Mathematical Biosciences and Engineering

MBE, 21(1): 1472–1488.
DOI: 10.3934/mbe.2024063
Received: 22 September 2023
Revised: 24 November 2023
Accepted: 04 December 2023
Published: 28 December 2023

http://www.aimspress.com/journal/MBE

Research article

NCSP-PLM: An ensemble learning framework for predicting non-classical secreted proteins based on protein language models and deep learning

Taigang Liu, Chen Song and Chunhua Wang*

College of Information Technology, Shanghai Ocean University, Shanghai 201306, China

* **Correspondence:** Email: chhwang@shou.edu.cn.

Abstract: Non-classical secreted proteins (NCSPs) refer to a group of proteins that are located in the extracellular environment despite the absence of signal peptides and motifs. They usually play different roles in intercellular communication. Therefore, the accurate prediction of NCSPs is a critical step to understanding in depth their associated secretion mechanisms. Since the experimental recognition of NCSPs is often costly and time-consuming, computational methods are desired. In this study, we proposed an ensemble learning framework, termed NCSP-PLM, for the identification of NCSPs by extracting feature embeddings from pre-trained protein language models (PLMs) as input to several fine-tuned deep learning models. First, we compared the performance of nine PLM embeddings by training three neural networks: Multi-layer perceptron (MLP), attention mechanism and bidirectional long short-term memory network (BiLSTM) and selected the best network model for each PLM embedding. Then, four models were excluded due to their below-average accuracies, and the remaining five models were integrated to perform the prediction of NCSPs based on the weighted voting. Finally, the 5-fold cross validation and the independent test were conducted to evaluate the performance of NCSP-PLM on the benchmark datasets. Based on the same independent dataset, the sensitivity and specificity of NCSP-PLM were 91.18% and 97.06%, respectively. Particularly, the overall accuracy of our model achieved 94.12%, which was 7~16% higher than that of the existing state-of-the-art predictors. It indicated that NCSP-PLM could serve as a useful tool for the annotation of NCSPs.

Keywords: protein language model; non-classical secreted protein; deep learning; imbalanced classification; ensemble learning

1. Introduction

As a fundamental mechanism for intercellular communication, protein secretion could occur in all living organisms and has an important role in many physiological processes. The majority of secretory proteins contain an N-terminal signal peptide that allows their translocation into the endoplasmic reticulum via the classical secretory system [1]. Nevertheless, several cytoplasmic proteins detected in the extracellular environment lacking a known signal peptide are secreted via the non-classical protein secretion pathway [2]. They are usually described as NCSPs and can play diverse roles in various biological processes including intercellular signaling, immune regulation, tissue repair and regeneration, cell communication and human diseases such as neurodegenerative disorders and cancer [3–5].

Accurate identification of NCSPs is important for unraveling the complexity of intercellular communication and the underlying mechanisms involved in the above physiological and pathological processes. Since experimental approaches are often costly and time-consuming, computational methods will be required to enable genome-wide annotation of NCSPs with high efficiency and low cost [6]. To date, various methods based on machine learning have been developed for predicting NCSPs, including SecretomeP [7], SecretP [8], NClassG+ [9], PeNGaRoo [10], NonClasGP-Pred [11], ASPIRER [12], iNSP-GCAAP [13] and so on. For instance, Bendtsen et al. proposed the first tool termed SecretomeP for predicting NCSPs in mammals by employing six sequence-based features as the input of the neural network [7]. The SecretP model trained a support vector machine (SVM) to distinguish the three types of secretory proteins by using both sequence and structural features [8]. The NClassG+ tool was designed for identifying NCSPs in Gram-positive bacteria, which adopted the nested k-fold cross-validation (CV) to select the best models from four different sequence transformation vectors and SVMs with linear, polynomial and Gaussian kernel functions [9]. Recently, Zhang et al. developed a two-layer LightGBM ensemble learning framework, termed PeNGaRoo, for predicting NCSPs in Gram-positive bacteria by extracting three groups of features, i.e., sequence-derived features, evolutionary information-based features and physicochemical property-based features [10]. Moreover, the NonClasGP-Pred model improved the performance of NCSPs prediction based on the same datasets with PeNGaRoo by handling the potential prediction bias arising from imbalanced data [11]. Additionally, ASPIRER trained a hybrid deep learning-based framework to enhance the identification of NCSPs by combining a whole amino acid sequence-based model and an N-terminal sequence-based model [12]. iNSP-GCAAP utilized the global composition of amino acid properties to encode protein sequences and then adopted the random forest algorithm to perform the prediction of NCSPs, which achieved the superior performance than the other state-of-the-art methods [13].

Most of existing techniques often depend on handcrafted features as the input of machine learning algorithms [14,15], such as the position-specific scoring matrix (PSSM) derived from time-consuming database searches [16]. In contrast, pre-trained PLMs could automatically learn efficient representations (also known as PLM embeddings) from the protein sequences in a self-supervised manner by treating the protein sequences as sentences in the field of natural language processing. These pre-trained models include ProtVec [17], SeqVec [18], ProSE [19], UniRep [20], Tape [21], ESM-1b [22], ProtBERT [23], ProtT5 [23], ProteinBERT [24] and so on. Recent studies have shown that PLM embeddings could be successfully applied for different protein-related downstream tasks, such as protein subcellular localization [25], peptide recognition [26,27], protein fold prediction [28], recognition of post-translational modification sites [29,30] and so on [31,32].

To the best of our knowledge, the PLM technique has not been systematically tested on the prediction of NCSPs. In this study, we proposed a novel computational approach, termed NCSP-PLM, to identify the NCSPs based on their protein sequences by selecting the optimal model from nine different PLM embeddings and three deep learning models. For each of these nine embeddings, we first trained three neural networks, i.e., MLP, attention mechanism and bidirectional long short-term memory network (BiLSTM), and then picked out the best one. Second, we selected the top five models with the accuracies higher than the average accuracy of these nine models. Third, the ensemble classifier was adopted to perform the final prediction of NCSPs using the weighted voting of these five optimal models. Benchmark experiments on the 5-fold CV and the independent test suggested that the proposed NCSP-PLM model outperformed existing tools based on the traditional handcrafted features and the PLM embeddings are particularly useful for the NCSPs prediction. Figure 1 illustrates the flow chart of the NCSP-PLM model.

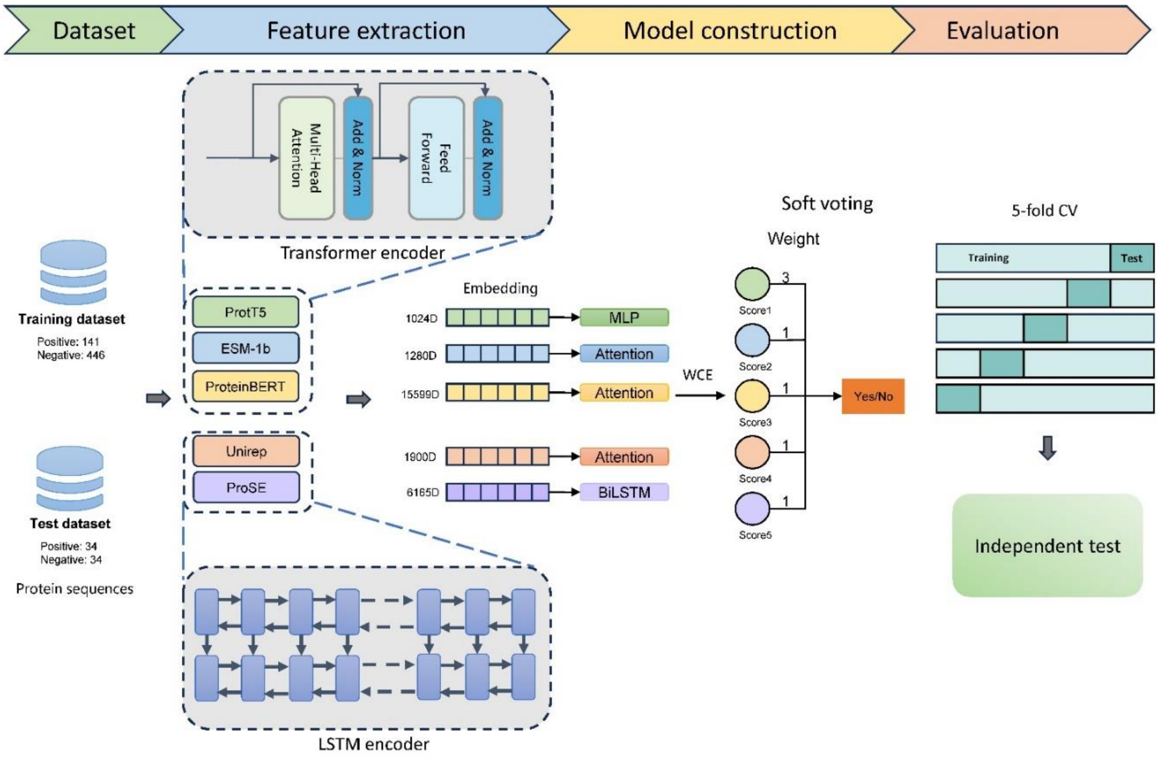

Figure 1. The flow chart of the NCSP-PLM model.

2. Materials and methods

2.1. Benchmark datasets

The critical first step in developing a robust and efficient classification model is the construction of a high-quality benchmark dataset. In this study, we used the benchmark datasets constructed by Zhang et al. [10] to train and evaluate the proposed model. The training dataset includes 141 positive samples (i.e., NCSPs) and 446 negative samples (i.e., cytoplasmic proteins), which was applied to perform the 5-fold CV. In addition, the independent test dataset consists of 34 positive and 34 negative samples, which was employed to compare our model with the other existing tools.

The reasons why we adopted these datasets were chiefly as follows. (1) All NCSPs were experimentally verified in literature. (2) The sequence similarity was reduced to 80% to avoid the homology bias. (3) There were no overlapping sequences between the training dataset and the independent test dataset.

2.2. Pre-trained protein language model embeddings

As protein representations, we directly extracted self-supervised embeddings from pre-trained PLMs without fine-tuning the training data. In this present work, nine popular PLMs were adopted, including ProtVec [17], SeqVec [18], ProSE [19], UniRep [20], Tape [21], ESM-1b [22], ProtBERT [23], ProtT5 [23] and ProteinBERT [24]. Given a protein with the length of L, the size of a PLM embedding is $L \times F$, where F denotes the dimension of the individual embedding for each amino acid. To obtain a fixed-length vector representation, we averaged the embedding matrix over the length L.

Table 1. The summary of the nine PLM embeddings adopted in this study.

Name	Dimension	Database
ProtVec	100	Swiss-Prot [33]
SeqVec	1024	UniRef50 [34]
ProSE	6165	UniRef50 + SCOP [35]
UniRep	1900	UniRef50
Tape	768	Pfam [36]
ESM-1b	1280	UniRef50
ProtBERT	1024	BFD [37] + UniRef100 [34]
ProtT5	1024	BFD + UniRef50
ProteinBERT	15599	UniRef90 [34]

Table 1 summarizes the nine PLM embeddings used in this study. (1) The ProtVec embedding is the first word vector-based protein representation, which was trained on the Swiss-Prot database [33] through a Skip-gram neural network and generated a 100-dimensional vector [17]. (2) The SeqVec was trained on the UniRef50 database [34] by using an architecture composed of a convolutional layer and two BiLSTM layers [18]. (3) The structure of ProSE is a three-layer BiLSTM similar to the SeqVec structure, with the difference that it uses not only the sequence data but also the structural information of the proteins [19]. (4) The UniRep model contains a layer of multiplicative LSTM with 1900 hidden units, which was trained on the UniRef50 database [20]. (5) The Tape model aims to leverage the power of transformers to capture long-range dependencies and context in protein sequences [21], trained on the Pfam database [36]. (6) The ESM-1b model has 33 transformer layers and was trained on the UniRef50 database by using the masked language modeling objective [22]. (7) The ProtBERT and ProtT5 models are based on two auto-encoder transformer structures, trained on data from the BFD [37] and UniRef databases. The difference between the two models is that ProtBERT trained only the encoder component, while ProtT5 consists of both an encoder and a decoder. (8) Unlike the classic transformers, ProteinBERT is a denoising auto-encoder model and contains both local and global representations [24]. The details of these nine PLM embeddings were also provided in Supplementary

Table S1.

2.3. Deep learning model architecture

In this study, we adopted three different deep learning architectures, i.e., MLP, attention mechanism and BiLSTM, to process the PLM embeddings and perform the prediction of NCSPs. Figure 2 shows the overall network structures of these three models. We implemented our models by using TensorFlow (1.15.5) and the specific parameters of these deep networks are available in Supplementary Table S2.

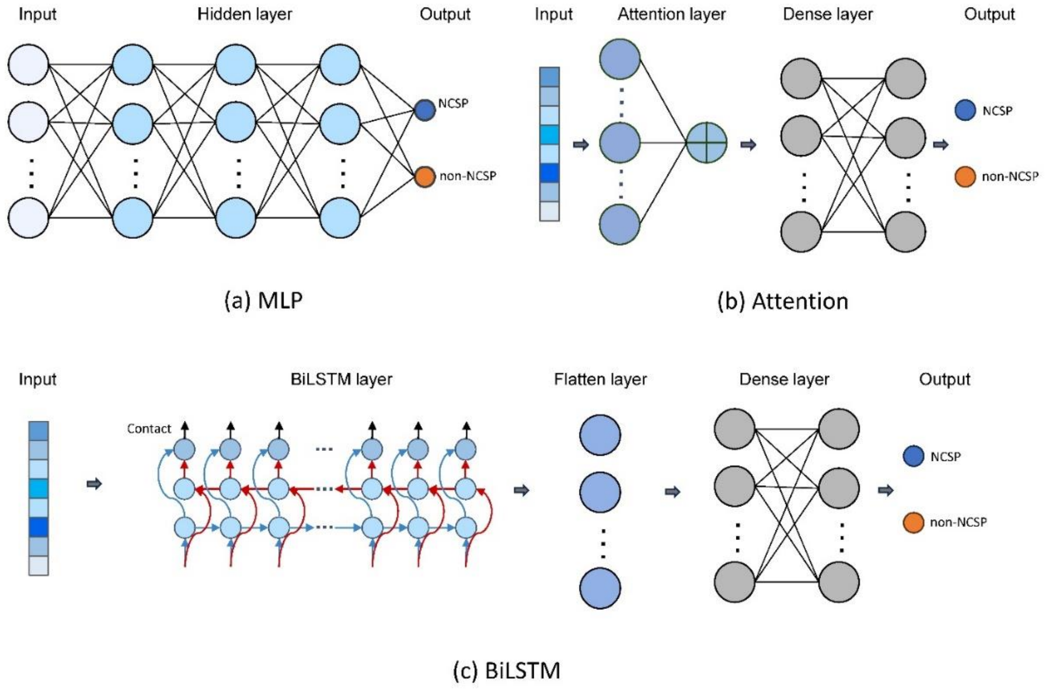

Figure 2. The network structures of three deep learning models. (a) The MLP model processes PLM embeddings through three dense layers. (b) The Attention model adds the attention mechanism before two dense layers. (c) The BiLSTM model uses a flatten layer after the output of BiLSTM, followed by two dense layers.

As shown in Figure 2(a), the MLP was employed as our baseline model, which consists of an input layer, three hidden layers and an output layer. Additionally, we applied the batch normalization (BatchNorm) to mitigate the overfitting after the hidden layers. In Figure 2(b), an attention layer before the MLP structure was introduced to amplify the influence of key input features. As the output of the attention layer, a weighted feature vector quantifying the importance of the embeddings was obtained and then passed to a dense layer consisting of 512 units. As illustrated in Figure 2(c), we designed a BiLSTM layer with 512 cells before the MLP to process the input PLM embeddings in both forward and backward directions simultaneously. The output of the BiLSTM layer was fed to a flatten layer, followed by two dense layers with 512 cells. With the aim of reducing the overfitting, we also applied the BatchNorm to both the flatten and dense layers.

2.4. Imbalanced classification problem solving

The imbalanced proportion of positive and negative samples could affect the prediction accuracy of the classifier. In this study, we explored three approaches to address this issue, i.e., synthetic minority oversampling technique (SMOTE) [38], focal loss [39] and weighted binary cross-entropy (WCE) [40].

SMOTE is an oversampling technique that allows us to create synthetic samples for our minority class on the lines connecting a sample point and one of its K-nearest neighbors [38]. Focal loss is an improved version of cross-entropy loss that specifically handles the imbalanced classification problem by assigning higher weights to hard or frequently misclassified instances, while down-weighting the easy instances [39]. WCE is also a variant of the binary cross-entropy loss function that assigns different weights to the positive and negative classes to balance their contributions to the loss function [40]. The weights are usually inversely proportional to the class frequencies, meaning that the weight of the minority class is higher than the weight of the majority class.

2.5. Performance assessment

In this study, the 5-fold CV and the independent test were performed to examine the performance of our models for the prediction of NCSPs. In addition, six common metrics were adopted to report the predictive ability of the proposed model [41,42], including sensitivity (SN), specificity (SP), precision (P), accuracy (ACC), F1 and Matthews correlation coefficient (MCC), defined with the following equations:

$$\mathrm{SN} = \frac{TP}{TP+FN}, \tag{1}$$

$$\mathrm{SP} = \frac{TN}{TN+FP}, \tag{2}$$

$$\mathrm{P} = \frac{TP}{TP+FP}, \tag{3}$$

$$\mathrm{ACC} = \frac{TP+TN}{TP+FP+TN+FN}, \tag{4}$$

$$\mathrm{F1} = 2 \times \frac{TP}{2TP+FP+FN}, \tag{5}$$

$$\mathrm{MCC} = \frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FN)\times(TN+FP)\times(TP+FP)\times(TN+FN)}}, \tag{6}$$

where *TP*, *FP*, *TN* and *FN* represent the numbers of the true positive, false positive, true negative and false negative samples, respectively.

Additionally, the area under the receiver operating characteristic (ROC) curve (AUC) and the area under the precision-recall (PR) curve (AUPRC) were calculated as another two reliable performance metrics for the comparison with existing algorithms.

3. Results and discussion

3.1. Performance of protein language model embedding with different deep learning models

In this section, we employed three deep learning models, i.e., MLP, attention mechanism and

BiLSTM, to compare the performance of nine PLM embeddings for the prediction of NCSPs. For each embedding, three neural networks were trained on the benchmark dataset, resulting in 27 base models. The results of the independent tests were shown in Figure 3 and those of the 5-fold CV were illustrated in Supplementary Figure S1.

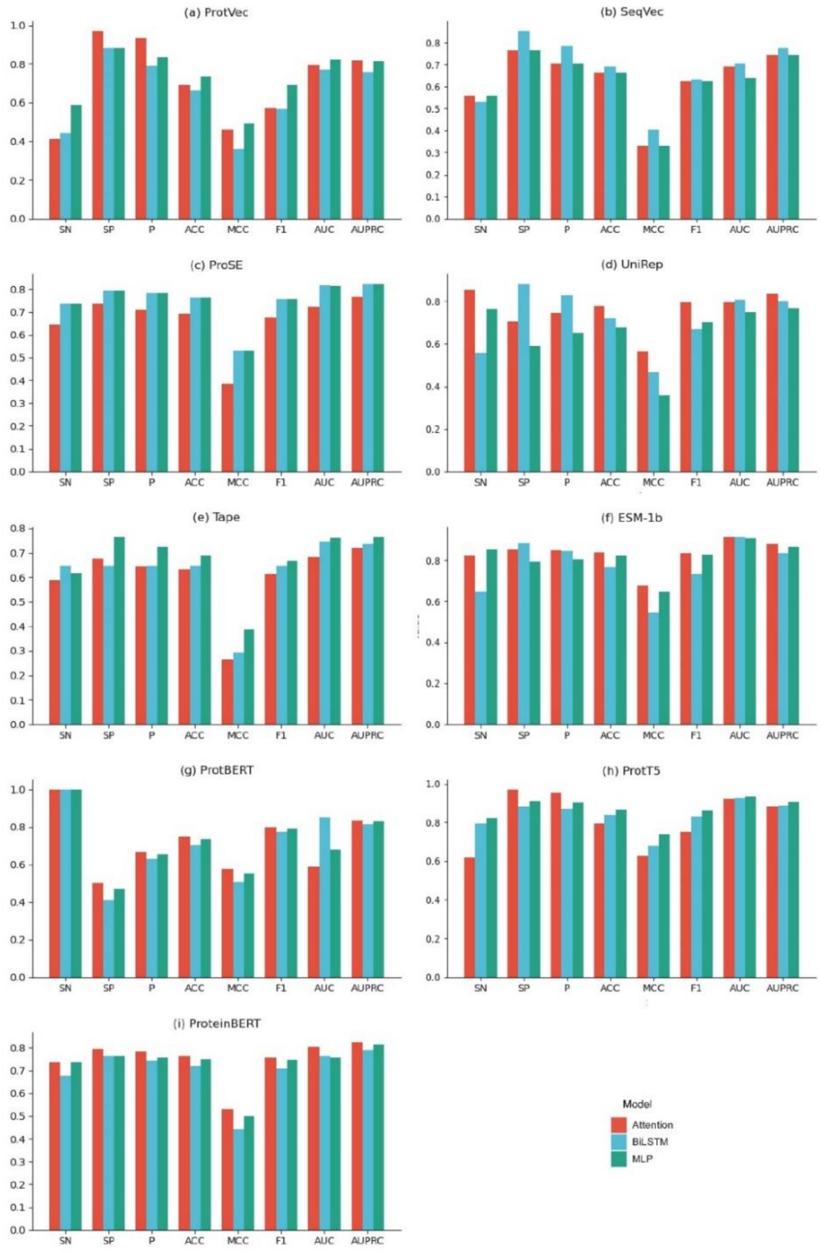

Figure 3. Performance comparison of nine PLM embeddings, i.e., (a) ProtVec, (b) SeqVec, (c) ProSE, (d) UniRep, (e) Tape, (f) ESM-1b, (g) ProtBERT, (h) ProtT5 and (i) ProteinBERT.

As seen from Figure 3, different embeddings achieved the best ACC values using different deep learning models. Specifically, the UniRep, ESM-1b, ProtBERT and ProteinBERT embeddings exhibited the outstanding ability of identifying the NCSPs by utilizing the attention mechanism model, with the ACC values of 0.7794, 0.8382, 0.7500 and 0.7647, respectively. The MLP models trained by the ProtVec, Tape and ProtT5 embeddings, respectively, outperformed the attention mechanism and BiLSTM models, with the AUC values of 0.8244, 0.7638 and 0.9325. The BiLSTM models obtained the highest ACC values (i.e., 0.6912 and 0.7647) when using the SeqVec and ProSE embeddings. Moreover, ProtT5 performed better than the other embeddings in terms of ACC, MCC, F1, AUC and AUPRC.

To further select the optimal models, the best ACC values for these nine models selected from 27 base models were plotted in Figure 4. The average ACC of nine models was 0.765. Four embeddings were discarded in the subsequent analysis due to their below-average ACC values. In other words, ProSE+BiLSTM, ProtT5+MLP, UniRep+Attention, ESM-1b+Attention and ProteinBERT+Attention were selected to build the ensemble classifier for the identification of the NCSPs.

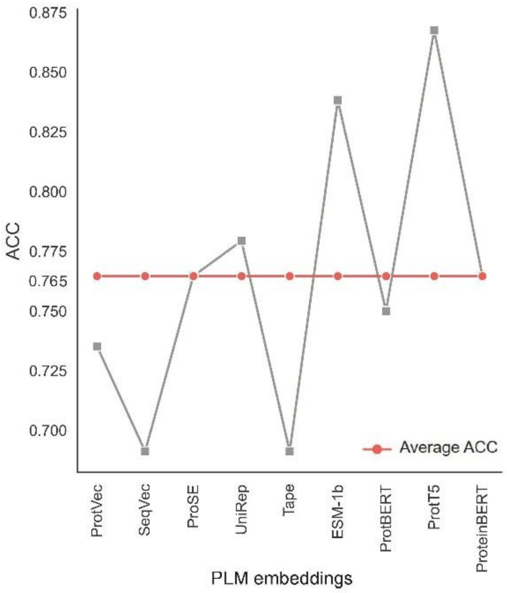

Figure 4. The line chart shows the ACC values for nine PLM embeddings.

3.2. Performance of ensemble approaches

In this section, the independent test was performed to assess the performance of the ensemble models, which adopted the soft voting strategy to integrate the output of the 5 optimal base models by assigning different weights. For the sake of simplicity, the weights of ProSE+BiLSTM, UniRep+Attention, ESM-1b+Attention and ProteinBERT+Attention were equally set to 1 due to their comparable levels. Moreover, ProtT5+MLP was assigned higher weights to strengthen its influence in the final results because of its remarkable performance. The five metrics, including SN, SP, ACC, MCC and F1, were adopted to evaluate the performance of these models, and the corresponding results were listed in Table 2.

As can be seen from Table 2, all ensemble models achieved the better and more stable performance compared with the corresponding individual models, indicating the effectiveness of the soft voting strategy. Besides, the ensemble model obtained the highest ACC, MCC and F1 values when ProtT5+MLP had a weight of 3. However, the ACC value witnessed a downward trend when increasing the weight of ProtT5+MLP higher than 3, indicating that the excessively high weight setting may lead to the overreliance on a single model and thus harm the overall performance.

Table 2. Performance of the soft voting by using different weights.

Weight	SN	SP	ACC	MCC	F1
1:1:1:1:1	0.8824	0.9412	0.9118	0.8250	0.9091
2:1:1:1:1	0.9118	0.9412	0.9265	0.8533	0.9254
3:1:1:1:1	**0.9118**	**0.9706**	**0.9412**	**0.8839**	**0.9394**
4:1:1:1:1	0.8824	0.9706	0.9265	0.8563	0.9231
5:1:1:1:1	0.9118	0.8824	0.8971	0.7945	0.8986

3.3. Effect of different strategies for handling sample imbalance

In this section, we investigated the effect of three different strategies for solving the data imbalance problem, including SMOTE, focal loss and WCE. Table 3 summarized the comparison results on the independent test dataset. The corresponding ROC and PR curves were shown in Figure 5.

Table 3. Effect of three balancing strategies.

Strategy	SN	SP	ACC	MCC	F1
No balancing	0.8824	0.9412	0.9118	0.8250	0.9143
SMOTE	0.7941	0.9118	0.8529	0.7108	0.8438
Focal loss	0.8824	**0.9796**	0.9265	0.8563	0.9231
WCE	**0.9118**	0.9706	**0.9412**	**0.8839**	**0.9394**

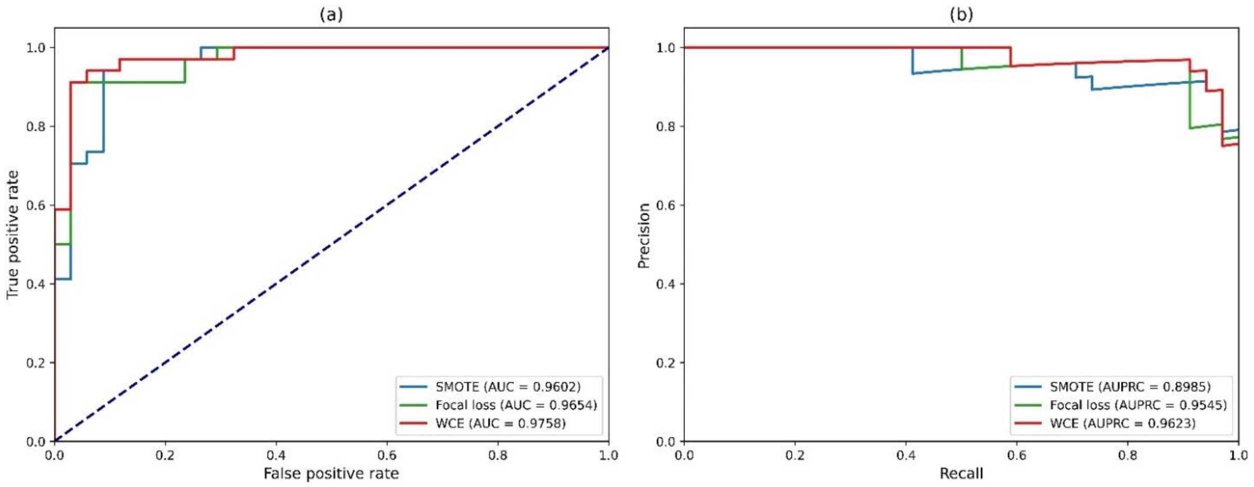

Figure 5. The ROC and PR curves based on three different balancing strategies. (a) ROC curves; and (b) PR curves.

Referring to Table 3, the SN values were always lower than the SP values in any case, caused by the low proportion of the positive samples in the training dataset compared to the negative samples. In addition, the SMOTE technique unexpectedly performed poorly in terms of ACC and MCC, suggesting that synthetic examples generated by the SMOTE did not retain the specific characteristics of the NCSPs. In contrast, the focal loss and WCE techniques, which were based on the cross-entropy loss function, markedly improved the model’s performance. The WCE method was superior to the focal loss method in terms of all evaluation metrics except for SP. Hence, we adopted the WCE method as the final scheme to handle the imbalanced classification in this study.

3.4. Comparison with existing methods

To the best of our knowledge, there are only four computational tools for the identification of the NCSPs on the same training dataset and the independent test dataset, including PeNGaRoo [10], NonClasGP-Pred [11], ASPIRER [12] and iNSP-GCAAP [13]. As mentioned above, these models relied on a variety of handcrafted features to train different supervised learning algorithms for predicting the NCSPs. Table 4 presented a comparison of our NCSP-PLM model with these methods using eight evaluation indices. The ROC and PR curves of NCSP-PLM were illustrated in Figure 6.

Table 4. Performance comparison with existing methods using the independent test.

Method	SN	SP	P	ACC	MCC	F1	AUC	AUPRC
PeNGaRoo	0.8235	0.7353	0.7568	0.7794	0.5610	0.7887	0.8521	0.9042
NonClasGP-Pred	0.8676	0.8529	0.8571	0.8676	0.7356	0.8696	0.9019	0.9177
ASPIRER	0.6471	0.9701	0.9565	0.8088	0.6528	0.7719	0.9533	0.9444
iNSP-GCAAP	0.6176	**0.9706**	-	0.7941	0.6287	-	0.9256	-
NCSP-PLM	**0.9118**	**0.9706**	**0.9688**	**0.9412**	**0.8839**	**0.9394**	**0.9758**	**0.9623**

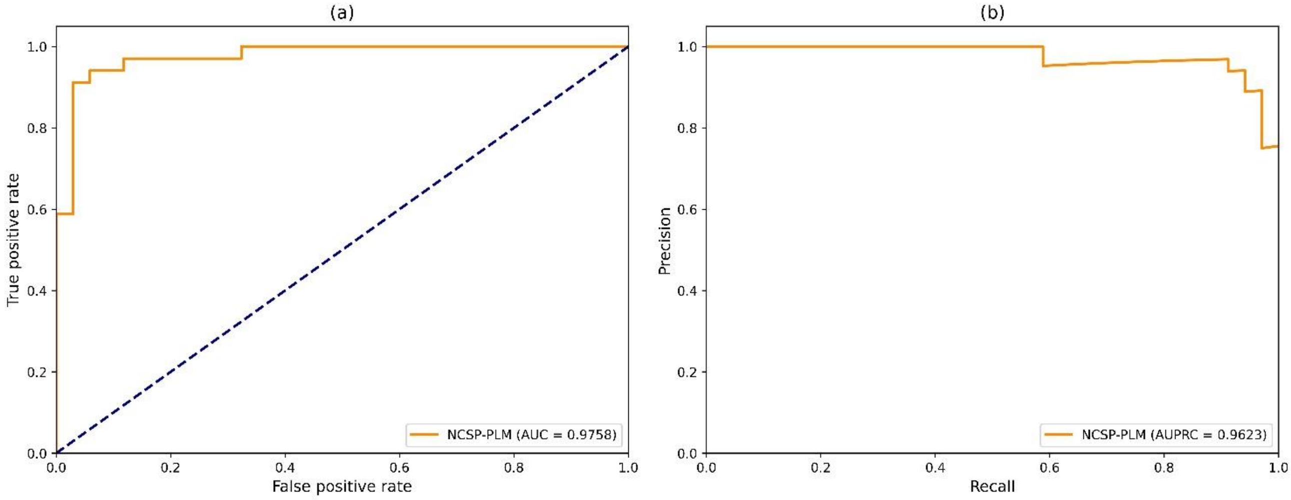

Figure 6. The ROC and PR curves of NCSP-PLM based on the independent test. (a) ROC curves; and (b) PR curves.

As shown in Table 4, the proposed NCSP-PLM predictor outperformed the listed state-of-the-art methods in terms of SN (0.9118), SP (0.9706), P (0.9688), ACC (0.9412), MCC (0.8839), F1 (0.9394), AUC (0.9758) and AUPRC (0.9623). This indicated that the performance of traditional protein representations can be reached or surpassed by the PLM embeddings for the NCSPs prediction task. Additionally, the NonClasGP-Pred tool achieved the balanced SN and SP values, which addressed the data imbalance issue by generating ten balanced datasets. Moreover, the ASPIRER and iNSP-GCAAP models yielded the comparable SP values higher than 0.97. However, the SN values of these two methods were lower than 0.65, probably caused by the data imbalance.

4. Conclusions

In this study, we presented a novel approach called NCSP-PLM for predicting the NCSPs in Gram-positive bacteria. First, we provided a comparative analysis of nine different PLM embeddings with three deep learning models, and picked out the five optimal base models. Then, we constructed the ensemble learning framework using the weighted soft voting scheme to improve the performance of the proposed model and adopted the WCE technique to handle the data imbalance issue. Finally, benchmark experiments demonstrated that NCSP-PLM performed remarkably well in the NCSPs identification task and obtained a significant performance boost over current state-of-the-art methods based on traditional protein feature representations. The source code and all the datasets are freely available at https://github.com/hollymmm/NCSP-PLM.

There are two aspects that highlight the novelty of our model: (1) The knowledge derived from the pre-trained PLMs was extracted as feature embeddings and adopted to predict the NCSPs for the first time; and (2) the comparison of nine PLMs was made to develop the most of their potential for the annotation of NCSPs. In our future endeavors, we aspire to continually improve our model through three major avenues. First, to mitigate the risk of overfitting, we will gather additional NCSP samples from published work and build a larger dataset for training our model. Second, we will explore the combined use of multi-view features to enhance the prediction of NCSPs such as sequence-derived features, PSSM-based features, physicochemical property-based features and PLMs-based features. Third, we will provide a user-friendly web server accessible to the public, offering more than just the source code of the model.

Use of AI tools declaration

The authors declare that they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (grant number 11601324).

Conflict of interest

The authors declare that there are no conflicts of interest.

References

1. M. Zhang, L. Liu, X. Lin, Y. Wang, Y. Li, Q. Guo, et al., A translocation pathway for vesicle-mediated unconventional protein secretion, *Cell*, **181** (2020), 637–652. https://doi.org/10.1016/j.cell.2020.03.031

2. Q. Kang, D. Zhang, Principle and potential applications of the non-classical protein secretory pathway in bacteria, *Appl. Microbiol. Biotechnol.*, **104** (2020), 953–965. https://doi.org/10.1007/s00253-019-10285-4
3. M. Jacopo, Unconventional protein secretion (UPS): Role in important diseases, *Mol. Biomed.*, **4** (2023), 2. https://doi.org/10.1186/s43556-022-00113-z
4. P. Broz, Unconventional protein secretion by gasdermin pores, *Semin. Immunol.*, **69** (2023), 101811. https://doi.org/10.1016/j.smim.2023.101811
5. G. Poschmann, J. Bahr, J. Schrader, I. Stejerean-Todoran, I. Bogeski, K. Stuehler, Secretomics-a key to a comprehensive picture of unconventional protein secretion, *Front. Cell. Dev. Biol.*, **10** (2022), 828027. https://doi.org/10.3389/fcell.2022.878027
6. W. Dai, J. Li, Q. Li, J. Cai, J. Su, C. Stubenrauch, et al., PncsHub: A platform for annotating and analyzing non-classically secreted proteins in Gram-positive bacteria, *Nucleic Acids Res.*, **50** (2022), D848–D857. https://doi.org/10.1093/nar/gkab814
7. J. D. Bendtsen, L. J. Jensen, N. Blom, G. von Heijne, S. Brunak, Feature-based prediction of non-classical and leaderless protein secretion, *Protein Eng. Des. Sel.*, **17** (2004), 349–356. https://doi.org/10.1093/protein/gzh037
8. L. Yu, Y. Guo, Z. Zhang, Y. Li, M. Li, G. Li, et al., SecretP: A new method for predicting mammalian secreted proteins, *Peptides*, **31** (2010), 574–578. https://doi.org/10.1016/j.peptides.2009.12.026
9. D. Restrepo-Montoya, C. Pino, L. F. Nino, M. E. Patarroyo, M. A. Patarroyo, NClassG+: A classifier for non-classically secreted Gram-positive bacterial proteins, *BMC Bioinf.*, **12** (2011), 21. https://doi.org/10.1186/1471-2105-12-21
10. Y. Zhang, S. Yu, R. Xie, J. Li, A. Leier, T.T. Marquez-Lago, et al., PeNGaRoo, a combined gradient boosting and ensemble learning framework for predicting non-classical secreted proteins, *Bioinf.*, **36** (2020), 704–712. https://doi.org/10.1093/bioinformatics/btz629
11. C. Wang, J. Wu, L. Xu, Q. Zou, NonClasGP-Pred: Robust and efficient prediction of non-classically secreted proteins by integrating subset-specific optimal models of imbalanced data, *Microb. Genom.*, **6** (2020), mgen000483. https://doi.org/10.1099/mgen.0.000483
12. X. Wang, F. Li, J. Xu, J. Rong, G. I. Webb, Z. Ge, et al., ASPIRER: A new computational approach for identifying non-classical secreted proteins based on deep learning, *Brief. Bioinf.*, **23** (2022), bbac031. https://doi.org/10.1093/bib/bbac031
13. T. T. Do, T. H. Nguyen-Vo, H. T. Pham, Q. H. Trinh, B. P. Nguyen, iNSP-GCAAP: Identifying nonclassical secreted proteins using global composition of amino acid properties, *Proteomics*, **23** (2023), e2100134. https://doi.org/10.1002/pmic.202100134
14. H. Zulfiqar, Z. Guo, B. K. Grace-Mercure, Z. Y. Zhang, H. Gao, H. Lin, et al., Empirical comparison and recent advances of computational prediction of hormone binding proteins using machine learning methods, *Comput. Struct. Biotechnol. J.*, **21** (2023), 2253–2261. https://doi.org/10.1016/j.csbj.2023.03.024
15. F. Y. Dao, M. L. Liu, W. Su, H. Lv, Z. Y. Zhang, H. Lin, et al., AcrPred: A hybrid optimization with enumerated machine learning algorithm to predict anti-CRISPR proteins, *Int. J. Biol. Macromol.*, **228** (2023), 706–714. https://doi.org/10.1016/j.ijbiomac.2022.12.250
16. S. F. Altschul, T. L. Madden, A. A. Schaffer, J. H. Zhang, Z. Zhang, W. Miller, et al., Gapped BLAST and PSI-BLAST: A new generation of protein database search programs, *Nucleic Acids Res.*, **25** (1997), 3389–3402. https://doi.org/10.1093/nar/25.17.3389

17. E. Asgari, M. R. K. Mofrad, Continuous distributed representation of biological sequences for deep proteomics and genomics, *PloS One*, **10** (2015), e0141287. https://doi.org/10.1371/journal.pone.0141287

18. M. Heinzinger, A. Elnaggar, Y. Wang, C. Dallago, D. Nechaev, F. Matthes, et al., Modeling aspects of the language of life through transfer-learning protein sequences, *BMC Bioinf.*, **20** (2019), 723. https://doi.org/10.1186/s12859-019-3220-8

19. T. Bepler, B. Berger, Learning the protein language: Evolution, structure, and function, *Cell Syst.*, **12** (2021), 654–669. https://doi.org/10.1016/j.cels.2021.05.017

20. E. C. Alley, G. Khimulya, S. Biswas, M. AlQuraishi, G. M. Church, Unified rational protein engineering with sequence-based deep representation learning, *Nat. Methods*, **16** (2019), 1315–1322. https://doi.org/10.1038/s41592-019-0598-1

21. R. Rao, N. Bhattacharya, N. Thomas, Y. Duan, X. Chen, J. Canny, et al., Evaluating protein transfer learning with TAPE, in *33rd Conference on Neural Information Processing Systems (NeurIPS),* **32** (2019), 9689–9701.

22. A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, et al., Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences, *Proc. Natl. Acad. Sci. U. S. A.*, **118** (2021), e2016239118. https://doi.org/10.1073/pnas.2016239118

23. A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, et al., ProtTrans: Toward understanding the language of life through self-supervised learning, *IEEE Trans. Pattern Anal. Mach. Intell.*, **44** (2022), 7112–7127. https://doi.org/10.1109/tpami.2021.3095381

24. N. Brandes, D. Ofer, Y. Peleg, N. Rappoport, M. Linial, ProteinBERT: A universal deep-learning model of protein sequence and function, *Bioinformatics*, **38** (2022), 2102–2110. https://doi.org/10.1093/bioinformatics/btac020

25. V. Thumuluri, J. J. A. Armenteros, A. R. Johansen, H. Nielsen, O. Winther, DeepLoc 2.0: Multi-label subcellular localization prediction using protein language models, *Nucleic Acids Res.*, **50** (2022), W228–W234. https://doi.org/10.1093/nar/gkac278

26. L. Wang, C. Huang, M. Wang, Z. Xue, Y. Wang, NeuroPred-PLM: An interpretable and robust model for neuropeptide prediction by protein language model, *Brief. Bioinf.*, **24** (2023), bbad077. https://doi.org/10.1093/bib/bbad077

27. Z. Du, X. Ding, W. Hsu, A. Munir, Y. Xu, Y. Li, pLM4ACE: A protein language model based predictor for antihypertensive peptide screening, *Food Chem.*, **431** (2024), 137162–137162. https://doi.org/10.1016/j.foodchem.2023.137162

28. A. Villegas-Morcillo, A. M. Gomez, V. Sanchez, An analysis of protein language model embeddings for fold prediction, *Brief. Bioinf.*, **23** (2022), bbac142. https://doi.org/10.1093/bib/bbac142

29. P. Pratyush, S. Pokharel, H. Saigo, D. B. Kc, pLMSNOSite: An ensemble-based approach for predicting protein S-nitrosylation sites by integrating supervised word embedding and embedding from pre-trained protein language model, *BMC Bioinf.*, **24** (2023), 41. https://doi.org/10.1186/s12859-023-05164-9

30. X. Wang, Z. Ding, R. Wang, X. Lin, Deepro-Glu: Combination of convolutional neural network and Bi-LSTM models using ProtBert and handcrafted features to identify lysine glutarylation sites, *Brief. Bioinf.*, **24** (2023), bbac631. https://doi.org/10.1093/bib/bbac631

31. E. Fenoy, A.A. Edera, G. Stegmayer, Transfer learning in proteins: Evaluating novel protein learned representations for bioinformatics tasks, *Brief. Bioinf.*, **23** (2022), bbac232. https://doi.org/10.1093/bib/bbac232
32. X. Peng, X. Wang, Y. Guo, Z. Ge, F. Li, X. Gao, et al., RBP-TSTL is a two-stage transfer learning framework for genome-scale prediction of RNA-binding proteins, *Brief. Bioinf.*, **23** (2022), bbac215. https://doi.org/10.1093/bib/bbac215
33. B. Boeckmann, A. Bairoch, R. Apweiler, M.C. Blatter, A. Estreicher, E. Gasteiger, et al., The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003, *Nucleic Acids Res.*, **31** (2003), 365–370. https://doi.org/10.1093/nar/gkg095
34. B. E. Suzek, Y. Wang, H. Huang, P. B. McGarvey, C. H. Wu, C. UniProt, UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches, *Bioinformatics*, **31** (2015), 926–932. https://doi.org/10.1093/bioinformatics/btu739
35. A. G. Murzin, S. E. Brenner, T. Hubbard, C. Chothia, SCOP-A structural classification of proteins database for the investigation of sequences and structures, *J. Mol. Biol.*, **247** (1995), 536–540. https://doi.org/10.1016/s0022-2836(05)80134-2
36. R. D. Finn, A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, et al., Pfam: The protein families database, *Nucleic Acids Res.*, **42** (2014), D222–D230. https://doi.org/10.1093/nar/gkt1223
37. M. Steinegger, M. Mirdita, J. Soeding, Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold, *Nat. Methods*, **16** (2019), 603–606. https://doi.org/10.1038/s41592-019-0437-4
38. N.V . Chawla, K. W. Bowyer, L. O. Hall, W.P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, *J. Artif. Intell. Res.*, **16** (2002), 321–357. https://doi.org/10.1613/jair.953
39. T. Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollar, Focal loss for dense object detection, *IEEE Trans. Pattern Anal. Mach. Intell.*, **42** (2020), 318–327. https://doi.org/10.1109/tpami.2018.2858826
40. S. Jadon, Ieee, A survey of loss functions for semantic segmentation, in *IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, (2020), 115–121. https://doi.org/10.1109/cibcb48159.2020.9277638
41. S. S. Yuan, D. Gao, X. Q. Xie, C. Y. Ma, W. Su, Z. Y. Zhang, et al., IBPred: A sequence-based predictor for identifying ion binding protein in phage, *Comput. Struct. Biotechnol. J.*, **20** (2022), 4942–4951. https://doi.org/10.1016/j.csbj.2022.08.053
42. Y. H. Wang, Y. F. Zhang, Y. Zhang, Z. F. Gu, Z. Y. Zhang, H. Lin, et al., Identification of adaptor proteins using the ANOVA feature selection technique, *Methods*, **208** (2022), 42–47. https://doi.org/10.1016/j.ymeth.2022.10.008

Supplementary

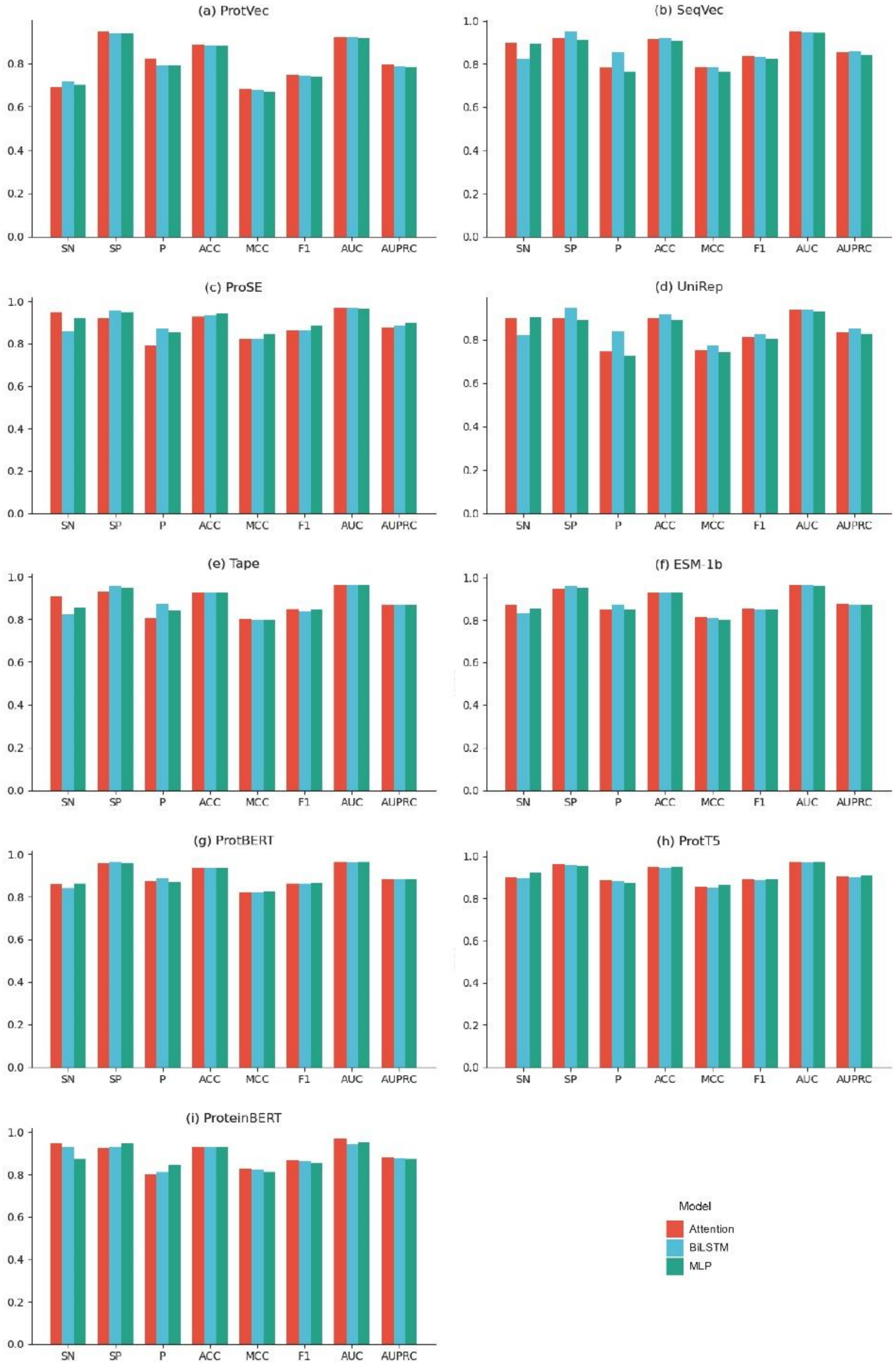

Figure S1. Performance comparison of nine PLM embeddings on the 5-fold CV.

Table S1. Description of the nine PLM feature embeddings.

Name	Description
ProtVec	ProtVec is the word vector-based protein representation, which was trained on the Swiss-Prot database through a Skip-gram neural network.
SeqVec	SeqVec was trained on the UniRef50 database by using an architecture composed of a convolutional layer and two BiLSTM layers, which was designed to reduce the risk of overfitting by sharing weights between the forward and the backward LSTMs. The encoding vector per residue obtained by two LSTMs were concatenated and then averaged to get a single vector per protein.
ProSE	The structure of ProSE is a three-layer BiLSTM similar to the SeqVec structure, with the difference that it uses not only the sequence data but also the structural information of the proteins.
UniRep	UniRep contains a layer of multiplicative LSTM with 1900 hidden units, which was trained on the UniRef50 database. UniRep can capture important information about the protein's structure, function, and evolutionary relationships.
Tape	Tape utilized a 12-layer Transformer with 512 hidden units and 8 attention heads, which were trained on the Pfam database. Tape aims to leverage the power of transformers to capture long-range dependencies and context in protein sequences.
ESM-1b	ESM-1b is a variant of the ESM model for protein sequence representation, which consists of 33 transformer layers with embedding dimension of 1280. The model was trained on the UniRef50 database via the masked language modeling objective.
ProtBERT	ProtBERT is a powerful protein language model generated by self-supervised training on the BFD and UniRef databases. It is a multi-layer bidirectional transformer encoder, which is leveraged for the more comprehensive representation of protein sequences.
ProtT5	ProtT5 is pretrained in a self-supervised manner on the BFD and UniRef50 database. Unlike the other PLMs, ProtT5 is composed of an encoder that converts a source language into an embedding space and a decoder that utilizes the encoder’s embedding to produce a translation in a target language.
ProteinBERT	ProteinBERT was pretrained on protein sequences and Gene Ontology annotations extracted from the UniRef database. Unlike the classic transformers, ProteinBERT is a denoising auto-encoder model and contains both local (residue level) and global (sequence level) representations.

Table S2. Description of the parameters required by NCSP-PLM.

Model	Layer	Configuration
MLP	Dense	Units: PLM embedding Dimension
	BatchNorm	-
	Dense	Units: 512
	BatchNorm	-
	Dense	Units:1
	Activation	Sigmoid
Attention	Attention	-
	Dense	Units: 512
	BatchNorm	-
	Dense	Units: 1
	Activation	Sigmoid
BiLSTM	Bidirectional LSTM	Units: 512
		Return Sequences: True

Model	Layer	Configuration
		Merge Mode: Concat
	Flatten	-
	BatchNorm	-
	Dense	Units: 512
	BatchNorm	-
	Dense	Units: 1
	Activation	Sigmoid