International Journal of Biological Macromolecules 308 (2025) 142309

Contents lists available at ScienceDirect

International Journal of Biological Macromolecules

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

CasPro-ESM2: Accurate identification of Cas proteins integrating pre-trained protein language model and multi-scale convolutional neural network

Chaorui Yan [a], Zilong Zhang [a], Junlin Xu [b], Yajie Meng [c], Shankai Yan [a], Leyi Wei [d,e], Quan Zou [f,g], Qingchen Zhang [a], Feifei Cui [a,*]

[a] School of Computer Science and Technology, Hainan University, Haikou 570228, China
[b] School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430081, Hubei, China
[c] School of Computer Science and Artificial Intelligence, Wuhan Textile University, Wuhan 430200, Hubei, China
[d] Centre for Artificial Intelligence driven Drug Discovery, Faculty of Applied Science, Macao Polytechnic University, Macao
[e] School of Informatics, Xiamen University, Xiamen, China
[f] Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu 610054, China
[g] Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou 324000, China

ARTICLE INFO

Keywords:
Cas protein
CRISPR-Cas
ESM-2
MSCNN
Deep learning

ABSTRACT

Cas proteins (CRISPR-associated protein) are the core components of the CRISPR-Cas system, playing critical roles in defending against foreign DNA and RNA invasions. Identifying Cas proteins can provide deeper insights into the immune mechanisms of the CRISPR-Cas system and help uncover the functional mechanisms of Cas proteins. In this study, we developed a computational tool named CasPro-ESM2, which combines the Pre-trained Protein Language Model ESM-2, multi-scale convolutional neural networks, and evolutionary information from protein sequences to identify Cas proteins. Experimental results demonstrate that CasPro-ESM2 outperforms existing models in Cas protein identification, achieving the highest values in metrics such as ACC, SP, SN, and MCC on two different datasets. Furthermore, we deployed this tool on a web server to enable direct access for users (http://www.bioai-lab.com/CasProESM-2).

1. Introduction

The CRISPR-Cas system (Clustered Regularly Interspaced Short Palindromic Repeats and its associated Cas proteins) is an RNA-guided immune mechanism developed by bacteria and archaea to defend against the invasion of exogenous DNA and RNA [1]. This system comprises short repetitive sequences in DNA (CRISPR sequences) and functionally diverse Cas proteins. Spacer sequences within the CRISPR loci store genetic fragments of previously invading viruses or plasmids, enabling the host to rapidly respond to similar invasions in the future by leveraging this “memory.” Utilizing this stored information, the CRISPR-Cas system precisely identifies and eliminates foreign genes, thus maintaining genomic stability within the cell [2].

The core of the CRISPR-Cas system is the Cas proteins, which perform diverse functions required to recognize and degrade foreign DNA or RNA. The immune process of the CRISPR-Cas system is generally divided into three stages: adaptation, expression, and interference [3]. In the adaptation stage, Cas proteins recognize exogenous DNA and integrate its fragments into CRISPR sequences. Subsequently, these sequences are transcribed into crRNA (CRISPR RNA) and form complexes with Cas proteins. Finally, during the interference stage, the crRNA collaborates with Cas proteins to specifically target and cleave foreign genes [4,5]. As shown in Fig. 1, the CRISPR-Cas system is primarily classified into Class 1 and Class 2 based on the composition and function of effector proteins. The key distinction between these two classes lies in whether the effector function is carried out by a multi-protein complex or a single protein [6]. Further research has led to a more detailed subdivision of these two classes into multiple subtypes. Class 1 CRISPR-Cas systems include Type I, Type III, Type IV, and the recently identified Type VII, whereas Class 2 CRISPR-Cas systems consist of Type II, Type V, and Type VI [7]. Class 1 CRISPR-Cas systems are the most prevalent in nature, requiring multiple Cas proteins to work together, forming a

[*] Corresponding author.
E-mail address: feifeicui@hainanu.edu.cn (F. Cui).

https://doi.org/10.1016/j.ijbiomac.2025.142309
Received 23 December 2024; Received in revised form 15 March 2025; Accepted 18 March 2025
Available online 22 March 2025

complex multi-protein effector complex to carry out CRISPR immune functions. For example, Cas8/Cas10 is responsible for target DNA binding and interference, while Cas3, a hallmark Cas protein of Class 1, functions as an exonuclease that degrades invading DNA. Cas7 provides structural support within the CRISPR complex and plays a role in crRNA processing and target interference. Additionally, Cas5 and Cas6 contribute to the assembly of the CRISPR complex and are involved in crRNA processing, ensuring the system's stable operation. During the CRISPR adaptation stage, Cas1 and Cas2 mediate the insertion of new spacer sequences, while Cas4 is involved in the preprocessing of these spacers. In contrast, Class 2 CRISPR-Cas systems are widely used in gene editing and detection applications due to their simpler structure, which requires only a single Cas protein and its corresponding RNA [8]. Cas proteins exhibit significant diversity, with each protein possessing unique characteristics and evolutionary origins. Since the Cas proteins in Class 2 can independently perform all effector functions, they are widely used in gene editing and molecular detection, for example, Cas9, one of the most well-known Cas proteins, is a nuclease extensively utilized in gene editing, achieving precise DNA cleavage by recognizing specific PAM (Protospacer Adjacent Motif) sequences [9]. Cas12, in addition to cleaving double-stranded DNA, demonstrates the ability to cleave non-targeted single-stranded DNA, showcasing its potential in molecular diagnostics [10]. Cas13 primarily targets RNA and has gained wide application in virus detection in recent years, including for COVID-19 testing [11]. As research progresses, CRISPR-Cas system technologies have found broad applications in gene editing, gene regulation, and pathogen detection. For instance, Cas9 is used not only for gene correction in human hereditary diseases but also for editing crops like rice and wheat to enhance yield and disease resistance [12,13]. Similarly, Cas12 has achieved groundbreaking results in plant gene editing. Cas13, due to its RNA-targeting specificity, is widely employed in virus detection, covering SARS-CoV-2 as well as early detection of other infectious diseases [14].

Despite significant advancements in understanding the functionality of Cas proteins, the diversity and functional mechanisms of the Cas family still require further exploration. Identifying and classifying different types of Cas proteins is critical for advancing novel CRISPR-Cas system applications and developing more precise and efficient gene-editing tools. Traditional methods for identifying Cas proteins primarily rely on sequence alignment-based technologies, such as BLAST and HMMER [15,16]. These tools depend heavily on known sequence databases, identifying Cas proteins by detecting similar sequences or conserved regions. For instance, HMMCAS [17] is a tool that identifies Cas proteins based on the HMMER method. While these methods perform well in identifying Cas proteins closely related to known sequences, their efficacy declines when dealing with distantly related Cas proteins, especially in poorly studied bacteria and archaea. This limitation arises due to the lack of sequences in existing databases resembling these Cas proteins, making it challenging to capture their functional characteristics [1].

Given the reliance of sequence homology-based methods on the quality and breadth of existing databases, recent years have seen the emergence of computational methods leveraging machine learning to identify Cas proteins [18]. CASPredict [19], developed by the Yang team, uses support vector machines (SVMs) with optimal dipeptide composition as features for predicting Cas proteins [19]. CRISPRCas-Stack [20], developed by the Zhang team, employs features such as sequence information and position-specific scoring matrix (PSSM) data, combined with the sequential forward search (SFS) method to select features for identifying Cas proteins. Currently, most Cas protein identification methods, apart from sequence homology-based approaches, rely on traditional machine learning techniques. However, the application of deep learning in recent years has revolutionized bioinformatics due to its superior performance in handling complex nonlinear data. Tasks like protein structure prediction, functional annotation, and sequence classification have benefited significantly from deep learning. For instance, convolutional neural networks (CNNs) and long short-term memory (LSTM) models have been applied to protein family classification and sequence pattern detection. These approaches automatically learn sequence features, overcoming the limitations of traditional methods that rely on manually designed features. For example, Zhou's team used CNNs for protein structure prediction [21], while Alakus's team applied LSTM models to predict protein-protein interactions [22].

Moreover, the success of large language models (LLMs) in natural language processing (NLP) has inspired researchers to introduce them into bioinformatics, particularly in protein sequence analysis [23,24]. Models such as BERT and GPT, based on the Transformer architecture, were initially developed for NLP tasks. Their powerful context-processing and sequence dependency capabilities make them well-suited for analyzing biological sequences [25]. In recent years, specialized LLMs for protein sequences, such as the ESM (Evolutionary Scale Modeling) series, have emerged. These models learn from millions of protein sequences, capturing evolutionary and functional features. ESM-1b, ESM-2, and similar models [26–28], based on the Transformer architecture, process large volumes of protein sequence data using deep neural networks, demonstrating exceptional capabilities in protein function prediction. Their advantage lies in the ability to uncover implicit sequence patterns and features through self-supervised learning without requiring labeled data. Compared to traditional machine learning models, these methods effectively handle vast amounts of sequence information and achieve higher accuracy when analyzing distantly related homologs.

Building on the rapid development of deep learning technologies and LLMs, we propose a new method for identifying Cas proteins. This method combines the powerful sequence representation capabilities of the ESM-2 protein language model with functional information derived from PSSM-based AATP features (Amino Acid Type Profile) [29]. Using a deep learning framework, the method identifies Cas proteins by

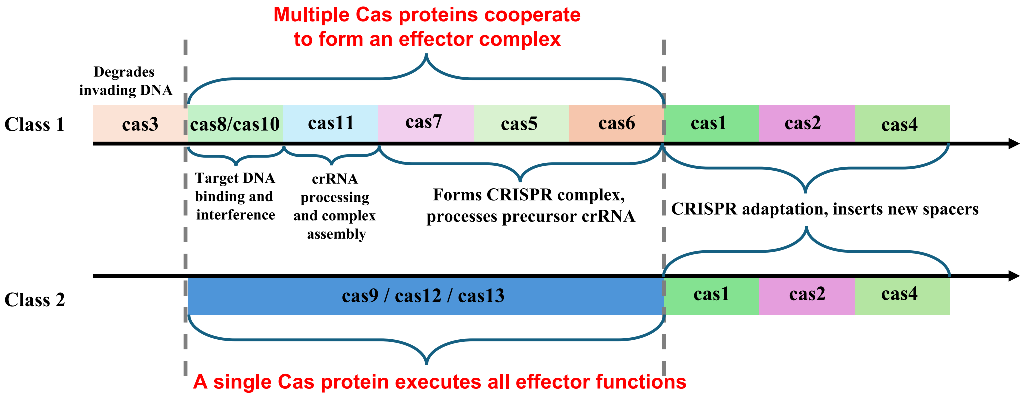
Multiple Cas proteins cooperate to form an effector complex
Degrades invading DNA
Class 1
cas3
cas8/cas10
cas11
cas7
cas5
cas6
cas1
cas2
cas4
Target DNA binding and interference
crRNA processing and complex assembly
Forms CRISPR complex, processes precursor crRNA
CRISPR adaptation, inserts new spacers
Class 2
cas9 / cas12 / cas13
cas1
cas2
cas4
A single Cas protein executes all effector functions

Fig. 1. Classification of CRISPR-Cas systems and related Cas proteins.

capturing deep sequence characteristics that traditional methods struggle to detect. Experimental results demonstrate that the ESM-2-based model achieves superior accuracy and sensitivity, especially when analyzing distantly related homologs, providing new directions for developing more precise and efficient tools for CRISPR-Cas system applications.

2. Methods and materials

2.1. Dataset

In this study, we utilized two high-quality datasets to train and evaluate our Cas protein classification model. These datasets were derived from CRISPRCasStack [20] and CASPredict [19], respectively. Both datasets were constructed based on the UniProt database.

For the CRISPRCasStack dataset, a total of 209 Cas protein sequences were collected. The negative dataset was composed of manually curated non-Cas proteins. To ensure that sequence similarity was below 70 % and to exclude sequences containing non-standard amino acids, the CD-HIT tool was applied with a 70 % similarity threshold for clustering. The final dataset consisted of 150 positive and 150 negative samples for the training set and 59 positive and 59 negative samples for the test set.

For the CASPredict dataset, an initial collection of 293 Cas protein sequences was extracted from the UniProt database. These sequences were manually inspected to remove those containing ambiguous residues (e.g., "X," "B," "Z"). To reduce sequence redundancy, CD-HIT was used with a 30 % similarity threshold, resulting in 155 Cas protein sequences retained as positive samples. The negative samples were carefully selected from manually reviewed non-Cas proteins in the UniProt database. CD-HIT was subsequently applied to remove sequences with homology >40 % in the initial negative dataset, ensuring sequence similarity below 40 % and matching the length distributions. The final dataset included 155 positive (Cas proteins) and 155 negative (non-Cas proteins) samples for the training set, with an additional 64 positive and 64 negative samples for the test set. For these two different datasets, we mainly use Dataset1 to determine our model hyperparameters and use both Dataset1 and Dataset2 to evaluate our model under different training conditions. Table 1 shows the information of the two datasets in detail.

2.2. Workflow architecture of CasPro-ESM2

To accurately identify Cas proteins, we designed a new web-based tool—CasPro-ESM2. The model's construction and protein analysis workflow are illustrated in Fig. 2. As shown in Fig. 2A, the raw amino acid sequence of the protein to be analyzed enters the model and flows through two distinct channels to extract the embedded information. One channel extracts features based on the AATP (Amino Acid Type Profile) derived from the Position-Specific Scoring Matrix (PSSM). The other channel processes information using the ESM-2 MSCNN Extractor module. These two refined feature sets are subsequently fused, forming the foundation for predicting Cas proteins. Fig. 2B depicts the ESM-2 MSCNN Extractor module, which uses multi-scale convolutional neural networks to process the information obtained after the protein sequence is input into the ESM-2 model. After designing the prediction model, we integrated a visualization module to evaluate which components of the CasPro-ESM2 model contribute to Cas protein predictions and to validate the interpretability of our model, as illustrated in Fig. 2C. Finally, the CasPro-ESM2 model was deployed on a server, creating an online tool for Cas protein identification. This tool enables users to perform protein sequence predictions conveniently through the web interface.

Table 1
Datasets used in this study.

	Training data		Testing data	
Dataset 1	Negative **150**	Positive **150**	Negative **59**	Positive **59**
Dataset 2	Negative **155**	Positive **155**	Negative **64**	Positive **64**

2.3. ESM-2 MSCNN extractor

In recent years, a significant breakthrough in natural language processing (NLP) has been the development and application of large language models (LLMs). In 2017, Vaswani et al. introduced the Transformer model, which became the cornerstone of LLMs [30]. The self-attention mechanism in the Transformer architecture eliminates the sequential processing limitations of recurrent neural networks (RNNs), allowing parallel sequence processing and excelling at capturing long-range dependencies [31]. This innovation greatly improved the efficiency and performance of models in handling textual data, laying the foundation for large-scale pre-trained language models. Subsequently, numerous pre-trained models based on the Transformer architecture emerged. In 2018, OpenAI introduced GPT (Generative Pretrained Transformer) [32], which used the encoder portion of the Transformer for pre-training and was applied to various downstream generative tasks such as text generation, translation, and summarization. In 2019, Google introduced BERT (Bidirectional Encoder Representations from Transformers) [33], a bidirectional Transformer-based model that achieved remarkable performance in multiple NLP tasks, further propelling the development of LLMs.

In bioinformatics, protein sequences share structural similarities with textual data, making language model-based representation methods directly applicable to protein sequences. Pre-trained models can learn semantic information from large-scale protein sequences, generating rich protein embeddings. Notable examples include the ESM (Evolutionary Scale Modeling) series and ProteinBERT [34]. These embeddings effectively capture both local and global features of amino acids, enabling downstream tasks like classification and functional prediction. To efficiently identify Cas proteins, this study adopted the flexible and accurate ESM-2 [35] LLM to generate embeddings for protein sequences, combining these with multi-scale convolutional neural networks (MSCNNs) for further information extraction and learning. ESM-2, developed under Meta AI's deep learning framework, is a type of protein language model. As a pre-trained Transformer model trained on extensive protein sequence data, ESM-2 employs unsupervised learning to capture contextual dependencies among amino acids, extracting semantic features from protein sequences. Unlike traditional sequence alignment methods, ESM-2 does not rely on evolutionary information but infers structural and functional insights directly from the sequence itself. ESM-2 builds upon its predecessors, ESM-1 (2019) and ESM-1b (2020) [26], offering a better balance between performance and computational efficiency. It excels at handling longer sequences and more complex tasks, demonstrating superior performance across different applications. Compared to traditional multiple sequence alignment (MSA)-based protein feature extraction methods [36], ESM-2 exhibits stronger generalization capabilities and higher computational efficiency.

MSCNNs [37], a variant of convolutional neural networks (CNNs), use filters of different scales to capture multi-level feature information from input data. By accommodating a diverse range of receptive fields, MSCNNs excel at identifying feature patterns across varying scales. This capability makes MSCNNs particularly effective for processing long-sequence data, image analysis, and complex biological data that require simultaneous recognition of both local and global patterns. The ESM-2 model adeptly captures long-range relationships between amino acids. By combining MSCNNs with ESM-2, we further investigate the important patterns within the high-dimensional features provided by ESM-2, refining the information from global to granular details. Specifically, protein sequences encoded by ESM-2 produce 320-dimensional

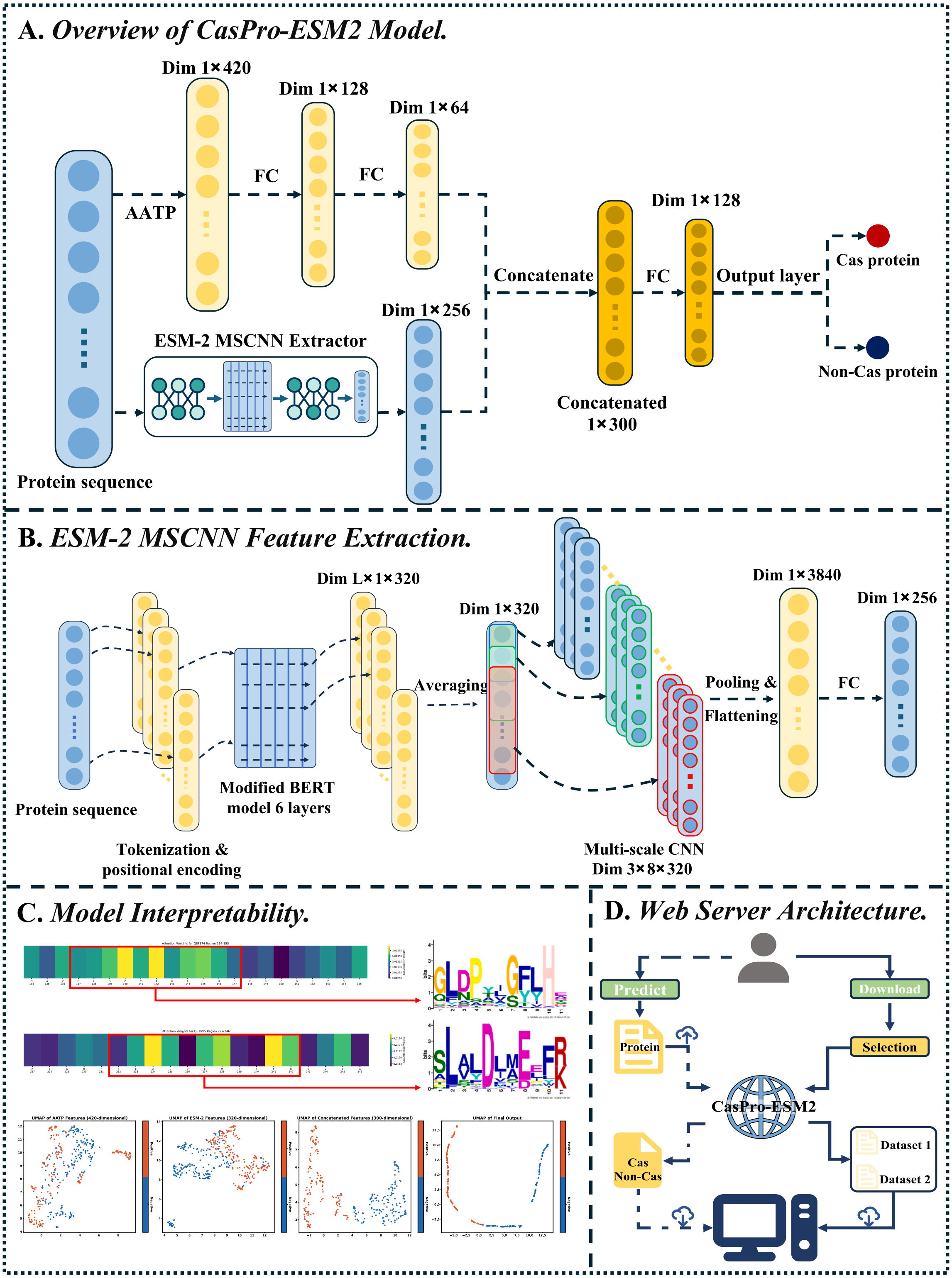
A. Overview of CasPro-ESM2 Model.
Dim 1×420
Dim 1×128
Dim 1×64
FC
FC
AATP
Protein sequence
ESM-2 MSCNN Extractor
Dim 1×256
Concatenate
Dim 1×128
FC
Output layer
Cas protein
Non-Cas protein
Concatenated 1×300
B. ESM-2 MSCNN Feature Extraction.
Dim L×1×320
Dim 1×320
Dim 1×3840
Dim 1×256
Protein sequence
Tokenization & positional encoding
Modified BERT model 6 layers
Averaging
Pooling & Flattening
FC
Multi-scale CNN Dim 3×8×320
C. Model Interpretability.
UMAP of AATP Features (420-dimensional)
UMAP of ESM-2 Features (320-dimensional)
UMAP of Concatenated Features (300-dimensional)
UMAP of Final Output
D. Web Server Architecture.
Predict
Download
Protein
Selection
CasPro-ESM2
Cas Non-Cas
Dataset 1
Dataset 2

Fig. 2. (A) Overall Structure of the CasPro-ESM2 Model. This module consists of AATP and the ESM-2 MSCNN components, which extract high-efficiency features from protein sequences. These features are combined into high-dimensional feature representations that enable the accurate identification of Cas proteins. (B) Details of the ESM-2 MSCNN Extractor Module. The features derived from the ESM-2 model, based on a six-layer BERT architecture, undergo pooling, flattening, and other processes. These operations enhance the model's ability to perceive multi-scale information within protein sequences. (C) Model Interpretability Analysis UMAP (Uniform Manifold Approximation and Projection) dimensionality reduction is used to visualize extracted features. Combined with attention mechanism heatmaps, this approach highlights the importance of specific amino acids in identifying Cas proteins, thereby enhancing the interpretability of the model. (D) Web Server Architecture Based on CasPro-ESM2. The application workflow of the CasPro-ESM2 model on the web server includes functionalities such as protein sequence prediction and dataset downloads, providing a convenient platform for researchers to utilize the model.

feature vectors, which are then processed through convolutional operations in MSCNNs. The MSCNN module uses three different kernel sizes to capture multi-level sequence information, ultimately producing 256-dimensional feature vectors. These features capture not only the local information of amino acids within the sequence but also the global structural characteristics, providing rich inputs for downstream multi-scale convolutional neural networks. This combination of ESM-2 and MSCNNs allows for more robust and nuanced analysis of protein sequences, offering a significant advantage in identifying Cas proteins with high accuracy and precision.

2.4. AATP features

Position-Specific Scoring Matrix (PSSM) is a matrix representing the evolutionary information of protein sequences [38]. It describes the probability of each amino acid appearing at different positions in a protein sequence, calculated through alignments with a set of similar protein sequences. PSSM captures the conservation and evolutionary information of different positions within the sequence [39]. Various feature representations can be derived from PSSM through different computational methods, such as AATP, Pse-PSSM, CTDT, and Tri-gram-PSSM. In this study, we utilized AATP [29] features as auxiliary features to assist CasPro-ESM2 in learning more diverse information. The core concept of AATP lies in analyzing and summarizing the values within the PSSM matrix to extract preference information of different types of amino acids in the sequence. By statistically capturing the tendency of specific amino acids to appear, AATP reduces data redundancy while retaining critical sequence information related to amino acid preferences. Furthermore, the calculation of AATP enables the transformation of the high-dimensional PSSM matrix into a fixed-length vector, thereby improving computational efficiency and making it suitable for large-scale protein data analysis. Previous experimental comparisons have also demonstrated [20] that, for the representation of Cas protein sequences, the AATP representation method based on PSSM is more effective.

The calculation of AATP features consists of two main feature vectors: Amino Acid Composition (AAC) [40] and Transition Probability Composition (TPC) [41], specifically as follows:

(1) AAC describes the average propensity of each amino acid across the entire protein sequence. It can be calculated as the mean value of each column in the PSSM.

$$x_j = \left(\frac{1}{L}\right)\sum_{i=1}^{L} P_{i,j}, \text{for} j = 1, 2, \dots, 20 \tag{1}$$

where L represents the length of the protein, and $P_{i,j}$ denotes the score in the PSSM matrix at position i corresponding to the j amino acid. These 20 values constitute a 20-dimensional vector $\mathbf{C} = (x_1, x_2, \dots, x_{20})^T$, which represents the AAC feature.

(2) TPC (Transition Probability Composition) reflects the probability of one amino acid transitioning to another within the sequence, resulting in a 400-dimensional vector. The specific calculation method is as follows:

$$X_{ij} = \left(\sum_{k=1}^{L-1}\left(P_{k,i} \times P_{k+1,j}\right)\right) \div \left(\sum_{j=1}^{20}\sum_{k=1}^{L-1}\left(P_{k,j} \times P_{k+1,j}\right)\right), \text{for} 1 \le i,j \le 20 \tag{2}$$

where $P_{k,i}$ and $P_{k+1,j}$ represent the scores of the i amino acid at position k and the j amino acid at position $k+1$ in the PSSM, respectively.

Finally, the AATP features are formed by combining the two components described above, resulting in a total of 20 (AAC) + 400 (TPC) = 420-dimensional features for each protein sequence.

2.5. Performance evaluation

To evaluate the performance of our model on Cas protein datasets, we adopted the same four evaluation metrics used in previous studies: ACC (Accuracy), SP (Specificity), SN (Sensitivity), and MCC (Matthews Correlation Coefficient). These metrics are calculated based on the values of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) [24,42,43]. The specific calculation formulas are as follows:

$$SN = \frac{TP}{TP + FN} \tag{3}$$

$$SP = \frac{TN}{TP + FN} \tag{4}$$

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

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

Among these four evaluation metrics, ACC, SP, and SN have values ranging from 0 to 1, representing the proportion of correctly predicted samples among all samples, the model's ability to identify negative samples, and the model's ability to identify positive samples, respectively. Higher values for these three metrics indicate better performance in the aspects they measure. MCC is a comprehensive metric that can better evaluate the classification performance of the model in cases of imbalanced datasets. MCC values range from −1 to 1, where 1 indicates perfect classification, 0 indicates random predictions, and − 1 indicates completely incorrect classification.

3. Results

3.1. Prediction performance of CasPro-ESM2

To ensure more robust evaluation, we applied a five-fold cross-validation method on Training Dataset 1. The data was divided into five subsets, and the model was trained and validated on these subsets in turn. This approach allowed for a more comprehensive evaluation of the model's performance while reducing the risk of overfitting due to limited data. Specifically, the training set of each dataset was divided into five folds. For each fold of training, four folds were used as the training set, while the remaining fold was used as the validation set. The validation set was utilized to evaluate the results of the current training. Finally, the performance of the validation sets from all folds was aggregated and averaged to produce the final training results on the dataset. Subsequently, the complete training data was used to train the model. The results are shown in Table 2.

3.2. Comparison with existing methods

Cas proteins are the core components of the CRISPR-Cas system, and numerous methods for Cas protein identification have been proposed. To fairly highlight the superiority of our approach, we conducted training on two separate datasets to avoid overlapping between training and testing sets across different datasets, which could lead to inaccurate evaluations. The detailed results are shown in Fig. 3. Fig. 3A presents the performance on the test set of Dataset 1, where the CasPro-ESM2 model

Table 2
Performance of the CasPro-ESM2 model on Dataset 1.

	ACC	SP	SN	MCC	AUC
Cross-validation	0.9400	0.9071	0.9749	0.8829	0.9779
Independent test set	0.9576	0.9153	1.0000	0.9186	0.9736

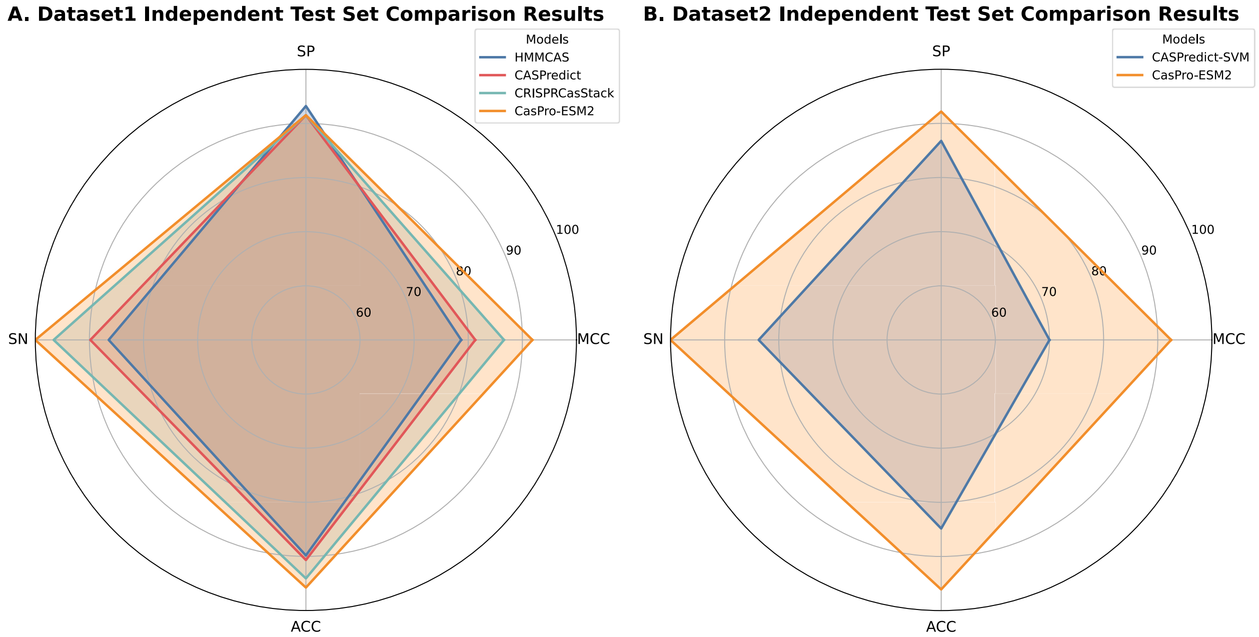

Fig. 3. Comparison of CasPro-ESM2 with other models. (A) Radar chart comparing CasPro-ESM2 with three other models on the independent test set of Dataset 1. (B) Radar chart comparing CasPro-ESM2 with the CASPredict-SVM model on the independent test set of Dataset 2.

achieves optimal results across all evaluation metrics. Specifically, its specificity (SP) and Matthews correlation coefficient (MCC) outperform the CRISPRCasStack method by 3.49 % and 5.26 %, respectively. Notably, the model demonstrates an exceptional ability to recognize non-Cas proteins in Dataset 1, achieving a specificity of 100 %.

Similarly, as shown in Fig. 3B, the CasPro-ESM2 model also exhibits excellent performance on Dataset 2. Compared to the CASPredict model, it achieves improvements in accuracy (ACC), specificity (SP), sensitivity (SN), and MCC by 11.25 %, 5.42 %, 16.29 %, and 22.47 %, respectively. These results demonstrate that our model has a strong capability for predicting Cas proteins and performs well across different datasets, highlighting its robustness and generalization ability.

3.3. Evaluating model performance on imbalanced data

To evaluate the model's performance under imbalanced data distribution in real-world scenarios, we collected additional non-Cas protein samples to simulate naturally imbalanced samples. Since no established research indicates the precise ratio of Cas to non-Cas proteins in nature, we selected the archaeon *Methanocaldococcus jannaschii*, which contains a typical CRISPR-Cas system, as the research subject. By analyzing the genomic data of this archaeon, we calculated the proportion of Cas proteins in its proteome, providing a reference for estimating the ratio of Cas to non-Cas proteins in nature [44]. The genome of this archaeon has been extensively studied and annotated, and its CRISPR-Cas system is well-documented, making it a suitable model for calculating the proportion of Cas and non-Cas proteins. Specifically, we downloaded the complete genome data of *Methanocaldococcus jannaschii* from the NCBI Genome database, and calculated the ratio by counting the number of Cas proteins and the total number of proteins in the genome. The genome of *Methanocaldococcus jannaschii* consists of 1812 proteins, of which 21 are Cas-related proteins. Based on this data, we estimated the ratio of Cas to non-Cas proteins to be approximately 1:86.

Subsequently, we selected reviewed non-Cas protein entries from bacteria or archaea in the UniProt database. The search results from UniProt provide two types of data: UniProtKB/Swiss-Prot (Reviewed) and UniProtKB/TrEMBL (Unreviewed). The UniProtKB/Swiss-Prot (Reviewed) entries are manually curated, having undergone expert validation based on experimental data, literature, and bioinformatics analysis. These entries are of high quality, with detailed annotations on function, structure, interactions, and more. On the other hand, UniProtKB/TrEMBL (Unreviewed) entries are automatically annotated, without human curation, and are of larger volume but relatively lower accuracy. To ensure the accuracy of this study, we applied the search condition "(taxonomy_id:2) OR (taxonomy_id:2157) NOT (keyword:cas) NOT (keyword:KW-1257)" for selecting non-Cas protein entries as negative controls. Here, 'taxonomy_id:2' and 'taxonomy_id:2157' refer to searches within bacteria and archaea, and 'KW-1257' represents CRISPR-Cas system-related entries. We then processed these entries following the same procedure as the original balanced dataset, removing sequences containing ambiguous residues (such as "X," "B," and "Z"). Afterward, we used the CD-HIT tool with a 30 % similarity threshold for clustering, ultimately constructing an imbalanced dataset with a positive-to-negative sample ratio of 1:86, as shown in Table 3.

We trained the model using the imbalanced training set and then evaluated its robustness and generalization ability on the imbalanced test set. Keeping the model architecture unchanged, we adjusted the "weight" parameter in the cross-entropy loss function due to the severe class imbalance. Specifically, we set the "weights = [0.2, 86]" to ensure that the loss computation included a weight factor, making the model more sensitive to the minority class. The final test results of the model on the imbalanced dataset are shown in Table 4.

From the final results, we can observe that although the positive samples used for training accounted for only 1.15 % of the total data samples, our model was still able to effectively capture the characteristics of Cas protein samples. These results demonstrate that the proposed model can efficiently distinguish between Cas and non-Cas proteins even under severe class imbalance. This proves that the model not only performs well on balanced datasets for standard tasks but also maintains high performance when dealing with highly imbalanced real-world data, providing strong support for efficient prediction in practical

Table 3
Imbalanced dataset samples.

	Training data		Testing data	
Imbalanced dataset	Negative **12,900**	Positive **150**	Negative **5074**	Positive **59**

Table 4
Model performance under imbalanced data distribution.

	ACC	SP	SN	MCC	AUC
Imbalanced data	0.9686	0.9696	0.8814	0.4662	0.9874

applications.

3.4. Ablation study

To better understand the contribution of each module to the overall performance of the model, we designed a series of ablation experiments. Specifically, we started with the complete CasPro-ESM2 model and progressively removed individual components to observe the impact of these changes on model performance. The experimental design includes the following model variants:

(1) AATP-only model: Using only AATP features while removing ESM-2 features and the multi-scale convolutional module.
(2) AATP+ESM-2 model: Retaining AATP and ESM-2 features but removing the multi-scale convolutional module.
(3) ESM-2+MSCNN model: Retaining ESM-2 features and the multi-scale convolutional module while removing AATP features.

By comparing the performance of these model variants, we aim to gain deeper insights into the role of each module in enhancing the overall performance of the CasPro-ESM2 model.

As shown in the Fig. 4, the complete CasPro-ESM2 model achieved the highest accuracy on the training set. On the independent test set, the complete model outperformed all ablation variants across all evaluation metrics, demonstrating significant superiority. Specifically, when the multi-scale convolutional module was removed, there was a noticeable drop in specificity and MCC, indicating the critical role of this module in capturing features across multiple scales. Similarly, removing the AATP features also led to a decline in MCC, highlighting the importance of AATP features in enhancing the classification accuracy of the model. These results underscore the significant contributions of both the multi-scale convolutional module and the AATP features to the model's performance. The complete CasPro-ESM2 model demonstrated strong capabilities in feature extraction and complementarity, suggesting that retaining the full model architecture is essential for achieving optimal performance in practical applications.

3.5. Optimization of MSCNN window sizes

When extracting embedding information from the ESM-2 model, we designed the ESM-2 MSCNN Extractor module. This module applies a multi-scale convolutional neural network (MSCNN) to the output layer of the ESM-2 model to further extract information. The choice of window size significantly affects feature extraction and model performance.

Smaller window sizes (e.g., 3, 5) are adept at capturing short-range local patterns or features, such as local amino acid sequence motifs. However, due to their limited scope, features extracted with smaller windows are more sensitive to noise in the input. In contrast, larger windows (e.g., 11, 13) are capable of capturing global patterns over a broader sequence range, making them suitable for identifying long-range features or motifs. However, larger windows increase computational complexity and may lead to higher model complexity, particularly for longer sequences. Moderately-sized windows (e.g., 7, 9) strike a balance between capturing local and global information, enabling the extraction of medium-length patterns in sequences. To achieve optimal prediction performance, we experimented with different window combinations. As illustrated in the Table 5, we compared the performance of four different window combinations on the training set of Dataset 1 using five-fold cross-validation, with accuracy (ACC) as the evaluation metric. Based on these results, we determined that the MSCNN module in the ESM-2 MSCNN Extractor should use window sizes of 9, 11, and 13.

Besides the four window combinations listed above, we tested several other window size combinations to evaluate the feasibility of the [9,11,13] configuration. Specifically, we tested ten other commonly

Table 5
Performance comparison of different convolution window combinations in the MSCNN module of the ESM-2 MSCNN Extractor.

	Convolution scales	ACC
1	(3, 5, 7)	0.933
2	(5, 7, 9)	0.926
3	(7, 9, 11)	0.930
4	(9, 11, 13)	0.943

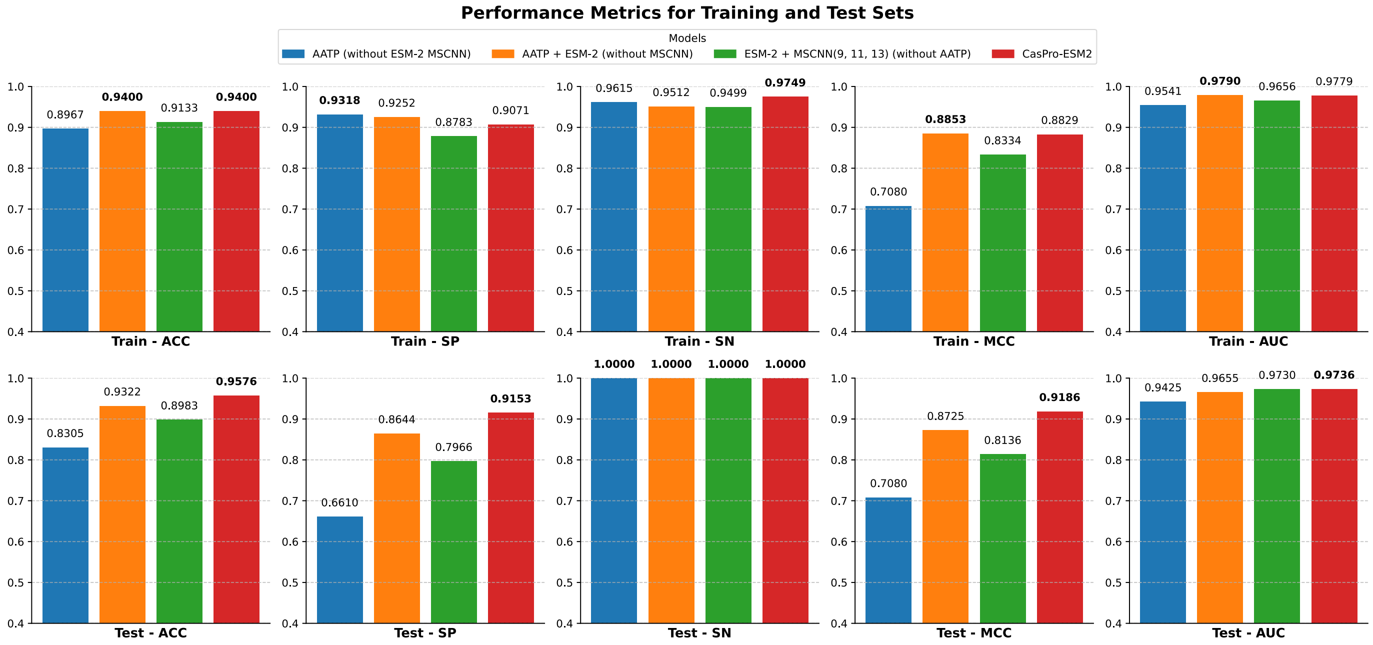

Fig. 4. Ablation study results on Dataset 1.

used window size combinations and found that the [9,11,13] configuration achieved the best overall performance. The detailed results are available at CasPro-ESM2/Result/results.csv at main · ChaoruiYan019/CasPro-ESM2 (github.com).

3.6. UMAP visualization analysis

As shown in Fig. 5, we performed UMAP dimensionality reduction and visualization on features extracted at different stages of the CasPro-ESM2 model. In the six subplots, orange represents the UMAP-space distribution of positive samples, while blue represents that of negative samples. Fig. 5A and D respectively display the distribution characteristics of AATP features before and after passing through a fully connected layer. Fig. 5B and E respectively show the distribution characteristics of ESM-2 features before and after MSCNN processing. Through pairwise comparisons, we can observe that after the initial learning by our designed modules, both AATP and ESM-2 features exhibit enhanced discriminative ability for separating positive and negative samples. This demonstrates the extraction capability of the AATP module and the learning and extraction capability of the ESM-2 MSCNN Extractor module within the CasPro-ESM2 model. Fig. 5C presents the UMAP visualization of the 300-dimensional features obtained by concatenating AATP and ESM-2 features processed by MSCNN. Fig. 5F shows the UMAP visualization of the final output layer, where positive and negative samples are almost entirely separated in the UMAP space. This fully validates the model's strong discriminative capability after multi-level feature extraction and fusion. Through UMAP visualization of features from different layers, we can clearly observe the progressive enhancement in the model's ability to distinguish sample classes. This indicates the model's effectiveness in learning features at different levels and its capability to integrate features for improved classification performance.

3.7. Mapping attention weights to protein sequences

To demonstrate that our model identifies Cas proteins based on the biological information embedded in protein sequences, we visualized the attention mechanism within the ESM-2 MSCNN Extractor module. Specifically, we utilized the STREME [45] tool to identify motifs present in the Cas proteins from our dataset. Subsequently, we extracted attention values from the sixth BERT layer of the Transformer module within the ESM-2 MSCNN Extractor. These attention values correspond to each amino acid in a protein sequence, reflecting the importance assigned to each amino acid during model training. We visualized the attention values for amino acids corresponding to the motifs identified by the STREME tool. For sequences such as Q8F874, Q53VV5, Q2RW61, and Q2RY12, we plotted the attention values for amino acids in regions surrounding potential motifs, as shown in Fig. 6. In the Fig. 6, the left panel displays attention heatmaps, where the intensity of the color reflects the magnitude of the attention values for individual amino acids. The right panel shows the motifs identified by the STREME tool. Notably, the regions enclosed within red boxes indicate the amino acids forming specific motifs. Within these regions, the amino acids exhibit significantly higher attention values compared to neighboring amino acids that do not form motifs. This observation provides indirect evidence that our model leverages biological information, such as motifs, to effectively identify Cas proteins.

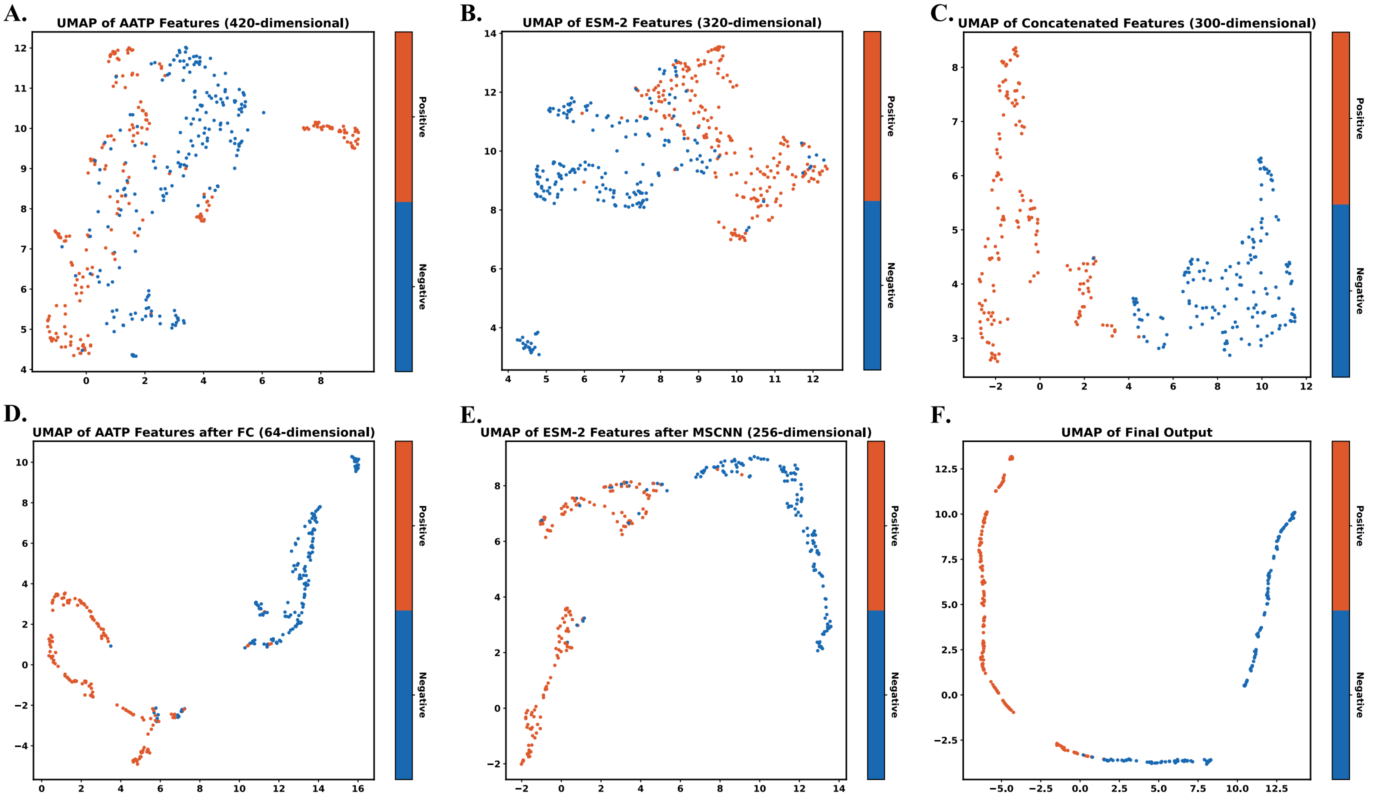

Fig. 5. UMAP Dimensionality Reduction Visualization Analysis. (A) Visualization of AATP features after UMAP dimensionality reduction. (B) Visualization of ESM-2 features after UMAP dimensionality reduction. (C) Visualization of concatenated 300-dimensional features (AATP and MSCNN-processed ESM-2 features) after UMAP dimensionality reduction. (D) Visualization of AATP features after passing through a fully connected layer and UMAP dimensionality reduction. (E) Visualization of ESM-2 features after MSCNN processing and UMAP dimensionality reduction. (F) Visualization of final output layer features after UMAP dimensionality reduction.

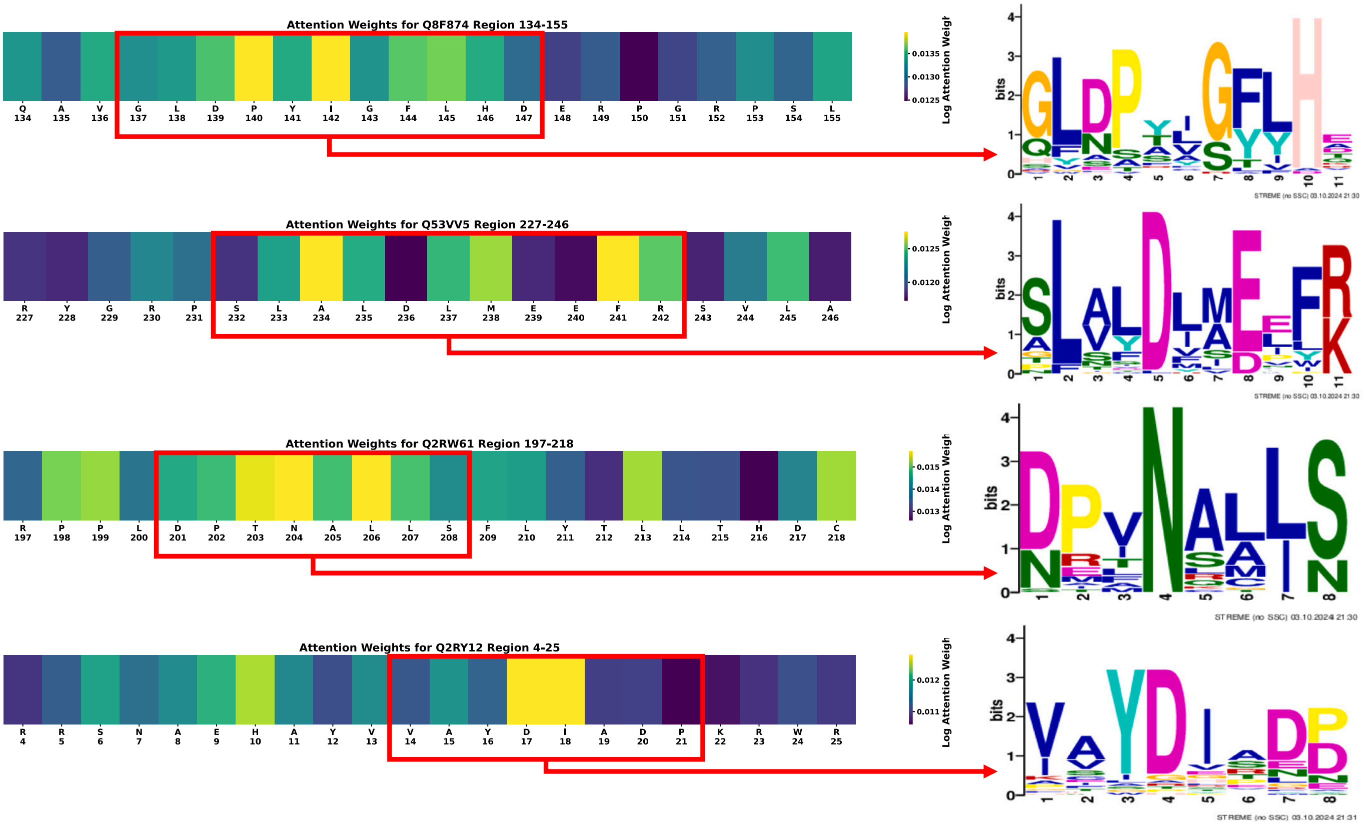

Fig. 6. Visualization of the attention weights mapped to motif regions in the CasPro-ESM2 model.

4. Webservers

To facilitate usage, we have deployed a web server for online implementation of the Cas protein identification model, which can be accessed via the provided website http://www.bioai-lab.com/CasProESM-2. For practical testing, we have also shared executable code on Google Colab, allowing users to directly reproduce the model from this manuscript online: https://drive.google.com/drive/folders/1NuEPezbY33iHP58d17lIb82T1lGXFiup?usp=drive_link. The website also includes a dataset download feature, allowing researchers to conveniently access the two datasets used in this study. For more information, please visit GitHub: https://github.com/ChaoruiYan019/CasPro-ESM2.

5. Discussion

In this study, we propose the CasPro-ESM2 model, which effectively distinguishes whether a given protein belongs to the Cas protein family. This model provides an efficient computational screening tool for CRISPR-related research, including CRISPR-based gene editing tool development, microbiology, evolutionary biology, and drug discovery [46] [47]. For instance, CasPro-ESM2 is primarily designed for large-scale Cas protein identification in extensive protein datasets, facilitating the rapid screening of potential CRISPR-Cas system proteins from uncharacterized sequences. Traditionally, Cas protein identification has relied on experimental validation methods, which can be labor-intensive and time-consuming. However, through genomic data analysis, CasPro-ESM2 enables researchers to automatically identify candidate Cas proteins, reducing the workload of experimental screening and improving the efficiency of discovering new CRISPR-Cas systems. In microbial genome annotation, CasPro-ESM2 can be utilized to identify Cas proteins in newly sequenced microorganisms, enhancing the accuracy of genome functional annotation. Additionally, it aids microbiologists in screening and classifying Cas proteins, facilitating research on the evolutionary patterns of Cas proteins across different species. Furthermore, CasPro-ESM2 has a significant impact in the field of drug discovery. Applications such as CRISPR-mediated gene editing, antimicrobial drug research, and Cas protein-based therapeutic strategies require specific Cas proteins [48]. The CasPro-ESM2 model accurately identifies these proteins, enabling researchers to efficiently screen experimentally suitable Cas proteins, thereby reducing experimental costs and improving research efficiency [49]. In antimicrobial drug development, CasPro-ESM2 can be used to identify Cas proteins within bacterial CRISPR immune systems, aiding in the study of their impact on antibiotic resistance mechanisms. For example, certain Cas proteins (such as Cas9 and Cas12a) have been found to regulate the expression of bacterial antibiotic resistance genes, influencing the evolution of resistant strains [50]. Through CasPro-ESM2, we can rapidly screen large-scale bacterial genome data to identify potential Cas proteins and validate their involvement in resistance regulation through experimental approaches, providing a theoretical foundation for the development of novel antimicrobial strategies. Additionally, in gene therapy research, different Cas proteins exhibit diverse behaviors across various cell types and genomic environments. CasPro-ESM2 can assist researchers in rapidly identifying potential Cas proteins from unannotated protein databases. For example, Cas13, due to its RNA target specificity, can be used for rapid disease diagnostics, enabling early treatment to save lives [51]. CasPro-ESM2 can predict which Cas proteins are more suitable for specific cellular or therapeutic environments, thereby accelerating the development of personalized gene-editing tools.

Experimental results on two independent datasets demonstrate that CasPro-ESM2 outperforms traditional sequence homology-based methods and machine learning-based identification approaches in terms of ACC, SP, SN, and MCC. Notably, it exhibits excellent performance in detecting distantly related homologs, highlighting the effectiveness of deep learning-based feature extraction and sequence representation in capturing functional information that traditional methods often overlook. The model incorporates a multi-scale convolutional neural network, leveraging convolutional kernels of varying sizes to capture both short-range and long-range sequence patterns, thereby enhancing the model's ability to understand Cas protein characteristics. Compared to traditional methods that rely solely on PSSM or handcrafted features, CasPro-ESM2 demonstrates superior performance in high-dimensional feature expression and structural information extraction. Additionally, the introduction of an attention mechanism enables the model to focus on key amino acid regions, providing interpretability support for Cas protein recognition and further enhancing result reliability. Moreover, experimental results indicate that CasPro-ESM2 maintains consistent performance across different datasets, reflecting its robustness.

However, our model still has some limitations. Specifically, while CasPro-ESM2 is highly effective in Cas protein classification, CRISPR-Cas systems can be categorized into various types and subtypes, and characterizing these system types and subtypes is equally important. Differences between CRISPR-Cas system types and subtypes significantly impact their adaptability and functionality in bacteria and archaea, as well as their potential applications in gene editing, biotechnology, and antiviral defense. However, CasPro-ESM2 is currently designed specifically for Cas protein identification and does not yet possess the functionality to classify CRISPR-Cas system types and subtypes.

6. Conclusion

In conclusion, this study proposes CasPro-ESM2, a deep learning-based model for Cas protein identification, which effectively integrates deep semantic features derived from ESM-2 with AATP features based on PSSM. By leveraging multi-scale convolutional neural networks and attention mechanisms, the model achieves superior performance in Cas protein identification across different datasets, outperforming traditional sequence-based and machine learning-based approaches. The results demonstrate its robustness and effectiveness, making it a valuable computational tool for CRISPR-related research, such as CRISPR gene editing tool development and drug discovery. Moreover, CasPro-ESM2 significantly reduces the workload of experimental validation and improves the efficiency of discovering new CRISPR-Cas systems. In the next phase of our research, the identification of different CRISPR-Cas system types and subtypes will also be a key focus. Moving forward, we plan to extend the CasPro-ESM2 framework by incorporating contextual information from neighboring Cas genes, aiming to develop a model capable of automatically classifying CRISPR-Cas system types and subtypes. This extension will provide a more comprehensive and precise tool for the panoramic analysis and application of CRISPR-Cas systems.

CRediT authorship contribution statement

Chaorui Yan: Writing – original draft, Visualization, Software, Project administration, Investigation, Formal analysis, Data curation, Conceptualization. **Zilong Zhang:** Writing – review & editing, Validation, Supervision, Resources, Methodology, Funding acquisition, Conceptualization. **Junlin Xu:** Writing – review & editing, Validation, Methodology, Data curation. **Yajie Meng:** Writing – review & editing, Validation, Methodology, Investigation. **Shankai Yan:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Leyi Wei:** Writing – review & editing, Software, Data curation, Conceptualization. **Quan Zou:** Validation, Supervision, Resources, Methodology, Conceptualization. **Qingchen Zhang:** Resources, Methodology, Investigation, Conceptualization. **Feifei Cui:** Writing – review & editing, Validation, Supervision, Resources, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used Generative AI tools ChatGPT in order to improve readability and language of the manuscript. The authors reviewed and verified all AI-assisted translations to ensure accuracy and consistency with the original content and take full responsibility for the content of the published manuscript.

Funding

The work is supported by the National Natural Science Foundation of China (No. 62450002), the Science and Technology special fund of Hainan Province ZDYF2024GXJS018 and the Innovation Platform for “New Star of South China Sea” of Hainan Province under Grant No. NHXXRCXM202306.

Declaration of competing interest

All authors declare that they have no conflicts of interest.

Data availability

Executable code is available on https://drive.google.com/drive/folders/1NuEPezbY33iHP58d17lIb82T1lGXFiup?usp=drive_link. More details on https://github.com/ChaoruiYan019/CasPro-ESM2.

References

[1] K.S. Makarova, D.H. Haft, R. Barrangou, S.J. Brouns, E. Charpentier, P. Horvath, S. Moineau, F.J. Mojica, Y.I. Wolf, A.F. Yakunin, Evolution and classification of the CRISPR–Cas systems, Nat. Rev. Microbiol. 9 (6) (2011) 467–477.

[2] R. Barrangou, C. Fremaux, H. Deveau, M. Richards, P. Boyaval, S. Moineau, D. A. Romero, P. Horvath, CRISPR provides acquired resistance against viruses in prokaryotes, Science 315 (5819) (2007) 1709–1712.

[3] J.K. Nuñez, A.S. Lee, A. Engelman, J.A. Doudna, Integrase-mediated spacer acquisition during CRISPR–Cas adaptive immunity, Nature 519 (7542) (2015) 193–198.

[4] L.A. Marraffini, E.J. Sontheimer, CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA, science 322 (5909) (2008) 1843–1845.

[5] M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J.A. Doudna, Charpentier E: a programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity, science 337 (6096) (2012) 816–821.

[6] K.S. Makarova, Y.I. Wolf, O.S. Alkhnbashi, F. Costa, S.A. Shah, S.J. Saunders, R. Barrangou, S.J.J. Brouns, E. Charpentier, D.H. Haft, et al., An updated evolutionary classification of CRISPR–Cas systems, Nat. Rev. Microbiol. 13 (11) (2015) 722–736.

[7] I. Yosef, M.G. Goren, U. Qimron, Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli, Nucleic Acids Res. 40 (12) (2012) 5569–5576.

[8] M. Yan, J. Li, The evolving CRISPR technology, Protein Cell 10 (11) (2019) 783–786.

[9] A. Le Rhun, A. Escalera-Maurer, M. Bratovič, E. Charpentier, CRISPR-Cas in streptococcus pyogenes, RNA Biol. 16 (4) (2019) 380–389.

[10] J.S. Chen, E. Ma, L.B. Harrington, M. Da Costa, X. Tian, J.M. Palefsky, J.A. Doudna, CRISPR-Cas12a target binding unleashes indiscriminate single-stranded DNase activity, Science 360 (6387) (2018) 436–439.

[11] O.O. Abudayyeh, J.S. Gootenberg, P. Essletzbichler, S. Han, J. Joung, J.J. Belanto, V. Verdine, D.B. Cox, M.J. Kellner, A. Regev, RNA targeting with CRISPR–Cas13, Nature 550 (7675) (2017) 280–284.

[12] X. Xiong, M. Chen, W.A. Lim, D. Zhao, L.S. Qi, CRISPR/Cas9 for human genome engineering and disease research, Annu. Rev. Genomics Hum. Genet. 17 (1) (2016) 131–154.

[13] S. Gupta, A. Kumar, R. Patel, V. Kumar, Genetically modified crop regulations: scope and opportunity using the CRISPR-Cas9 genome editing approach, Mol. Biol. Rep. 48 (5) (2021) 4851–4863.

[14] Y. Yang, D. Wang, P. Lü, S. Ma, K. Chen, Research progress on nucleic acid detection and genome editing of CRISPR/Cas12 system, Mol. Biol. Rep. 50 (4) (2023) 3723–3738.

[15] D.H. Haft, J. Selengut, E.F. Mongodin, K.E. Nelson, A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes, PLoS Comput. Biol. 1 (6) (2005) e60.

[16] R.D. Finn, J. Clements, S.R. Eddy, HMMER web server: interactive sequence similarity searching, Nucleic Acids Res. 39(suppl_2):W29-W37 (2011).

[17] G. Chai, M. Yu, L. Jiang, Y. Duan, J. Huang, HMMCAS: a web tool for the identification and domain annotations of Cas proteins, IEEE/ACM Trans. Comput. Biol. Bioinform. 16 (4) (2017) 1313–1315.

[18] C. Ao, S. Jiao, Y. Wang, L. Yu, Q. Zou, Biological sequence classification: a review on data and general methods, Research 2022 (2022) 0011.

[19] S. Yang, J. Huang, B. He, CASPredict: a web service for identifying Cas proteins, PeerJ 9 (2021) e11887.

[20] T. Zhang, Y. Jia, H. Li, D. Xu, J. Zhou, G. Wang, CRISPRCasStack: a stacking strategy-based ensemble learning framework for accurate identification of Cas proteins, Brief. Bioinform. 23(5):bbac335 (2022).

[21] Y. Zhou, K. Tan, X. Shen, Z. He, H. Zheng, A protein structure prediction approach leveraging transformer and CNN integration, in: 2024 7th International Conference on Advanced Algorithms and Control Engineering (ICAACE), IEEE, 2024, pp. 749–753.

[22] T.B. Alakus, I. Turkoglu, Prediction of protein-protein interactions with LSTM deep learning model, in: 2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), IEEE, 2019, pp. 1–5.

[23] Khan W, Leem S, See KB, Wong JK, Zhang S, Fang R: A Comprehensive Survey of Foundation Models in Medicine. arXiv preprint arXiv:240610729 2024.

[24] C. Xiao, Z. Zhou, J. She, J. Yin, F. Cui, Z. Zhang, PEL-PVP: application of plant vacuolar protein discriminator based on PEFT ESM-2 and bilayer LSTM in an unbalanced dataset, Int. J. Biol. Macromol. 277 (2024) 134317.

[25] Y. Li, X. Wei, Q. Yang, A. Xiong, X. Li, Q. Zou, F. Cui, Z. Zhang, msBERT-promoter: a multi-scale ensemble predictor based on BERT pre-trained model for the two-stage prediction of DNA promoters and their strengths, BMC Biol. 22 (1) (2024) 126.

[26] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C.L. Zitnick, J. Ma, Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences, Proc. Natl. Acad. Sci. 118 (15) (2021) e2016239118.

[27] J. Meier, R. Rao, R. Verkuil, J. Liu, T. Sercu, A. Rives, Language models enable zero-shot prediction of the effects of mutations on protein function, Adv. Neural Inf. Proces. Syst. 34 (2021) 29287–29303.

[28] Rao R, Liu J, Verkuil R, Meier J, Canny JF, Abbeel P, Sercu T, Rives A: MSA Transformer. bioRxiv 2021:2021.2002.2012.430858.

[29] S. Zhang, F. Ye, X. Yuan, Using principal component analysis and support vector machine to predict protein structural class for low-similarity sequences via PSSM, J. Biomol. Struct. Dyn. 29 (6) (2012) 1138–1146.

[30] A. Vaswani, Attention is all you need, Adv. Neural Inf. Proces. Syst 30 (2017) 6000–6010.

[31] J. Yuan, Z. Wang, Z. Pan, A. Li, Z. Zhang, F. Cui, DPNN-ac4C: a dual-path neural network with self-attention mechanism for identification of N4-acetylcytidine (ac4C) in mRNA, Bioinformatics 40(11):btae625 (2024).

[32] A. Radford, Improving Language Understanding by Generative Pre-Training, 2018.

[33] Kenton JDM-WC, Toutanova LK: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of naacL-HLT: 2019. Minneapolis, Minnesota: 2.

[34] N. Brandes, D. Ofer, Y. Peleg, N. Rappoport, M. Linial, ProteinBERT: a universal deep-learning model of protein sequence and function, Bioinformatics 38 (8) (2022) 2102–2110.

[35] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, Evolutionary-scale prediction of atomic-level protein structure with a language model, Science 379 (6637) (2023) 1123–1130.

[36] R.C. Edgar, S. Batzoglou, Multiple sequence alignment, Curr. Opin. Struct. Biol. 16 (3) (2006) 368–373.

[37] Cui Z, Chen W, Chen Y: Multi-scale convolutional neural networks for time series classification. arXiv preprint arXiv:160306995 2016.

[38] X. Fu, Y. Yuan, H. Qiu, H. Suo, Y. Song, A. Li, Y. Zhang, C. Xiao, Y. Li, L. Dou, et al., AGF-PPIS: a protein–protein interaction site predictor based on an attention mechanism and graph convolutional networks, Methods 222 (2024) 142–151.

[39] S. Ahmad, A. Sarai, PSSM-based prediction of DNA binding sites in proteins, BMC bioinformatics 6 (2005) 1–6.

[40] C. Yan, A. Geng, Z. Pan, Z. Zhang, F. Cui, MultiFeatVotPIP: a voting-based ensemble learning framework for predicting proinflammatory peptides, Brief. Bioinform. 25 (6) (2024).

[41] R. Muhammod, S. Ahmed, D. Md Farid, S. Shatabda, A. Sharma, A. Dehzangi, PyFeat: a Python-based effective feature generation tool for DNA, RNA and protein sequences. Bioinformatics 35 (19) (2019) 3831–3833.

[42] S. Jiao, X. Ye, T. Sakurai, Q. Zou, R. Liu, Integrated convolution and self-attention for improving peptide toxicity prediction, Bioinformatics 40 (5) (2024).

[43] Y. Wang, Y. Zhai, Y. Ding, Q. Zou, SBSM-pro: support bio-sequence machine for proteins, SCIENCE CHINA Inf. Sci. 67 (11) (2024) 212106.

[44] C.J. Bult, O. White, G.J. Olsen, L. Zhou, R.D. Fleischmann, G.G. Sutton, J.A. Blake, L.M. FitzGerald, R.A. Clayton, J.D. Gocayne, et al., Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii, Science (New York, N.Y.) 273 (5278) (1996) 1058–1073.

[45] T.L. Bailey, STREME: accurate and versatile sequence motif discovery, Bioinformatics 37 (18) (2021) 2834–2840.

[46] H. Manghwar, K. Lindsey, X. Zhang, S. Jin, CRISPR/Cas system: recent advances and future prospects for genome editing, Trends Plant Sci. 24 (12) (2019) 1102–1125.

[47] S. Serajian, E. Ahmadpour, S.M.R. Oliveira, Pereira MdL, Heidarzadeh S: CRISPR-Cas technology: emerging applications in clinical microbiology and infectious diseases, Pharmaceuticals 14 (11) (2021) 1171.

[48] G. Liu, Q. Lin, S. Jin, C. Gao, The CRISPR-Cas toolbox and gene editing technologies, Mol. Cell 82 (2) (2022) 333–347.

[49] E.V. Koonin, K.S. Makarova, Origins and evolution of CRISPR-Cas systems, Philos. Trans. R. Soc. B 374 (1772) (2019) 20180087.

[50] W.L. Chew, Immunity to CRISPR Cas9 and Cas12a Therapeutics, Systems Biology and Medicine, Wiley Interdisciplinary Reviews, 2018, p. 10.

[51] Z. Huang, J. Fang, M. Zhou, Z. Gong, T. Xiang, CRISPR-Cas13: a new technology for the rapid detection of pathogenic microorganisms, Front. Microbiol. 13 (2022).