Computational and Structural Biotechnology Journal 27 (2025) 1350–1358

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research Article

iNClassSec-ESM: Discovering potential non-classical secreted proteins through a novel protein language model

Yizhou Shao, Taigang Liu *

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

ARTICLE INFO

Keywords:
Non-classical secreted protein
Protein language model
Embeddings
Ensemble learning
Protein representation

ABSTRACT

Non-classical secreted proteins (NCSPs) are a class of proteins lacking signal peptides, secreted by Gram-positive bacteria through non-classical secretion pathways. With the increasing demand for highly secreted proteins in recent years, non-classical secretion pathways have received more attention due to their advantages over classical secretion pathways (Sec/Tat). However, because the mechanisms of non-classical secretion pathways are not yet clear, identifying NCSPs through biological experiments is expensive and time-consuming, making it imperative to develop computational methods to address this issue. Existing NCSP prediction methods mainly use traditional handcrafted features to represent proteins from sequence information, which limits the models' ability to capture complex protein characteristics. In this study, we proposed a novel NCSP predictor, iNClassSec-ESM, which combined deep learning with traditional classifiers to enhance prediction performance. iNClassSec-ESM integrates an XGBoost model trained on comprehensive handcrafted features and a Deep Neural Network (DNN) trained on hidden layer embeddings from the protein language model (PLM) ESM3. The ESM3 is the recently proposed multimodal PLM and has not yet been fully explored in terms of protein representation. Therefore, we extracted hidden layer embeddings from ESM3 as inputs for multiple classifiers and deep learning networks, and compared them with existing PLMs. Benchmark experiments indicate that iNClassSec-ESM outperforms most of existing methods across multiple performance metrics and could serve as an effective tool for discovering potential NCSPs. Additionally, the ESM3 hidden layer embeddings, as an innovative protein representation method, show great potential for the application in broader protein-related classification tasks. The source code of iNClassSec-ESM and the ESM3 embeddings extraction script are publicly available at https://github.com/AmamiyaHoshie/iNClassSec-ESM/.

1. Introduction

Protein secretion is indispensable for cellular communication in eukaryotic organisms and plays a pivotal role in the vast majority of their physiological processes. Most eukaryotic secreted proteins are born with an amino-terminal signal peptide, termed a leader sequence, that directs them into the endoplasmic reticulum, where they mature and are transported to the plasma membrane through the Golgi apparatus [1]. Protein secretion by Gram-positive bacteria through such processes is commonly referred to as the classical secretion pathway, which is further divided into two systems, i.e., the general secretory (Sec) pathway [2] and the Twin-arginine translocation (Tat) pathway [3]. Recent studies have found that certain secreted proteins can be released into the external environment without the use of signal peptides, following a route distinct from the classical secretion pathway. This alternative mechanism is known as the non-classical secretory pathway [4]. *B.subtilis*, as the most studied Gram-positive bacterium, has been extensively utilized as a microbial cell factory and holds a significant position in both clinical research and the food industry. In studies involving *B.subtilis* and other organisms [5,6], an increasing number of proteins have been discovered in experiments to be capable of having their secretion mediated by a non-classical pathway. These proteins are referred to as non-classical secreted proteins (NCSPs) [5,7]. Further studies have revealed that the expression of enzymes with significant roles, such as Pullulanase (EC 3.2.1.41) and 1,4-α-Glucan Branching Enzyme (GBE; EC 2.4.1.18), via the non-classical secretion pathway in *B.subtilis* results in a significant increase in both secretion rate and activity compared to secretion processes guided by signal peptides [8,9]. The underlying reason is that

* Corresponding author.
E-mail address: tgliu@shou.edu.cn (T. Liu).

https://doi.org/10.1016/j.csbj.2025.03.043
Received 19 December 2024; Received in revised form 15 March 2025; Accepted 26 March 2025

Available online 28 March 2025

each step in the classical secretion pathway requires the involvement of translocation components, a strategy that affects the efficiency of protein secretion [10].

Benefiting from the significant role of the non-classical secretion pathway and the increasing demand for high-level protein secretion, researchers are dedicating efforts to identifying NCSPs in various microorganisms [5,7,11,12]. Given that the mechanisms of the non-classical secretion pathway remain unclear to date, the identification of NCSPs often requires fusing Sec- or Tat-dependent signal peptides and deleting Tat-related genes to block the Tat pathway. Additionally, researchers must ensure that secretion is not a consequence of cell lysis [10]. Such experimental procedures are cumbersome and resource-intensive. In conclusion, computational identification methods urgently need to be developed, characterized by speed and low cost, to efficiently and accurately identify NCSPs.

To date, several computational tools have been specifically designed to identify NCSPs based on their sequences. Bendtsen et al. [13] proposed SecretomeP, the first computational tool designed to identify NCSPs in mammals using sequence-based features. Shortly after, Bendtsen et al. [4] developed SecretomeP 2.0, which expanded the repertoire of predicted NCSPs across a wide range of bacterial species. Yu et al. developed SecretP [14] to simultaneously distinguish the classical secreted proteins (CSPs), non-secreted proteins, and NCSPs, which utilizes pseudo amino acid composition (PseAAC) and five additional features to train a support vector machine (SVM) classifier. Similarly, NClassG+ [15] was designed to classify NCSPs in Gram-positive bacteria by using feature vectors derived from frequencies, dipeptides, physicochemical factors, and Position-Specific Scoring Matrix (PSSM)-based features. With the continuous improvement of identification systems for non-classical secretion pathways in biological experiments, an increasing number of NCSPs have been recognized as important research targets, and their unique roles in various fields are gradually being uncovered [16].

As the number of the newly discovered NCSPs grows, a strong impetus has emerged to construct relatively large and high-quality datasets to improve the prediction of NCSPs. On the basis of a recent work [17], Zhang et al. [18] first constructed a high-quality benchmark dataset which includes more experimentally verified NCSPs and then adopted a two-layer Light Gradient Boosting Machine (LightGBM) to identify the NCSPs, named PeNGaRoo. Based on the same dataset, additional predictive models for NCSPs have been developed, including NonClasGP-Pred [19], ASPIRER [20], and iNSP-GCAAP [21].

Although the aforementioned prediction methods have achieved promising results, several challenges still remain to be addressed. First, the performance on the independent test dataset was far less impressive. Second, almost all predictors rely on handcrafted feature descriptors extracted from amino acid sequences by using specific algorithms, such as Amino Acid Composition (AAC), Dipeptide Composition (DPC) [22], Moran Autocorrelation [23] and various PSSM-based variants [24]. In recent years, protein language models (PLMs) are designed to understand and predict the properties and behaviors of proteins by leveraging natural language processing techniques [25]. These models were pre-trained on the large-scale databases such as UniProt [26] by using unsupervised or self-supervised learning and could capture complex patterns and structure-function relationships within amino acid sequences. Currently, several PLMs have been released, including UniRep [27], TAPE [28], ESM-1b [29], ProtT5 [30], and ESM-2 [31]. Recent studies have demonstrated that embeddings extracted from PLMs can be used as feature representation of proteins and applied in the various downstream tasks, yielding promising results. These applications encompass predicting interactions between drugs and protein targets [32], assessing conservation and variant effects [33], drug design [34], controllable protein design [35], enzyme function annotation [36], and so on. These advancements highlight the versatility and potential of PLMs in transforming various aspects of bioinformatics and protein engineering.

ESM3 is a brand-new PLM released by EvolutionaryScale [37]. Unlike existing PLMs, ESM3 is trained on a vast dataset of protein sequences, incorporating evolutionary information to improve its predictions. Although ESM3 has demonstrated outstanding performance on various tasks, its capacity for protein representation has not been widely tested.

In this study, we developed the iNClassSec-ESM model which extracted the hidden layers of ESM3 architecture to represent proteins and examined its ability for the identification of NCSPs. The overview of the proposed iNClassSec-ESM framework is illustrated in Fig. 1. This approach integrates two primary sub-models: an XGBoost classifier and a deep neural network (DNN). The XGBoost component is trained on a comprehensive set of handcrafted features, while the DNN utilizes hidden layer embeddings derived from ESM3. The output probabilities of both sub-models are then fused by a Logistic Regression (LR) meta-learner to perform the final classification. Benchmark experiments demonstrate that the iNClassSec-ESM model outperforms most of existing NCSP prediction tools on the independent test. Moreover, the hidden layer embeddings of ESM3 can serve as a novel and effective method for protein representation.

2. Materials and methods

2.1. Datasets

The dataset utilized in this study primarily follows the approach of Zhang et al. [18]. All NCSP samples were obtained from the UniProt database, and each NCSP was identified by at least three research groups in three different bacterial species [38]. The CSP samples for the training set were adopted from Bendtsen et al. [13], derived from proteins in the *Firmicutes* phylum explicitly annotated as cytoplasmic. Subsequently, we used the CD-HIT program [39] to reduce sequence similarity within the dataset to 80%, yielding 157 positive samples and 446 negative samples. For the independent test set, one-tenth of the NCSPs from previous studies and Zhang et al.'s dataset [18] were randomly chosen as positive samples. In contrast, negative samples were obtained from the UniProt database, focusing on proteins annotated as "cytoplasm" or "cytoplasmic" yet not labeled as "secreted", thereby excluding any form of secreted protein. The sequence lengths of the positive and negative samples had similar distributions to avoid potential bias. As a result, 141 validated NCSPs and 446 cytoplasmic proteins were designated as the training set, while the final independent test set consisted of 34 positive and 34 negative samples, consistent with previous research.

2.2. Feature extraction

To comprehensively extract the characteristic related to NCSPs, we fused traditional handcrafted features and hidden layer embeddings from popular PLMs.

2.2.1. Handcrafted features

In this study, we adopted 11 types of handcrafted features, including Amino Acid Composition (AAC), Dipeptide Composition (DPC), Composition of K-Spaced Amino Acid Pairs (CKSAAP), Composition, Transition, and Distribution (CTD), Composition of Triads (CTriad), Grouped Amino Acid Composition (GAAC), Grouped Dipeptide Composition (GDPC), Moran Autocorrelation (Moran), and Pseudo PSSM (Pse-PSSM). The dimensions and parameter values of all feature descriptors are shown in Table 1. All feature descriptors were computed using the iFeature tool [40] and the POSSUM software package [41].

2.2.2. PLM embeddings

To obtain consistent representations, we adopted the latest ESM3 model [37] along with ESM-1b [29], ESM-2-650M [31], and ProtT5 [30] to extract embedding matrices. Protein sequences were first encoded by using each model's specific tokenizer and then input into the

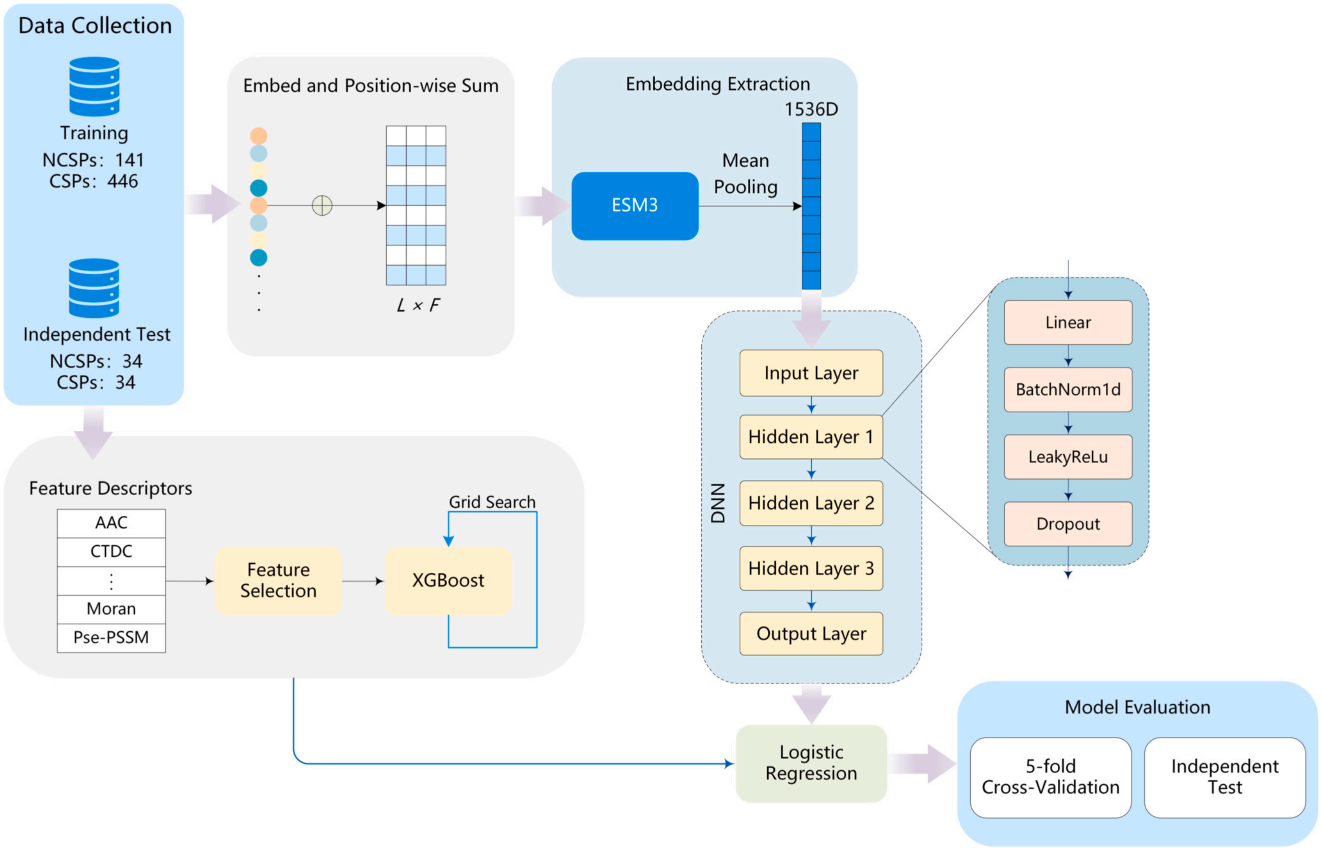

Fig. 1. The overall framework of the iNClassSec-ESM.

Table 1
Description of handcrafted features.

Feature Name	Feature Dimension	Parameter Value
AAC	20	—
DPC	400	—
CKSAAP	1600	$K = 3$
CTD	273	—
CTriad	343	—
GAAC	5	—
GDPC	25	—
Moran	1062	$Lag = 2$
Pse-PSSM	40	$Alpha = 1$

Table 2
Overview of PLMs used in this study.

Name	Dimension	Database
ProtT5	1024	UniRef50
ESM-1b	1280	UniRef50 + MSA
ESM-2-650M	1280	UniRef50 (Sample UniRef90)
ESM3	1536	UniProt + MGnify + JGI + OAS

models for forward propagation. As a result, the embedding matrix was generated with the size of $L \times F$, where L denotes the length of query protein and F denotes the dimension of the individual embedding for each amino acid. Finally, we applied the average pooling to transform the embedding matrix into a feature vector with the uniform dimension.

Table 2 summarizes the training data sources and hidden layer embedding dimensions of the four PLMs used in this study, each characterized by distinct architectures and training data sources. ProtT5 is a transformer-based model inspired by the T5 architecture, featuring an embedding dimension of 1024, and consisting of 24 layers of encoder and decoder stacks with 16 attention heads per layer. It was trained on the UniRef50 database [26], which clusters UniRef100 sequences at the 50% sequence identity levels. ESM-1b and ESM-2-650M, developed by Meta AI, are both built on advanced transformer architectures with an embedding dimension of 1280. ESM-1b comprises 33 transformer encoder layers with 20 attention heads per layer and was trained on the UniRef50 database along with Multiple Sequence Alignment (MSA) data [42] to enhance sequence diversity [29]. ESM-2-650M maintains the same embedding dimension and foundational architecture but incorporates additional layers and optimized attention mechanisms to improve performance, utilizing the UniRef50 database along with a subset sampled from the UniRef90 database [31]. ESM3, the latest in the Evolutionary Scale Modeling series, boasts an increased embedding dimension of 1536 and features a more complex and deeper transformer architecture with 48 encoder layers and 24 attention heads per layer, supporting its multimodal data integration capabilities. ESM3 was trained on a combination of the UniProt [26], MGnify [43], Joint Genome Institute (JGI) [44], Observed Antibody Space (OAS) [45], Protein Data Bank (PDB) [46], AlphaFoldDB [47], and ESMAtlas [48] databases.

2.3. Machine learning classifiers

In this study, we employed six distinct machine learning classifiers to perform the prediction of NCSPs based on both handcrafted features and PLM embeddings, including K-Nearest Neighbors (KNN), Random Forest (RF), SVM, XGBoost, LightGBM, and CatBoost. Each classifier brings unique advantages, enabling effective handling of diverse data distributions and feature interactions, thereby serving as robust benchmarks for our analysis. KNN, as a non-parametric and distance-based method, is particularly well-suited for datasets with clear and distinct patterns [49]. RF enhances classification stability and robustness by aggregating multiple decision trees, making it ideal for datasets with numerous features and complex nonlinear relationships [50]. SVM aims to find a hyperplane that maximizes the margin between classes, demonstrating

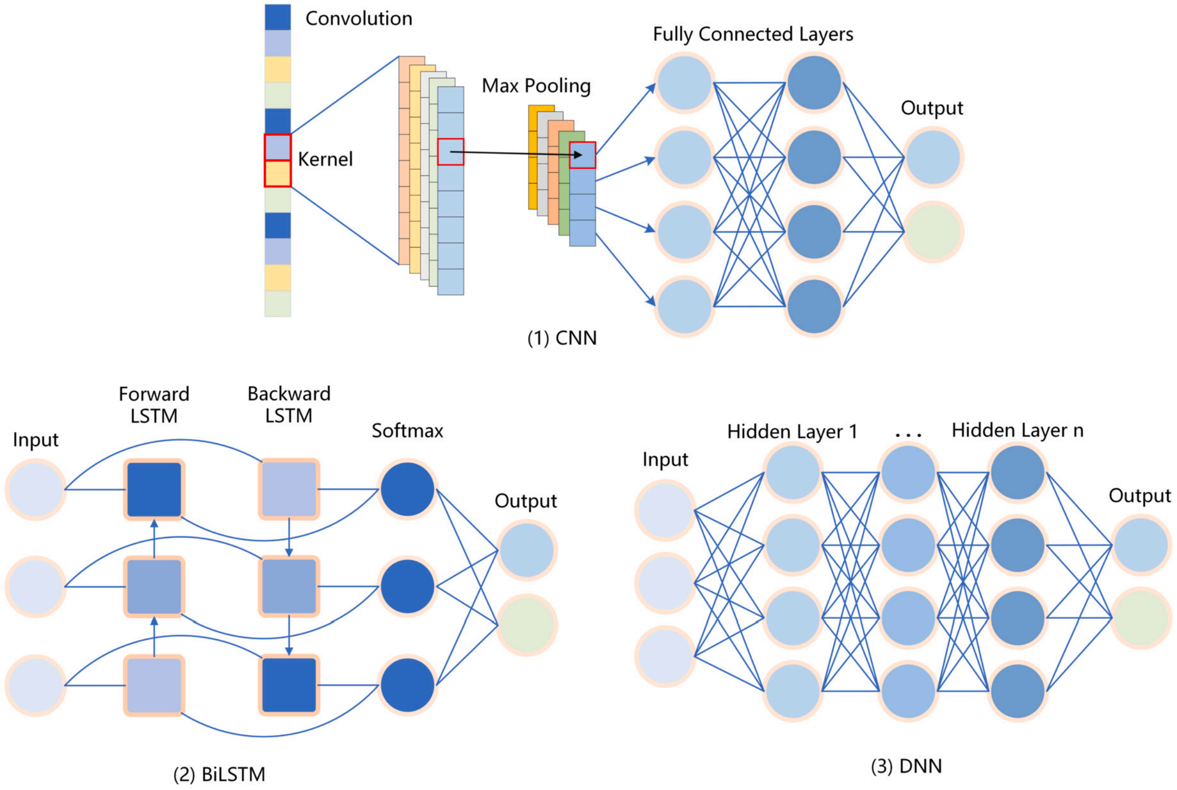

Fig. 2. Overview of three neural network architectures used in this study.

exceptional performance in high-dimensional spaces and scenarios with limited sample sizes [51]. XGBoost, as a gradient boosting algorithm, combines high training efficiency with regularization techniques to mitigate overfitting [52]. LightGBM optimizes training speed and memory efficiency through histogram-based bucketing and a leaf-wise tree growth strategy, making it suitable for large-scale datasets [53]. CatBoost natively supports categorical features and employs symmetric tree structures to reduce data leakage and overfitting, particularly benefiting datasets with high-cardinality categorical variables [54]. By leveraging these diverse classifiers, our study ensures comprehensive coverage of different machine learning paradigms, facilitating a thorough evaluation of their performance based on both handcrafted features and PLM embeddings.

2.4. Deep learning network architectures

To further investigate the effectiveness of hidden layer embeddings from PLMs across different classifier types, we employed three distinct neural network architectures, including Convolutional Neural Network (CNN), Bidirectional Long Short-Term Memory (BiLSTM), and DNN. As shown in Fig. 2, each network architecture possesses unique capabilities in learning and representing sequential and structural information. BiLSTM introduces bidirectionality in processing sequential data, enabling the model to utilize both past and future context within a sequence, which is particularly beneficial for understanding dependencies within the embeddings. CNN employs convolutional kernels to extract local patterns from feature vectors, making it especially effective in identifying spatial relationships within fixed-size embeddings. DNN is an artificial neural network composed of multiple layers of neurons that automatically learn and extract complex patterns from large datasets. Its deep, layered structure enables the modeling of intricate non-linear relationships, making DNNs highly effective for various downstream tasks. By implementing these diverse neural network architectures, our study aims to explore how different network structures impact the classification performance when using PLM embeddings as inputs.

2.5. Feature engineering

Imbalanced data classification often leads to prediction biases that favor the majority class. Additionally, directly concatenating handcrafted features results in high-dimensional feature vectors, which significantly increases training time and adversely impacts the model's overall performance. Therefore, feature engineering is essential for our study.

2.5.1. Data imbalance problem solving

To address the issue of imbalanced data, we employed three different approaches: synthetic minority oversampling technique (SMOTE) strategy, weighted cross-entropy (WCE) method and focal loss. SMOTE strategy balances the class distribution by generating new minority class samples [55]. WCE method adjusts the loss function by assigning different weights to different classes [56]. Focal loss modifies the loss calculation by assigning higher weights to hard-to-classify samples [57]. These three methods address the data imbalance problem from different perspectives to improve the training effectiveness of the model. The detailed information of these methods is provided in Supplementary Text S1.

2.5.2. Feature selection techniques

To address the issue of high-dimensional feature vectors, we adopted three viable feature selection methods to reduce the redundant and irrelevant features, including model-based feature selection, mutual information, and chi-square (Chi2) test. Model-based feature selection utilizes machine learning models to evaluate the importance of each feature and selects the most representative features by using an incremental stepwise greedy method [58]. Mutual information method measures the amount of mutual information between features and the target variable, selecting features that have a high correlation with the target [59]. Chi2 test employs statistical methods to assess the independence between each feature and the target variable, selecting features that exhibit strong dependence [60]. These three methods evaluate and select features from different perspectives to reduce feature dimensionality, thereby enhancing both the training efficiency and classification performance of the model. Detailed descriptions of these methods are provided in Supplementary Text S2.

2.6. Performance assessment

To comprehensively evaluate the classification performance of our models, we performed the 5-fold cross-validation and the independent test on the benchmark datasets and reported six commonly used performance metrics [20,61], including Precision, Recall, Accuracy, Matthews Correlation Coefficient (MCC), F1-Score, and Specificity. The formulas for these metrics are as follows:

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

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

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

where TP, TN, FP, and FN denote the numbers of true positives, true negatives, false positives, and false negatives, respectively.

We also computed the area under the receiver operating characteristic (ROC) curve (AUROC) and the area under the Precision-Recall (PR) curve (AUPRC) as additional evaluation metrics to understand the models' performance from different perspectives. These metrics work together to ensure that we can comprehensively and meticulously assess the models' classification performance across different classes, thereby providing reliable foundations for model optimization and selection.

3. Results and discussion

3.1. Performance of handcrafted features

Traditional handcrafted features have been widely utilized in previous studies to build reliable learning models. However, comprehensive test has not yet been conducted to examine the performance of these features for the identification of NCSPs. To address this gap, we adopted 11 types of features and 6 machine learning classifiers to construct 66 baseline models for the prediction of NCSPs. The performance of these models was evaluated by performing the 5-fold cross-validation. Detailed experimental results are provided in Supplementary Table S1. All baseline models achieved MCC scores exceeding 0.5 except for Moran-based models, indicating that these features could effectively characterize NCSPs. In addition, We utilized the RF method to assess the importance of these features. From Fig. 3, all features contributed positively to the performance of the classifiers, suggesting that combining all features could further enhance the model's performance. So we evaluated the prediction ability of each classifier combined with all features. The experimental results indicated that the fused features could indeed achieve better performance than single feature category. In particular, XGBoost outperformed other classifiers across most evaluation metrics, leading us to select XGBoost as the final classifier combined with all handcrafted features.

To further enhance model performance, we adopted three feature selection methods to reduce dimensionality, including model-based feature selection, mutual information, and Chi2 test. Leveraging the XGBoost classifier, the selected features based on the Chi2 test yielded the best performance on the 5-fold cross-validation (see Supplementary Table S2), with an average MCC of 0.6829, F1-Score of 0.7508, AUROC of 0.9409. Additional test on the independent test set reconfirmed the efficacy of the Chi2 feature selection (see Supplementary Table S3). Furthermore, we examined the proportion of each feature category in these features selected by the Chi2 test. As shown in Fig. 4, all feature categories contributed to the final selected feature set and their distribution aligned closely with their importance scores computed by the XGBoost classifier.

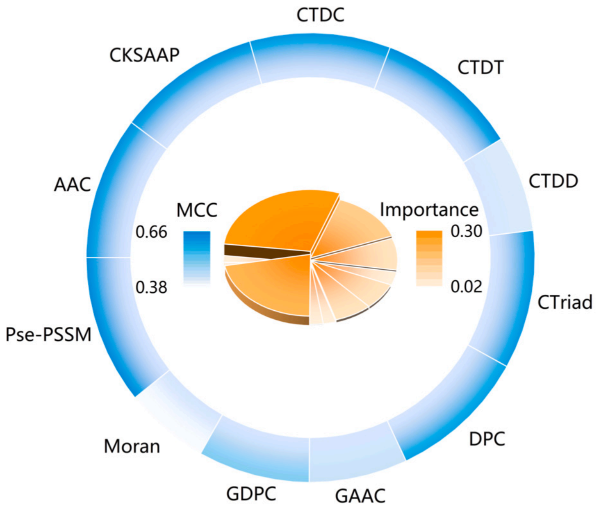

Fig. 3. Comparison of the average MCC scores across six classifiers and the feature importance scores.

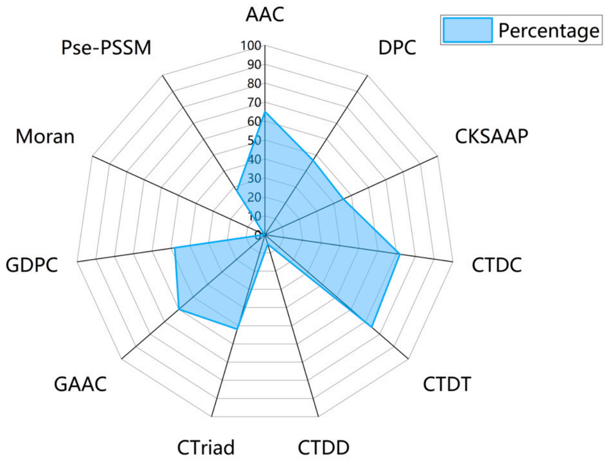

Fig. 4. Proportion of each feature category within the final feature set selected by the Chi2 test.

Moreover, we applied the SMOTE strategy to address the issue of data imbalance. As can be seen from Supplementary Table S4, the model only based on the SMOTE technique achieved the slightly improvement with an averaged AUROC of 0.9409 on the 5-fold cross-validation. The best Recall value (0.7591) was obtained by simultaneously applying the Chi2 feature selection and the SMOTE strategy. A similar conclusion was also reached on the independent test (see Supplementary Table S5). This indicates that the combined strategy facilitates the model to identify more NCSPs. In addition, grid search was performed to optimize the hyperparameters of the final XGBoost model based on the 5-fold cross-validation. For the sake of convenience, this model was named "M1".

3.2. Performance of different PLM embeddings

In this section, we compared the performance of four popular PLM embeddings for the NCSP classification, including ProtT5, ESM-1b,

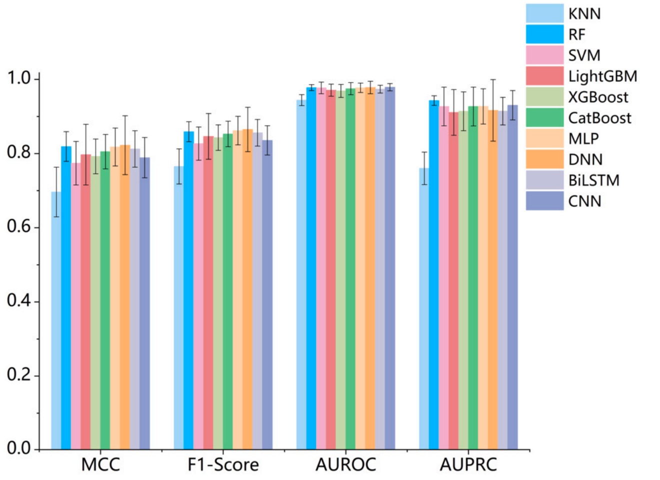

Fig. 5. Performance comparison of ESM3 embeddings across different models on the 5-fold cross-validation.

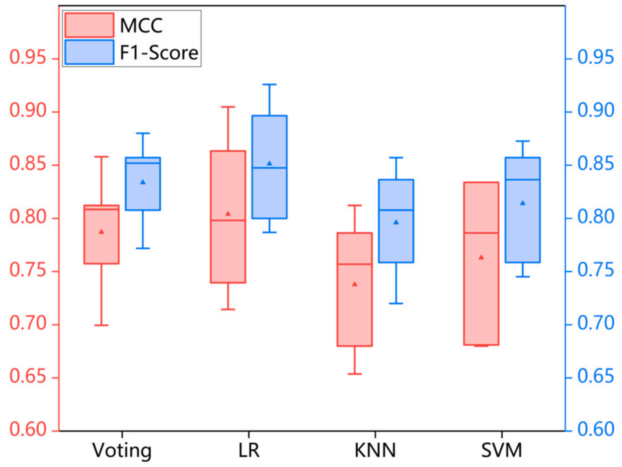

Fig. 6. Performance comparison of different ensemble strategies on the 5-fold cross-validation.

ESM-2, and ESM3. All the experiments were conducted on the 5-fold cross-validation by training XGBoost models and the corresponding results were reported in Supplementary Table S6. We observed that the ESM3 embeddings outperformed the other three widely used embeddings across all evaluation metrics. Additionally, leveraging the SMOTE strategy to keep data balance, the ESM3 embeddings remained significantly superior to the other embeddings. This reconfirmed that the ESM3 embeddings could effectively capture intricate patterns for NCSPs (see Supplementary Table S7).

To assess the potential of the ESM3 embeddings for representing NCSPs, we also trained six machine learning classifiers and three deep learning s on the 5-fold cross-validation. As illustrated in Fig. 5 and Supplementary Table S8, the DNN model achieved the best performance, with the average MCC of 0.8226, F1-Score of 0.8650. In contrast, the CNN and BiLSTM models also demonstrated good performance. Notably, the CNN model slightly outperformed the DNN in terms of AUROC and AUPRC metrics. However, CNN and BiLSTM still showed gaps in performance compared to the DNN across other evaluation metrics. Consequently, the DNN architecture was adopted as the final classifier for the subsequent analysis. In addition, we adjusted the loss function to further enhance the model's sensitivity to minority classes and improve its generalization ability on imbalanced datasets. The prediction results of three loss functions were listed in Supplementary Table S9, including WCE, focal loss, and default. The WCE was eventually adopted as the loss function due to its superior performance.

Then, a thorough parameter search was conducted to identify an optimal DNN structure and training strategy. The finalized model comprises three fully connected layers with 128 hidden units each, batch normalization, and LeakyReLU activations. LeakyReLU activations were selected over standard ReLU for their stable training performance, particularly when handling numerical uncertainties inherent in ESM3 embeddings. To mitigate overfitting, a dropout rate of 0.5 was applied in deeper layers. The AdamW optimization algorithm was chosen due to its adaptive moment estimation and decoupled weight decay mechanism, ensuring stable convergence and effective regularization. Additionally, a plateau-based scheduling mechanism automatically reduces the learning rate if improvement stalls, complemented by an early stopping criterion to further prevent overfitting. A batch size of 64 and a training duration of 50 epochs were determined sufficient for achieving optimal model performance within reasonable computational time. Collectively, these design choices consistently outperformed alternative configurations during validation. The optimized DNN model was named “M2” for convenience.

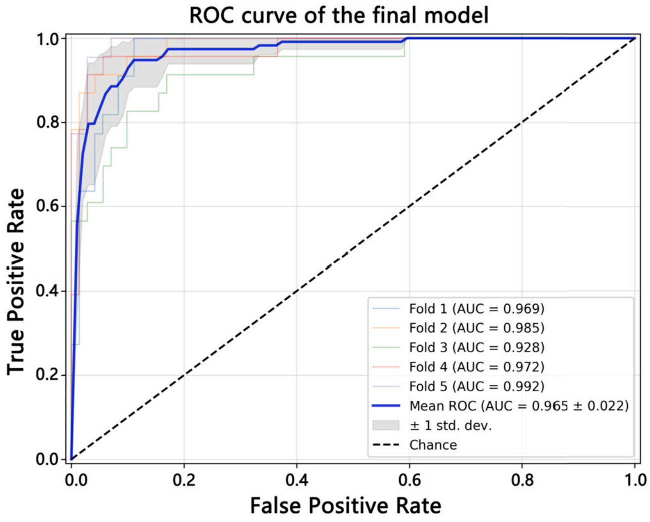

Fig. 7. ROC curves of the final model on the 5-fold cross-validation.

3.3. Performance comparison of different ensemble methods

In this section, the stacking-based ensemble learning together with the voting strategy was adopted to improve the overall performance of the NCSPs identification by combining the output of two base learners mentioned above (i.e., M1 and M2). Specially, LR, KNN, and SVM were employed as the meta learner in the stacking-based models, respectively. These ensemble models were evaluated on the 5-fold cross-validation and the corresponding metrics were reported in Fig. 6 and Supplementary Table S10. As can be seen, the LR model achieved the best overall performance, with the average Recall of 0.8655, Accuracy of 0.9268, MCC of 0.8040, F1-Score of 0.8514, and AUPRC of 0.9224. Additionally, the voting-based model obtained the highest Specificity, Precision and AUROC values. By comparison, we ultimately selected LR as the meta learner to construct the ensemble model (termed “M3”) for the accurate and robust prediction of NCSPs. The ROC curves of the final ensemble model on the 5-fold cross-validation were illustrated in Fig. 7, indicating that the stacking-based ensemble approach delivers stable and reliable predictive performance.

Next, we compared the meta learner with two base learners on the 5-fold cross-validation. As summarized in Table 3, the meta learner M3 achieved superior results across all evaluation metrics except for Recall. Although previous analyses indicated that ESM3 embeddings offer distinct advantages in representing NCSPs, the complex architecture of the

Table 3
Performance of the meta learner and two base learners on the 5-fold cross-validation.

Model	Recall	Specificity	Precision	Accuracy	MCC	F1-Score	AUROC	AUPRC
M1	0.7591 ± 0.0327	0.9350 ± 0.0336	0.7968 ± 0.1038	0.8928 ± 0.0300	0.7070 ± 0.0810	0.7751 ± 0.0570	0.9403 ± 0.0229	0.8443 ± 0.0791
M2	**0.8775 ± 0.1174**	0.9147 ± 0.0346	0.7955 ± 0.0871	0.9105 ± 0.0227	0.7751 ± 0.0564	0.8235 ± 0.0473	0.9649 ± 0.0225	0.8996 ± 0.0574
M3	0.8655 ± 0.0515	**0.9462 ± 0.0260**	**0.8399 ± 0.0712**	**0.9268 ± 0.0275**	**0.8040 ± 0.0720**	**0.8514 ± 0.0537**	**0.9694 ± 0.0221**	**0.9247 ± 0.0431**

[*] Performance is expressed as mean ± standard deviation while the bold values indicate the best performance.

DNN component (M2) raised concerns about potential overfitting. The meta learner M3 mitigated the overfitting issue to some extent and thus improved overall predictive performance by integrating two base learners.

Furthermore, we further optimized the classification threshold of the ensemble model M3 to 0.3 by maximizing the MCC value on the 5-fold cross-validation. The adjusted M3 model was also named iNClassSec-ESM, providing a more reliable prediction for the unbalanced classification.

We also applied the t-distributed stochastic neighbor embedding (t-SNE) technique to visualize high-dimensional input features of three models. As can be seen from Supplementary Figure S1, the PLM embeddings of the M2 model formed tighter and more distinct class clusters than the traditional handcrafted features of the M1 model, demonstrating that the ESM3 embeddings have the capacity to capture inherently discriminative patterns of NCSPs. Notably, the feature distribution input to the meta learner M3 revealed clear separations between positive and negative samples, suggesting a synergistic effect between ESM3 embeddings and handcrafted features.

3.4. Performance comparison with state-of-the-art methods on the independent test

We compared iNClassSec-ESM with four state-of-the-art methods trained on the same dataset by performing the independent test, including PeNGaRoo [18], NonClasGP-Pred [19], ASPIRER [20], and iNSP-GCAAP [21]. These models all extracted traditional handcrafted features to train the machine learning or deep learning classifiers for the identification of NCSPs. Details of the comparative results are provided in Table 4. The ROC and PR curves of iNClassSec-ESM were illustrated in Figs. 8 and 9.

As shown in Table 4, iNClassSec-ESM attains AUROC of 0.9654 and AUPRC of 0.9646, outperforming ASPIRER, which specifically optimized for these two metrics. iNClassSec-ESM also achieves Recall of 0.9412, Accuracy of 0.9118, MCC of 0.8250, and F1-score of 0.9143, surpassing the performance of other methods. Although Specificity and Precision are slightly lower than the best values from other models, iNClassSec-ESM demonstrates robust and competitive performance across nearly all evaluation metrics.

The superiority of iNClassSec-ESM over previous approaches could be largely attributed to the high-quality sequence representations provided by ESM3 embeddings, which enrich the feature space with biologically meaningful information that conventional feature engineering struggles to extract. Therefore, ESM3 embeddings offer a robust foundation for ongoing model optimization, demonstrating the potential of ESM3 to advance the applicability of other protein-related classification tasks.

Moreover, the development of iNClassSec-ESM provides biological researchers with a reliable computational tool to effectively prioritize candidate NCSPs, significantly reducing the laborious and resource-intensive experimental procedures currently required. Although this study primarily leverages sequence-based features, incorporating structural information into future models could yield valuable insights, as protein structure often provides critical clues regarding secretion mechanisms, interaction interfaces, and localization signals that sequence alone cannot fully reveal. Given that ESM3 inherently functions as a multimodal protein language model, systematically exploring its potential in structural feature extraction could further enhance the accuracy and biological interpretability of NCSP prediction.

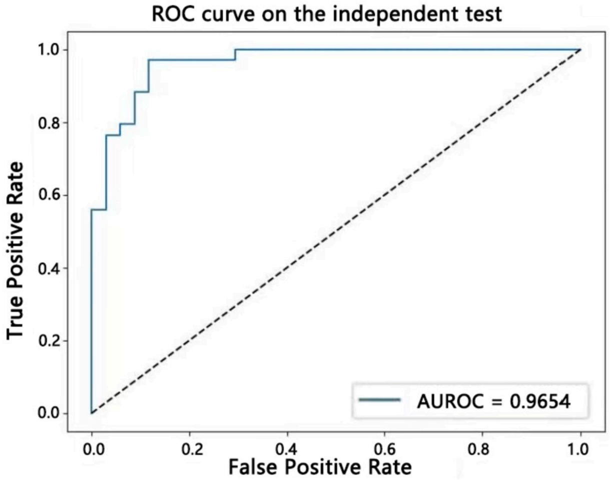

Fig. 8. ROC curve of iNClassSec-ESM based on the independent test.

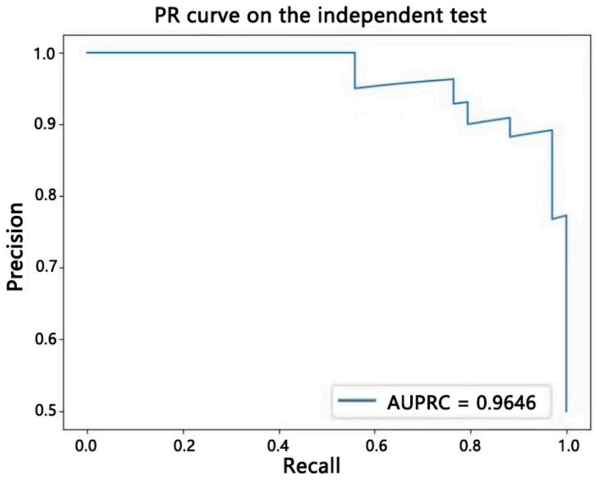

Fig. 9. PR curve of iNClassSec-ESM based on the independent test.

Future directions for improving iNClassSec-ESM will involve expanding the dataset through the systematic collection of newly reported NCSP samples from literature and databases, creating larger and more representative datasets that enhance model reliability and generalizability. We also plan to further investigate the structural representation capabilities of ESM3 embeddings, aiming to fully leverage the power of multimodal protein language modeling and to deepen our understanding of non-classical secretion mechanisms. Moreover, we are committed to developing an accessible web server based on iNClassSec-ESM that will provide researchers with a practical, rapid, and cost-effective tool for identifying NCSP candidates, ultimately accelerating efforts in exploring and engineering protein secretion systems across various microorganisms.

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

Method	Recall	Specificity	Precision	Accuracy	MCC	F1-Score	AUROC	AUPRC
SecretomeP	0.3529	0.8235	-	0.5882	0.2000	-	0.6799	-
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	-
iNClassSec-ESM	**0.9412**	0.8824	0.8889	**0.9118**	**0.8250**	**0.9143**	**0.9654**	**0.9646**

* The bold values indicate the best performance.

4. Conclusion

In this study, we developed a novel NCSP predictor, named iNClassSec-ESM, which integrates a DNN model trained on ESM3 embeddings and an XGBoost model trained on handcrafted features. Benchmark experimental results from the 5-fold cross-validation and the independent test demonstrate that iNClassSec-ESM outperforms most of existing state-of-the-art classifiers across multiple performance metrics, showcasing superior predictive capabilities. Furthermore, comparative and ablation experiments reveal that ESM3 hidden layer embeddings could significantly enhance the representation of protein sequence information compared to traditional handcrafted features, effectively capturing the intrinsic characteristics and distribution patterns of proteins. We anticipate that the developed iNClassSec-ESM will serve as an effective tool for the discovery and study of potential NCSPs in the future. Additionally, the ESM3 hidden layer embeddings exhibit substantial potential as an innovative protein representation method, which can be applied to a broader range of protein-related classification tasks, thereby advancing protein research and related fields.

5. Code and data availability

The code and datasets are publicly available at https://github.com/AmamiyaHoshie/iNClassSec-ESM/. Additionally, the derivative tool has been released at https://github.com/AmamiyaHoshie/iNClassSec-ESM-Prediction-Tool/.

CRediT authorship contribution statement

Yizhou Shao: Writing – original draft, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Taigang Liu:** Writing – review & editing, Supervision, Resources, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csbj.2025.03.043.

References

[1] Cavalli G, Cenci S. Autophagy and protein secretion. J Mol Biol 2020;432(8):2525–45.
[2] Wickner W, Driessen A, Hartl F-U. The enzymology of protein translocation across the escherichia coli plasma membrane. Annu Rev Biochem 1991;60(1):101–24.
[3] Berks BC, Palmer T, Sargent F. Protein targeting by the bacterial twin-arginine translocation (tat) pathway. Curr Opin Microbiol 2005;8(2):174–81.
[4] Bendtsen JD, Kiemer L, Fausbøll A, Brunak S. Non-classical protein secretion in bacteria. BMC Microbiol 2005;5(1):58.
[5] Xin C, Ban X, Gu Z, Li C, Cheng L, Hong Y, et al. Non-classical secretion of 1, 4-α-glucan branching enzymes without signal peptides in escherichia coli. Int J Biol Macromol 2019;132:759–65.
[6] Niu J, Meng F, Zhou Y, Zhang C, Lu Z, Lu F, et al. Non-classical secretion of a type i l-asparaginase in bacillus subtilis. Int J Biol Macromol 2021;180:677–83.
[7] Zhao X, Wang J, Li D, Ma F, Fang Y, Lu J, et al. Investigation of non-classical secretion of oxalate decarboxylase in bacillus mojavensis xh1 mediated by exopeptide yydf: mechanism and application. Int J Biol Macromol 2024;264:130662.
[8] Zhen J, Zheng H, Zhao X, Fu X, Yang S, Xu J, et al. Regulate the hydrophobic motif to enhance the non-classical secretory expression of pullulanase pula in bacillus subtilis. Int J Biol Macromol 2021;193:238–46.
[9] Xu T, Li Z, Gu Z, Li C, Cheng L, Hong Y, et al. The n-terminus of 1, 4-α-glucan branching enzyme plays an important role in its non-classical secretion in bacillus subtilis. Food Biosci 2023;52:102491.
[10] Chen J, Zhao L, Fu G, Zhou W, Sun Y, Zheng P, et al. A novel strategy for protein production using non-classical secretion pathway in bacillus subtilis. Microb Cell Fact 2016;15(1):69.
[11] Wang G, Chen H, Xia Y, Cui J, Gu Z, Song Y, et al. How are the non-classically secreted bacterial proteins released into the extracellular milieu? Curr Microbiol 2013;67(6):688–95.
[12] Pasztor L, Ziebandt A-K, Nega M, Schlag M, Haase S, Franz-Wachtel M, et al. Staphylococcal major autolysin (atl) is involved in excretion of cytoplasmic proteins. J Biol Chem 2010;285(47):36794–803.
[13] Bendtsen JD, Jensen LJ, Blom N, Von Heijne G, Brunak S. Feature-based prediction of non-classical and leaderless protein secretion. Protein Eng Des Sel 2004;17(4):349–56.
[14] Yu L, Guo Y, Zhang Z, Li Y, Li M, Li G, et al. Secretp: a new method for predicting mammalian secreted proteins. Peptides 2010;31(4):574–8.
[15] Restrepo-Montoya D, Pino C, Nino LF, Patarroyo ME, Patarroyo MA. Nclassg+: a classifier for non-classically secreted gram-positive bacterial proteins. BMC Bioinform 2011;12(1):21.
[16] Kang Q, Zhang D. Principle and potential applications of the non-classical protein secretory pathway in bacteria. Appl Microbiol Biotechnol 2020;104(3):953–65.
[17] Wang G, Xia Y, Song X, Ai L. Common non-classically secreted bacterial proteins with experimental evidence. Curr Microbiol 2016;72(1):102–11.
[18] Zhang Y, Yu S, Xie R, Li J, Leier A, Marquez-Lago TT, et al. PeNGaRoo, a combined gradient boosting and ensemble learning framework for predicting non-classical secreted proteins. Bioinformatics 2019;36(3):704–12.
[19] Wang C, Wu J, Xu L, Zou Q. Nonclasgp-pred: robust and efficient prediction of non-classically secreted proteins by integrating subset-specific optimal models of imbalanced data. Microbial Genom 2020;6(12):e000483.
[20] Wang X, Li F, Xu J, Rong J, Webb GI, Ge Z, et al. ASPIRER: a new computational approach for identifying non-classical secreted proteins based on deep learning. Brief Bioinform 2022;23(2):bbac031.
[21] Do TT, Nguyen-Vo T-H, Pham HT, Trinh QH, Nguyen BP. insp-gcaap: identifying nonclassical secreted proteins using global composition of amino acid properties. Proteomics 2023;23(1):2100134.
[22] Bhasin M, Raghava GP. Classification of nuclear receptors based on amino acid composition and dipeptide composition. J Biol Chem 2004;279(22):23262–6.
[23] Feng Z-P, Zhang C-T. Prediction of membrane protein types based on the hydrophobic index of amino acids. J Protein Chem 2000;19:269–75.
[24] Cai Y, Huang T, Hu L, Shi X, Xie L, Li Y. Prediction of lysine ubiquitination with mrmr feature selection and analysis. Amino Acids 2012;42:1387–95.
[25] Devlin J. Bert: pre-training of deep bidirectional transformers for language understanding. preprint. arXiv:1810.04805, 2018.
[26] Consortium U. Uniprot: a worldwide hub of protein knowledge. Nucleic Acids Res 2019;47(D1):D506–15.
[27] Alley EC, Khimulya G, Biswas S, AlQuraishi M, Church GM. Unified rational protein engineering with sequence-based deep representation learning. Nat Methods 2019;16(12):1315–22.
[28] Rao R, Bhattacharya N, Thomas N, Duan Y, Chen P, Canny J, et al. Evaluating protein transfer learning with tape. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R, editors. Advances in neural information processing systems, vol. 32. Curran Associates Inc.; 2019.
[29] Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci 2021;118(15):e2016239118.
[30] Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, et al. Prottrans: toward understanding the language of life through self-supervised learning. IEEE Trans Pattern Anal Mach Intell 2021;44(10):7112–27.

[31] Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. BioRxiv 2022;2022:500902.

[32] Singh R, Sledzieski S, Bryson B, Cowen L, Berger B. Contrastive learning in protein language space predicts interactions between drugs and protein targets. Proc Natl Acad Sci 2023;120(24):e2220778120.

[33] Marquet C, Heinzinger M, Olenyi T, Dallago C, Erckert K, Bernhofer M, et al. Embeddings from protein language models predict conservation and variant effects. Hum Genet 2022;141(10):1629–47.

[34] Moret M, Pachon Angona I, Cotos L, Yan S, Atz K, Brunner C, et al. Leveraging molecular structure and bioactivity with chemical language models for de novo drug design. Nat Commun 2023;14(1):114.

[35] Ferruz N, Höcker B. Controllable protein design with language models. Nat Mach Intell 2022;4(6):521–32.

[36] Thurimella K, Mohamed AM, Graham DB, Owens RM, La Rosa SL, Plichta DR, et al. Protein language models uncover carbohydrate-active enzyme function in metagenomics. bioRxiv 2023.

[37] Hayes T, Rao R, Akin H, Sofroniew NJ, Oktay D, Lin Z, et al. Simulating 500 million years of evolution with a language model. bioRxiv 2024.

[38] Wang G, Xia Y, Song X, Ai L. Common non-classically secreted bacterial proteins with experimental evidence. Curr Microbiol 2016;72:102–11.

[39] Huang Y, Niu B, Gao Y, Fu L, Li W. Cd-hit suite: a web server for clustering and comparing biological sequences. Bioinformatics 2010;26(5):680–2.

[40] Chen Z, Zhao P, Li F, Leier A, Marquez-Lago TT, Wang Y, et al. ifeature: a python package and web server for features extraction and selection from protein and peptide sequences. Bioinformatics 2018;34(14):2499–502.

[41] Wang J, Yang B, Revote J, Leier A, Marquez-Lago TT, Webb G, et al. Possum: a bioinformatics toolkit for generating numerical sequence feature descriptors based on pssm profiles. Bioinformatics 2017;33(17):2756–8.

[42] Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Res 2014;42(D1):D222–30.

[43] Mitchell AL, Almeida A, Beracochea M, Boland M, Burgin J, Cochrane G, et al. Mgnify: the microbiome analysis resource in 2020. Nucleic Acids Res 2020;48(D1):D570–8.

[44] Chen I-MA, Chu K, Palaniappan K, Ratner A, Huang J, Huntemann M, et al. The img/m data management and analysis system v. 7: content updates and new features. Nucleic Acids Res 2023;51(D1):D723–32.

[45] Olsen TH, Boyles F, Deane CM. Observed antibody space: a diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. Protein Sci 2022;31(1):141–6.

[46] Burley SK, Berman HM, Bhikadiya C, Bi C, Chen L, Di Costanzo L, et al. Rcsb protein data bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. Nucleic Acids Res 2019;47(D1):D464–74.

[47] Varadi M, Bertoni D, Magana P, Paramval U, Pidruchna I, Radhakrishnan M, et al. Alphafold protein structure database in 2024: providing structure coverage for over 214 million protein sequences. Nucleic Acids Res 2024;52(D1):D368–75.

[48] Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 2023;379(6637):1123–30.

[49] Guo G, Wang H, Bell D, Bi Y, Greer K. Knn model-based approach in classification. In: Meersman R, Tari Z, Schmidt DC, editors. On the move to meaningful Internet systems 2003: CoopIS, DOA, and ODBASE. Berlin, Heidelberg: Springer Berlin Heidelberg; 2003. p. 986–96.

[50] Breiman L. Random forests. Mach Learn 2001;45(1):5–32.

[51] Hearst M, Dumais S, Osuna E, Platt J, Scholkopf B. Support vector machines. IEEE Intell Syst Appl 1998;13(4):18–28.

[52] Chen T, Guestrin C. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, KDD '16. New York, NY, USA: Association for Computing Machinery; 2016. p. 785–94.

[53] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. Lightgbm: a highly efficient gradient boosting decision tree. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, editors. Advances in neural information processing systems, vol. 30. Curran Associates, Inc.; 2017.

[54] Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. Catboost: unbiased boosting with categorical features. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, et al., editors. Advances in neural information processing systems, vol. 31. Curran Associates, Inc.; 2018.

[55] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. Smote: synthetic minority oversampling technique. J Artif Intell Res 2002;16:321–57.

[56] Jadon S. A survey of loss functions for semantic segmentation. In: 2020 IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB); 2020. p. 1–7.

[57] Lin T-Y, Goyal P, Girshick R, He K, Dollar P. Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision (ICCV); 2017.

[58] Ng AY. Feature selection, l1 vs. l2 regularization, and rotational invariance. In: Proceedings of the twenty-first international conference on machine learning, ICML '04. New York, NY, USA: Association for Computing Machinery; 2004. p. 78.

[59] Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 2005;27(8):1226–38.

[60] Liu H, Setiono R. Chi2: feature selection and discretization of numeric attributes. In: Proceedings of 7th IEEE international conference on tools with artificial intelligence; 1995. p. 388–91.

[61] Gu Z-F, Hao Y-D, Wang T-Y, Cai P-L, Zhang Y, Deng K-J, et al. Prediction of blood–brain barrier penetrating peptides based on data augmentation with augur. BMC Biol 2024;22(1):86.