed the raw data, and contributed to the manuscript editing process. YC and NI contributed their technical expertise to provide comprehensive support for the experiments. They played a critical role in evaluating the experimental procedures employed in the research study, offering constructive feedback, and suggesting potential modifications. And, JSun actively participated in the execution of the experiments and offered objective insights during the revision process of the manuscript, as well as contributed valuable suggestions for enhancing the experimental setup. All authors read and approved the final manuscript. All authors have participated sufficiently in the work and agreed to be accountable for all aspects of the work.

Ethics Approval and Consent to Participate

Not applicable.

Acknowledgment

Not applicable.

Funding

This work was supported in part by the Shandong Provincial Natural Science Foundation (ZR2021MF036) and the National Natural Science Foundation of China (31872415).

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Zheng P, Obara CJ, Szczesna E, Nixon-Abell J, Mahalingan KK, Roll-Mecak A, *et al*. ER proteins decipher the tubulin code to regulate organelle distribution. Nature. 2022; 601: 132–138.

[2] Schrader M, Godinho LF, Costello JL, Islinger M. The different facets of organelle interplay-an overview of organelle interactions. Frontiers in Cell and Developmental Biology. 2015; 3: 56.

[3] Zhou H, Chen C, Wang M, Ma Q, Yu B. Predicting golgi-resident protein types using conditional covariance minimization with XGBoost based on multiple features fusion. Ieee Access. 2019; 7: 144154144164.

[4] Lv Z, Jin S, Ding H, Zou Q. A Random Forest Sub-Golgi Protein Classifier Optimized via Dipeptide and Amino Acid Composition Features. Frontiers in Bioengineering and Biotechnology. 2019; 7: 215.

[5] Yu B, Qiu W, Chen C, Ma A, Jiang J, Zhou H, *et al*. SubMito-XGBoost: predicting protein submitochondrial localization by fusing multiple feature information and eXtreme gradient boosting. Bioinformatics (Oxford, England). 2020; 36: 1074–1081.

[6] Ahmad J, Hayat M. MFSC: Multi-voting based feature selection for classification of Golgi proteins by adopting the general form of Chou's PseAAC components. Journal of Theoretical Biology. 2019; 463: 99–109.

[7] Qiu W, Li S, Cui X, Yu Z, Wang M, Du J, *et al*. Predicting protein submitochondrial locations by incorporating the pseudo-position specific scoring matrix into the general Chou's pseudo-amino acid composition. Journal of Theoretical Biology. 2018; 450: 86–103.

[8] Savojardo C, Bruciaferri N, Tartari G, Martelli PL, Casadio R. DeepMito: accurate prediction of protein sub-mitochondrial localization using convolutional neural networks. Bioinformatics (Oxford, England). 2020; 36: 56–64.

[9] Wanders RJA. Metabolic functions of peroxisomes in health and disease. Biochimie. 2014; 98: 36–44.

[10] Cai M, Sun X, Wang W, Lian Z, Wu P, Han S, *et al*. Disruption of peroxisome function leads to metabolic stress, mTOR inhibition, and lethality in liver cancer cells. Cancer Letters. 2018; 421: 82–93.

[11] Benjamin DI, Cozzo A, Ji X, Roberts LS, Louie SM, Mulvihill MM, *et al*. Ether lipid generating enzyme AGPS alters the balance of structural and signaling lipids to fuel cancer pathogenicity. Proceedings of the National Academy of Sciences of the United States of America. 2013; 110: 14912–14917.

[12] Zhou M, Chinnaiyan AM, Kleer CG, Lucas PC, Rubin MA. Alpha-Methylacyl-CoA racemase: a novel tumor marker over-expressed in several human cancers and their precursor lesions. The American Journal of Surgical Pathology. 2002; 26: 926–931.

[13] Hartmann T, Bergsdorf C, Sandbrink R, Tienari PJ, Multhaup G, Ida N, *et al*. Alzheimer's disease betaA4 protein release and amyloid precursor protein sorting are regulated by alternative splicing. The Journal of Biological Chemistry. 1996; 271: 13208–13214.

[14] Berger J, Dorninger F, Forss-Petter S, Kunze M. Peroxisomes in brain development and function. Biochimica et Biophysica Acta. 2016; 1863: 934–955.

[15] Trompier D, Vejux A, Zarrouk A, Gondcaille C, Geillon F, Nury T, *et al*. Brain peroxisomes. Biochimie. 2014; 98: 102–110.

[16] Ding H, Liu L, Guo FB, Huang J, Lin H. Identify Golgi protein types with modified Mahalanobis discriminant algorithm and pseudo amino acid composition. Protein and Peptide Letters. 2011; 18: 58–63.

[17] Yang C, Mo YS, Chen HF, Huang YH, Li SL, Wang H, *et al*. The effects of Danggui-Shaoyao-San on neuronal degeneration and amyloidosis in mouse and its molecular mechanism for the treatment of Alzheimer's disease. Journal of Integrative Neuroscience. 2021; 20: 255–264.

[18] Anteghini M, Martins Dos Santos V, Saccenti E. In-Pero: Exploiting Deep Learning Embeddings of Protein Sequences to Predict the Localisation of Peroxisomal Proteins. International Journal of Molecular Sciences. 2021; 22: 6409.

[19] Alley EC, Khimulya G, Biswas S, AlQuraishi M, Church GM. Unified rational protein engineering with sequence-based deep representation learning. Nature Methods. 2019; 16: 1315–1322.

[20] Heinzinger M, Elnaggar A, Wang Y, Dallago C, Nechaev D, Matthes F, *et al*. Modeling aspects of the language of life through transfer-learning protein sequences. BMC Bioinformatics. 2019; 20: 723.

[21] Rao R, Bhattacharya N, Thomas N, Duan Y, Chen X, Canny J, *et al*. Evaluating Protein Transfer Learning with TAPE. Advances in Neural Information Processing Systems. 2019; 32: 9689–9701.

[22] Bepler T, Berger B. Learning the protein language: Evolution, structure, and function. Cell Systems. 2021; 12: 654–669.e3.

[23] St L, Wold S. Analysis of variance (ANOVA). Chemometrics and Intelligent Laboratory Systems. 1989; 6: 259–272.

[24] Morgat A, Lombardot T, Coudert E, Axelsen K, Neto TB, Gehant S, *et al*. Enzyme annotation in UniProtKB using Rhea. Bioinformatics (Oxford, England). 2020; 36: 1896–1901.
[25] Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics (Oxford, England). 2006; 22: 1658–1659.
[26] Yadav AK, Singla D. VacPred: Sequence-based prediction of plant vacuole proteins using machine-learning techniques. Journal of Biosciences. 2020; 45: 106.
[27] Lv Z, Cui F, Zou Q, Zhang L, Xu L. Anticancer peptides prediction with deep representation learning features. Briefings in Bioinformatics. 2021; 22: bbab008.
[28] Lv Z, Wang P, Zou Q, Jiang Q. Identification of sub-Golgi protein localization by use of deep representation learning features. Bioinformatics (Oxford, England). 2021; 36: 5600–5609.
[29] Fang Z, Feng T, Zhou H, Chen M. DeePVP: Identification and classification of phage virion proteins using deep learning. GigaScience. 2022; 11: giac076.
[30] Cui F, Zhang Z, Zou Q. Sequence representation approaches for sequence-based protein prediction tasks that use deep learning. Briefings in Functional Genomics. 2021; 20: 61–73.
[31] Long H, Sun Z, Li M, Fu HY, Lin MC. Predicting protein phosphorylation sites based on deep learning. Current Bioinformatics. 2020; 15: 300–308.
[32] Zhang Y, Yan J, Chen S, Gong M, Gao D, Zhu M, *et al*. Review of the applications of deep learning in bioinformatics. Current Bioinformatics. 2020; 15: 898–911.
[33] Iuchi H, Matsutani T, Yamada K, Iwano N, Sumi S, Hosoda S, *et al*. Representation learning applications in biological sequence analysis. Computational and Structural Biotechnology Journal. 2021; 19: 3198–3208.
[34] Angermueller C, Pärnamaa T, Parts L, Stegle O. Deep learning for computational biology. Molecular Systems Biology. 2016; 12: 878.
[35] Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. A primer on deep learning in genomics. Nature Genetics. 2019; 51: 12–18.
[36] Tang Y, Zhang YQ, Chawla NV, Krasser S. SVMs modeling for highly imbalanced classification. IEEE Transactions on Systems, Man, and Cybernetics. Part B, Cybernetics: a Publication of the IEEE Systems, Man, and Cybernetics Society. 2009; 39: 281–288.
[37] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, *et al*. Scikit-learn: Machine learning in Python. The Journal of Machine Learning Research. 2011; 12: 2825–2830.
[38] Zhang T, You F. Research on Short Text Classification Based on Textcnn. Journal of Physics: Conference Series. 2021; 1757: 012092.
[39] Busta M, Neumann L, Matas J. Fastext: Efficient Unconstrained Scene Text Detector. Proceedings of the IEEE international conference on computer vision. 2015. Available at: https://openaccess.thecvf.com/content_iccv_2015/html/Busta_FASText_Efficient_Unconstrained_ICCV_2015_paper.html (Accessed: 11 May 2023).
[40] Siami-Namini S, Tavakoli N, Namin AS. The Performance of Lstm and Bilstm in Forecasting Time Series. 2019. Available at: https://ieeexplore.ieee.org/abstract/document/9005997 (Accessed: 11 May 2023).
[41] Rhanoui M, Mikram M, Yousfi S, Barzali S. A Cnn-Bilstm Model for Document-Level Sentiment Analysis. Machine Learning and Knowledge Extraction. 2019; 1: 832–847.
[42] Zeng X, Lin W, Guo M, Zou Q. A comprehensive overview and evaluation of circular RNA detection tools. PLoS Computational Biology. 2017; 13: e1005420.
[43] Wei L, Xing P, Zeng J, Chen J, Su R, Guo F. Improved prediction of protein-protein interactions using novel negative samples, features, and an ensemble classifier. Artificial Intelligence in Medicine. 2017; 83: 67–74.
[44] Wei L, Xing P, Su R, Shi G, Ma ZS, Zou Q. CPPred-RF: A Sequence-based Predictor for Identifying Cell-Penetrating Peptides and Their Uptake Efficiency. Journal of Proteome Research. 2017; 16: 2044–2053.
[45] Hu Y, Zhao T, Zhang N, Zang T, Zhang J, Cheng L. Identifying diseases-related metabolites using random walk. BMC Bioinformatics. 2018; 19: 116.
[46] Zhang M, Li F, Marquez-Lago TT, Leier A, Fan C, Kwoh CK, *et al*. MULTiPly: a novel multi-layer predictor for discovering general and specific types of promoters. Bioinformatics (Oxford, England). 2019; 35: 2957–2965.
[47] Song T, Zeng X, Zheng P, Jiang M, Rodriguez-Paton A. A Parallel Workflow Pattern Modeling Using Spiking Neural P Systems with Colored Spikes. IEEE Transactions on Nanobioscience. 2018; 17: 474–484.
[48] Zhang C, Hicks GR, Raikhel NV. Molecular Composition of Plant Vacuoles: Important but Less Understood Regulations and Roles of Tonoplast Lipids. Plants (Basel, Switzerland). 2015; 4: 320–333.
[49] Kolb C, Nagel MK, Kalinowska K, Hagmann J, Ichikawa M, Anzenberger F, *et al*. FYVE1 is essential for vacuole biogenesis and intracellular trafficking in Arabidopsis. Plant Physiology. 2015; 167: 1361–1373.
[50] Cui Y, Zhao Q, Hu S, Jiang L. Vacuole Biogenesis in Plants: How Many Vacuoles, How Many Models? Trends in Plant Science. 2020; 25: 538–548.
[51] Kataoka T, Watanabe-Takahashi A, Hayashi N, Ohnishi M, Mimura T, Buchner P, *et al*. Vacuolar sulfate transporters are essential determinants controlling internal distribution of sulfate in Arabidopsis. The Plant Cell. 2004; 16: 2693–2704.
[52] Martinoia E, Meyer S, De Angeli A, Nagy R. Vacuolar transporters in their physiological context. Annual Review of Plant Biology. 2012; 63: 183–213.
[53] Martinoia E, Maeshima M, Neuhaus HE. Vacuolar transporters and their essential role in plant metabolism. Journal of Experimental Botany. 2007; 58: 83–102.
[54] Jiao S, Zou Q. Identification of plant vacuole proteins by exploiting deep representation learning features. Computational and Structural Biotechnology Journal. 2022; 20: 2921–2927.
[55] Hoyer S. Is sporadic Alzheimer disease the brain type of non-insulin dependent diabetes mellitus? A challenging hypothesis. Journal of Neural Transmission (Vienna, Austria: 1996). 1998; 105: 415–422.
[56] Rose DR. Structure, mechanism and inhibition of Golgi α-mannosidase II. Current Opinion in Structural Biology. 2012; 22: 558–562.
[57] Su LJ, Auluck PK, Outeiro TF, Yeger-Lotem E, Kritzer JA, Tardiff DF, *et al*. Compounds from an unbiased chemical screen reverse both ER-to-Golgi trafficking defects and mitochondrial dysfunction in Parkinson's disease models. Disease Models & Mechanisms. 2010; 3: 194–208.
[58] Arendt T, Zvegintseva HG, Leontovich TA. Dendritic changes in the basal nucleus of Meynert and in the diagonal band nucleus in Alzheimer's disease–a quantitative Golgi investigation. Neuroscience. 1986; 19: 1265–1278.
[59] Majrashi M, Altukri M, Ramesh S, Govindarajulu M, Schwartz J, Almaghrabi M, *et al*. β-hydroxybutyric acid attenuates oxidative stress and improves markers of mitochondrial function in the HT-22 hippocampal cell line. Journal of Integrative Neuroscience. 2021; 20: 321–329.
[60] Burbulla LF, Song P, Mazzulli JR, Zampese E, Wong YC, Jeon S, *et al*. Dopamine oxidation mediates mitochondrial and lysosomal dysfunction in Parkinson's disease. Science (New York, N.Y.). 2017; 357: 1255–1261.

[61] Shi SP, Qiu JD, Sun XY, Huang JH, Huang SY, Suo SB, *et al*. Identify submitochondria and subchloroplast locations with pseudo amino acid composition: approach from the strategy of discrete wavelet transform feature extraction. Biochimica et Biophysica Acta. 2011; 1813: 424–430.
[62] Gerbitz KD, Gempel K, Brdiczka D. Mitochondria and diabetes. Genetic, biochemical, and clinical implications of the cellular energy circuit. Diabetes. 1996; 45: 113–126.
[63] Poveda-Huertes D, Mulica P, Vögtle FN. The versatility of the mitochondrial presequence processing machinery: cleavage, quality control and turnover. Cell and Tissue Research. 2017; 367: 73–81.