Research Article

PLPTP: A Motif-based Interpretable Deep Learning Framework Based on Protein Language Models for Peptide Toxicity Prediction

Shun Gao [1], Yanna Jia [1], Feifei Cui [1], Junlin Xu [2], Yajie Meng [3], Leyi Wei [4,5], Qingchen Zhang [1], Quan Zou [6,7], and Zilong Zhang [1,*]

*1 - **School of Computer Science and Technology,** Hainan University, Haikou 570228, China*
*2 - **School of Computer Science and Technology,** Wuhan University of Science and Technology, Wuhan 430081 Hubei, China*
*3 - **School of Computer Science and Artificial Intelligence,** Wuhan Textile University, Wuhan 430200 Hubei, China*
*4 - **Centre for Artificial Intelligence Driven Drug Discovery,** Faculty of Applied Science, Macao Polytechnic University, Macao Special Administrative Region of China*
*5 - **School of Informatics,** Xiamen University, Xiamen, China*
*6 - **Institute of Fundamental and Frontier Sciences,** University of Electronic Science and Technology of China, Chengdu 610054, China*
*7 - **Yangtze Delta Region Institute (Quzhou),** University of Electronic Science and Technology of China, Quzhou 324000, China*

***Correspondence to Zilong Zhang:** zhangzilong@hainanu.edu.cn (Z. Zhang)*
https://doi.org/10.1016/j.jmb.2025.169115
Edited by Rita Casadio

Abstract

Peptide toxicity prediction holds significant importance in drug development and biotechnology, as accurately identifying toxic peptide sequences is crucial for designing safer peptide-based drugs. This study proposes a deep learning-based model for peptide toxicity prediction, integrating Evolutionary Scale Modeling (ESM2), Bidirectional Long Short-Term Memory (BiLSTM), and Deep Neural Network (DNN). The ESM2 model captures evolutionary information from peptide sequences, providing a rich context for the sequences; the BiLSTM network focuses on extracting contextual dependencies, thereby capturing long-range dependencies within the sequence; and the DNN further classifies the extracted features to achieve the final toxicity prediction. To enhance the reliability and transparency of the model, we also conducted motif analysis to identify key patterns in the data, which helps to explain the model's attention mechanism and its classification performance. To address the class imbalance in the dataset, we employed Focal Loss as the loss function, which enhances the model's ability to identify minority class samples by reducing the contribution of easily classified samples. Experimental results demonstrate that the proposed model performs exceptionally well across multiple evaluation metrics, particularly in handling imbalanced data, achieving significant improvements over traditional methods. This result highlights the model's potential to improve the accuracy of peptide toxicity prediction and its valuable role in drug development and biotechnology research. The PLPTP web server is available at https://www.bioai-lab.com/PLPTP.

Introduction

Peptides are short chains of amino acids linked by peptide bonds, typically consisting of 2–50 amino acids. They occupy a unique position in biology due to their length and complexity, which lie between individual amino acids and longer proteins.[1–4] Over 7,000 naturally occurring peptides

have been identified, each playing a crucial role in various biological processes. For instance, peptides can function as hormones to regulate physiological processes, as neurotransmitters to transmit nerve signals, or as antimicrobial agents to protect the body from pathogens.[5–8] The simple structure and ease of synthesis of peptides make them highly valuable in biological research and medical applications. In drug development, many natural peptides are utilized for their demonstrated biological activities, such as antimicrobial, antiviral, and immunomodulatory properties.[9–12] These characteristics make peptides promising candidates for the development of new therapeutic agents.[13,14] Thus, understanding the structure and function of peptides is essential for advancing drug discovery and treatment strategies.

However, while peptides can perform their biological functions, they may also pose toxic effects on organisms, known as peptide toxicity. Peptide toxicity refers to the phenomenon where peptide molecules interact with cell membranes or intracellular targets, leading to cell damage or death. The toxicological properties of peptides are influenced by various factors, including their sequence, structure, dosage, and route of administration.[15,16] Therefore, despite the therapeutic potential of peptides, the evaluation of peptide toxicity remains a crucial step in peptide development, ensuring that therapeutic peptides do not cause harmful side effects when administered to patients.[17–19]

Although the importance of peptide toxicity prediction is increasingly recognized, existing prediction methods still face several limitations. Traditional experimental methods are often time-consuming, labor-intensive, and costly, making them unsuitable for high-throughput screening. In recent years, with the advancements in computational biology and artificial intelligence, computational approaches for peptide toxicity prediction have become a research hotspot.[20,21] These methods primarily rely on machine learning models to predict the toxicity of unknown peptides by extracting and recognizing patterns from known toxic peptides. For instance, ATSE is a peptide toxicity prediction tool based on graph neural networks and attention mechanisms, utilizing peptide molecular graphs and evolutionary information to predict potential toxicity.[22] ToxIBTL combines convolutional neural networks (CNN)[23–26] and bidirectional gated recurrent units (BiGRU)[27] in a hybrid network to automatically learn evolutionary information within peptide sequences, and employs a model based on graph and statistical features to generate physicochemical features for predicting the toxicity of peptides and proteins.[28] ToxinPred2 integrates multiple techniques, including BLAST-based similarity searches, motif searches using the MERCI tool, and machine learning prediction models, to identify peptides and proteins.[29] CSM-Toxin is a peptide toxicity prediction model based on Transformer models that encode protein sequence information using the deep learning natural language model ProteinBert.[30,31] CAPTP employs a novel encoder combining convolutional modulation and self-attention to automatically learn peptide sequence representations, improving the accuracy of peptide toxicity prediction from sequences.[32]

Nonetheless, these methods still face challenges in practical applications, such as insufficient model generalization, limited quality and scale of datasets, and uncertainty in feature selection. These issues limit the accuracy and reliability of peptide toxicity prediction methods, necessitating further research and improvements.

To address these issues, this study developed a peptide toxicity prediction method called Protein Language Model-based and BiLSTM-enhanced Peptide Toxicity Prediction (PLPTP), which utilizes a combination of Evolutionary Scale Modeling (ESM2)[33] and Bidirectional Long Short-Term Memory (BiLSTM).[34] The ESM2 model, through deep learning on a large corpus of protein sequence data, captures evolutionary information within protein sequences, demonstrating excellent performance in processing biological sequences. BiLSTM, a type of recurrent neural network suitable for sequence data, effectively captures long-term dependencies within sequences. By integrating protein sequence embeddings generated by ESM2 with a BiLSTM network, our model can better understand the complex features within peptide sequences, thereby improving the accuracy and robustness of peptide toxicity prediction. Compared to traditional methods, the ESM2 with BiLSTM model demonstrates higher efficiency in handling large-scale sequence data and possesses stronger generalization capabilities, enabling more accurate predictions of the toxicity of peptides.

Methods and Materials

Benchmark dataset

In this study, we used a dataset constructed by Jiao et al.[32] This dataset was initially collected by Shi et al. from the literature on CSM-Toxin,[30] ToxinPred2,[29] and ATSE,[22] and the samples were filtered to ensure that their lengths were all within 50 amino acids. To supplement the non-toxic samples, they also retrieved manually annotated and reviewed non-toxic peptides from the UniProt[35] database, excluding sequences containing non-standard amino acids or those longer than 50 amino acids. After merging these non-toxic peptides with the existing dataset, samples with conflicting labels were removed. Following the deduplication process, the final dataset included 2,138 toxic peptides and 5,375 non-toxic peptides. Of these samples, 85% were randomly allocated for training, while the remaining 15% were used as an independent

Table 1 The Training set and Test set for peptide toxicity.

Dataset	Negative sample	Positive samples	Total
Training Set	4569	1818	6387
Test Set	806	320	1126

test set. Table 1 presents the Benchmark dataset used for peptide toxicity prediction.

In this study, we performed statistical analysis on the amino acid frequency distribution and peptide length distribution in the dataset. First, we observed significant differences in the amino acid frequency distribution between positive and negative samples, with certain amino acids showing notably different frequencies across the two groups. This suggests that these amino acids may play an important role in distinguishing peptide toxicity. Furthermore, we found that the amino acid frequency and peptide length distributions in both the training and test sets are highly similar, demonstrating the rationality of the dataset split. This consistency ensures the reliability and validity of model evaluation, while also mitigating potential biases from an imbalanced or improperly divided dataset. Figure 1 shows the distributions of amino acid frequencies and peptide lengths in in the training and test sets.

Focal Loss

In this study, the dataset has a significant class imbalance issue, with a large disparity between the number of positive and negative samples. This imbalance may adversely affect the training process of the model. Traditional cross-entropy loss functions tend to be dominated by the majority class when dealing with imbalanced data, leading the model to favor predicting the majority class and thus neglecting the correct classification of minority class samples.

To address this issue, we have adopted Focal Loss[36] as an alternative to the traditional cross-entropy loss function. Focal Loss is designed to handle imbalanced classification problems by introducing a modulation factor that reduces the contribution of easily classified samples to the loss and puts more focus on hard-to-classify samples.[37] The definition of Focal Loss is as follows:

$$\mathrm{FL}(p_t) = \mathrm{f}(x) = \begin{cases} -\alpha_t (1-p_t)^{\gamma} \log(p_t), & y = 1 \\ -(1-\alpha_t) p_t^{\gamma} \log(1-p_t), & y = 0 \end{cases} \quad (1)$$

where p_t is the model's estimated probability for the true class, α_t is a weighting factor used to balance class importance, and γ is a focusing parameter that adjusts the weight of easy-to-classify samples. Specifically, the Focal Loss dynamic reduces the loss weight of correctly classified samples, thereby allowing the model to focus more on minority and hard-to-classify samples. This approach effectively addresses the class imbalance issue and improves model performance on minority classes.

In practical application, adopting Focal Loss significantly improved the model's performance on minority class samples. Compared to traditional cross-entropy loss, Focal Loss is more effective in handling imbalanced sample distributions, leading to more accurate predictions across different classes. The test set results for various loss functions are shown in Table 2.

The architecture of the proposed method PLPTP

In this study, we propose a hybrid model architecture based on ESM2, double-layer BiLSTM, and Deep Neural Networks (DNN)[30] for the task of peptide toxicity prediction. First, the ESM2 model serves as a feature extraction module, generating high-dimensional context-aware embeddings from peptide sequences. Built on the Transformer architecture, ESM2 leverages its self-attention mechanism to capture long-range dependencies between amino acid residues, producing feature representations that encompass both local characteristics and global semantic relationships. Thanks to pre-training on large-scale protein sequence data, ESM2 also encodes evolutionary and structural insights, providing rich inputs for subsequent layers.

Next, the double-layer BiLSTM module processes the sequence features output by ESM2 through bidirectional modeling, capturing both forward and backward temporal dependencies within the sequence. Its memory cells and gating mechanisms effectively address the vanishing gradient problem encountered by traditional RNNs in long sequences, enhancing the representation of comprehensive global and local features while preserving critical information.

Finally, the features processed by BiLSTM are fed into the DNN, which performs nonlinear transformations and integration through a multi-layer fully connected network. The DNN progressively increases the level of feature abstraction, yielding highly expressive feature representations, and outputs a probability vector for peptide toxicity prediction. This architecture combines ESM2's sequence comprehension, BiLSTM's temporal feature extraction, and DNN's deep integration capabilities to achieve precise toxicity predictions.

Given the widespread use and proven efficacy of ESM2, BiLSTM, and DNN in sequence modeling tasks, we provide only a brief overview here. Detailed model structures and algorithmic specifics can be found in the original papers.[33,34,38] And Figure 2 illustrates the PLPTP framework.

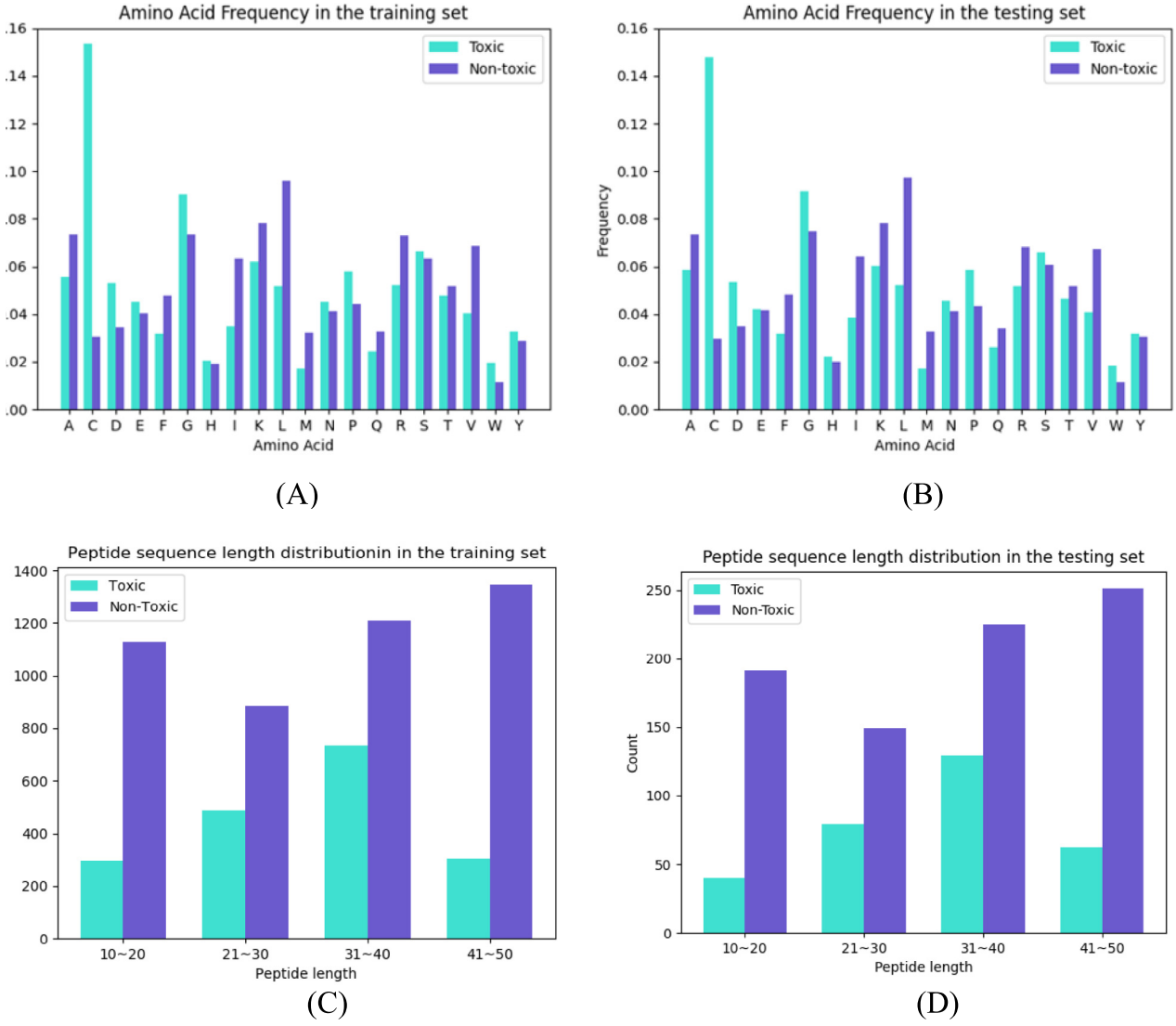

Figure 1. (A–B) The amino acid frequency distribution shows the prevalence of each amino acid in the dataset, illustrating their occurrence across the peptide sequences. (C–D) The peptide sequence length distribution shows the distribution of peptide lengths, revealing the range and distribution characteristics of the peptides.

Table 2 Performance of different loss functions.

Methods	BACC	SN	SP	MCC	AUC
Focal-Loss	**0.976**	**0.975**	**0.978**	0.944	**0.997**
Cross-Entropy Loss	0.929	0.903	0.956	**0.969**	0.969

Note: The highest score in each column is shown in bold.

Evaluation metrics

To comprehensively evaluate the model's performance in peptide toxicity prediction, this study employed five key evaluation metrics[39,40]: Sensitivity (SN), Specificity (SP), Balanced Accuracy (BACC), Matthews Correlation Coefficient (MCC), and Area Under the Curve (AUC). These metrics offer a multifaceted view of the model's performance, ensuring that the results are broadly applicable and reliable. The formula is as follows:

$$\begin{cases} SN = \frac{TP}{TP+FN} \\ SP = \frac{TN}{TN+FP} \\ BACC = \frac{1}{2}(SN + SP) \\ MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FP) \times (TN+FN) \times (TP+FN) \times (TN+FP)}} \end{cases} \quad (2)$$

SN reflects the model's ability to identify toxic peptides by calculating the proportion of actual toxic peptides that are correctly predicted as toxic. This metric is crucial as it helps minimize the number of missed detections, thereby enhancing the model's practical utility. SP assesses the model's ability to identify non-toxic peptides, calculating the proportion of actual non-toxic peptides that are correctly predicted as non-toxic.

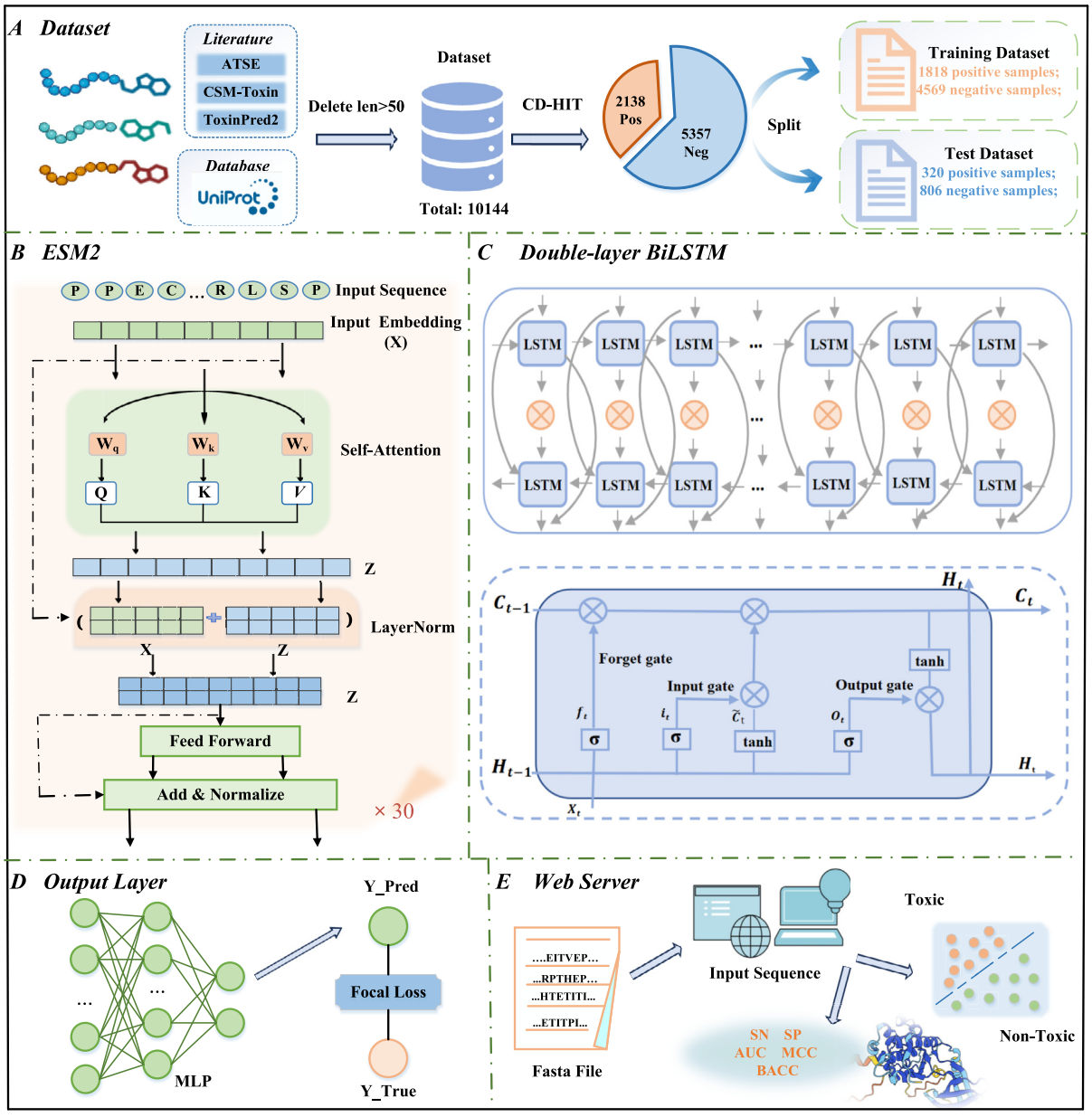

Figure 2. The model architecture of PLPTP contains four main modules. (A) Dataset module. (B) ESM2 module. (C) Double-layer BiLSTM module. (D) Classification module. (E) Web Server module.

High specificity indicates that the model effectively distinguishes non-toxic peptides, reducing the likelihood of misclassifying non-toxic peptides as toxic.

BACC combines sensitivity and specificity to provide a comprehensive evaluation of the model's performance in situations with class imbalance. This metric is particularly important as, in real-world applications, there is often a significant disparity between the number of toxic and non-toxic peptides. BACC balances the assessment of both positive and negative samples, offering a fair evaluation of the model's ability to handle imbalanced datasets. MCC provides a more comprehensive assessment by considering all four types of predictions: true positives, true negatives, false positives, and false negatives. MCC ranges from −1 to 1, with 1 indicating perfect prediction, 0 indicating performance equivalent to random chance, and −1 indicating completely incorrect prediction. MCC is particularly valuable for imbalanced datasets as it provides an accurate evaluation of the model's performance across both positive and negative classes.

AUC measures the model's discriminative ability by assessing the area under the ROC curve. AUC values range from 0 to 1, with values closer to 1

indicating a stronger ability to differentiate between positive and negative classes. AUC is a comprehensive metric that evaluates the model's performance across various classification thresholds.

By integrating these evaluation metrics—SN, SP, BACC, MCC, and AUC—this study provides a thorough analysis of the model's performance in peptide toxicity prediction from multiple angles.

Experiment and Discussions

Comparative analysis of PLPTP with classical models

In this study, we compared various classical deep learning models, including Transformer, Transformer with GRU, ESM2, and combinations of ESM2 with different feature extraction modules such as CNN, GRU, and BiLSTM. Each model exhibited varying degrees of performance improvement in the task of peptide toxicity prediction from protein sequences. The classical Transformer model effectively captures global features in sequences, but its ability to model structural information is limited, leaving room for improvement in fine-grained feature extraction.

To address this limitation, we combined the Transformer with modules like GRU and CNN to further enhance the model's ability to capture temporal dependencies. The transformer with GRU excelled in capturing long-range dependencies, while ESM2 with CNN effectively extracted local features through convolutional operations, significantly improving the model's overall performance. However, despite the improvements in BACC, SN, and SP achieved by these combined models, there remain some limitations in capturing the complex dependencies within sequences.

In contrast, the final ESM2 with the BiLSTM model achieved the best results across all evaluation metrics. By combining the strong representational power of ESM2 with BiLSTM's ability to capture bidirectional sequence dependencies, the model not only retained the global feature extraction advantages of the Transformer but also optimized the processing of temporal information. As a result, the model performed exceptionally well in terms of SN, SP, and BACC. Notably, it significantly outperformed other models in the MCC, demonstrating its robustness on imbalanced datasets. The results are summarized in Table 3 and visualized in Figure 3.

Table 3 Results of common deep learning methods on the Test set.

Model	BACC	SN	SP	MCC	AUC
Transformer	0.909	0.878	0.940	0.812	0.854
Transformer + GRU	0.914	0.866	0.963	0.839	0.935
ESM2	0.924	0.924	0.924	0.823	0.975
ESM2 + CNN	0.930	0.900	0.960	0.862	0.910
ESM2 + GRU	0.969	0.975	0.920	0.920	0.990
ESM2 + BiLSTM	**0.976**	**0.975**	**0.978**	**0.944**	**0.997**

Note: The highest score in each column is shown in bold.

Performance of ablation experiments on the baseline model

In the ablation study, we evaluated the impact of removing specific components from our model on its performance across various metrics. The results are summarized in Table 4 and visualized in Figure 4.

First, removing the ESM2 component resulted in a significant performance decline, highlighting the crucial role of ESM2 in the model's ability to capture relevant features. There was a noticeable decrease in all metrics. The BACC dropped from 0.976 in the full model to 0.901, while the AUC fell significantly to 0.925, indicating a reduced capacity for feature extraction. Additionally, the SN decreased from 0.975 to 0.844, and the SP dropped from 0.978 to 0.958, suggesting the model's ability to distinguish between positive and negative samples was impaired. The MCC, an indicator of overall prediction quality in binary classification tasks, also fell from 0.944 to 0.814, further underscoring the critical role of ESM2 in the model.

On the other hand, the removal of the BiLSTM component resulted in a smaller performance reduction, demonstrating that its impact was less pronounced compared to ESM2. Specifically, the model with BiLSTM removed had a balanced accuracy of 0.924, an AUC of 0.975, and both SN and SP of 0.924, with an MCC of 0.823. These metrics are considerably better than the model without ESM2, indicating that while BiLSTM improves performance, its contribution is not as vital as that of ESM2. BiLSTM primarily captures sequential information, whereas ESM2 is able to capture more complex evolutionary features, which may explain why the removal of BiLSTM leads to a smaller drop in performance.

The complete PLPTP model achieved the highest performance across all evaluation metrics. The balanced accuracy reached 0.976, AUC was nearly perfect at 0.997, and SN and SP were 0.975 and 0.978, respectively. The MCC was 0.944, demonstrating that the combination of ESM2 and BiLSTM enables the model to effectively capture the complexity of sequences and make more accurate predictions in the classification task.

Comparative performance of PLPTP with existing models

In comparison with existing prediction models CSM-Toxin,[30] ToxinPred2,[29] ToxIBTL,[28] CAPTP,[32]

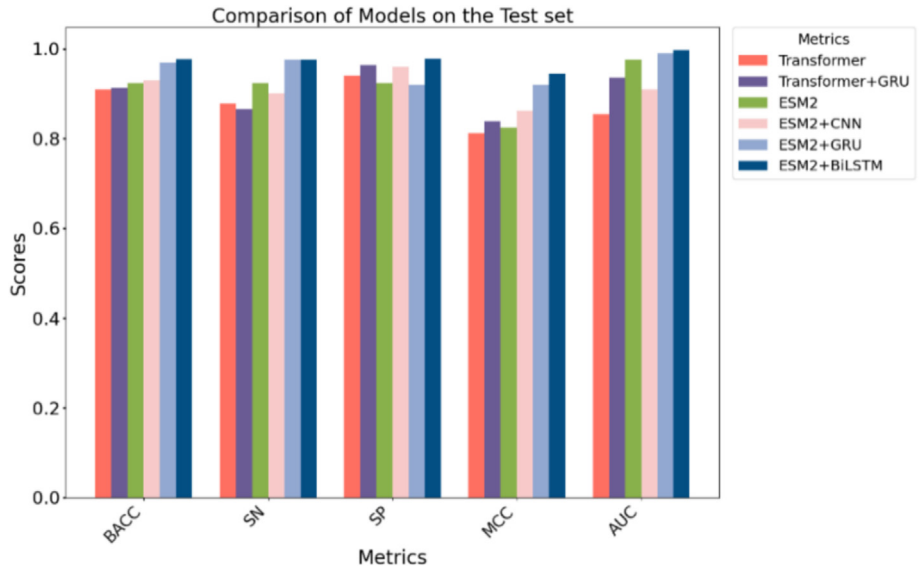

Figure 3. Results of common deep learning methods on the Test set.

Table 4 Results of removing different modules in the model on the Test set.

Model	BACC	SN	SP	MCC	AUC
Remove ESM2	0.901	0.844	0.958	0.814	0.925
Remove BiLSTM	0.924	0.924	0.924	0.823	0.975
PLPTP(our)	**0.976**	**0.975**	**0.978**	**0.944**	**0.997**

Note: The highest score in each column is shown in bold.

the PLPTP model exhibits superior performance across all evaluation metrics, significantly surpassing other approaches. As shown in Table 5 and Figure 5, the PLPTP model demonstrates exceptional performance, with a BACC of 0.976 and an AUC nearly reaching 1.0 (0.997). The model's SN and SP were 0.975 and 0.978, respectively, demonstrating its high precision and stability in peptide toxicity prediction tasks. In contrast, although other methods perform well in certain metrics, they still fall short compared to PLPTP overall. For instance, while ToxIBTL and CAPTP excel in SP, they do not match PLPTP in SN and overall balance. Notably, CSM-Toxin shows poor performance across all metrics, with a BACC of only 0.478 and an SN as low as 0.041, highlighting its limitations in peptide toxicity prediction.

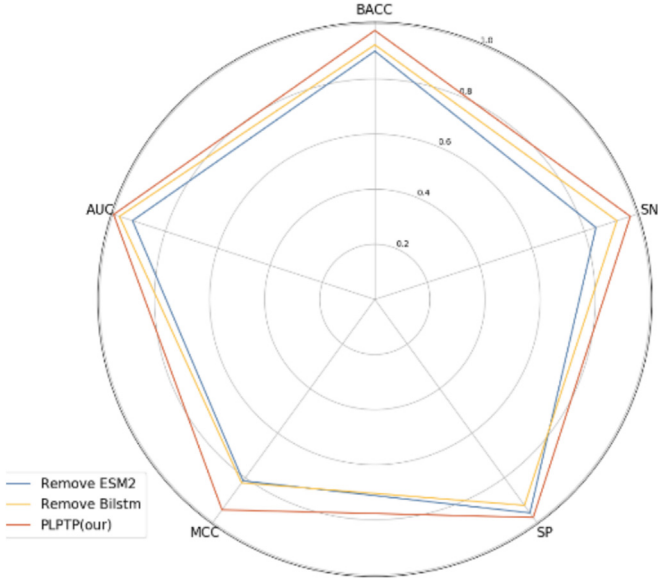

Figure 4. The shows the results of removing different modules in the model on the Test set.

Table 5 Comparison of existing methods on the Test set.

Model	BACC	SN	SP	MCC	AUC
CSM-Toxin	0.478	0.041	0.916	−0.076	0.400
ToxinPred2	0.642	0.959	0.325	0.299	0.874
ToxIBTL	0.916	0.916	0.916	0.803	0.916
CAPTP	0.916	0.906	0.926	0.811	0.959
PLPTP(our)	**0.976**	**0.975**	**0.978**	**0.944**	**0.997**

Note: The highest score in each column is shown in bold.

Furthermore, although ToxinPred2 achieved an SN of 0.959, its performance in other important metrics like SP and MCC is less satisfactory, with a BACC of only 0.642. This indicates a shortcoming in balancing prediction results. In this context, PLPTP demonstrates a clear advantage, not only showing balanced performance in SN and SP but also maintaining a high level of model balance, as evidenced by its MCC of 0.944. Overall, PLPTP represents a significant improvement in peptide toxicity prediction tasks through its innovative architectural design.

Meanwhile, to further validate the model's performance, we adopted a more stringent threshold to divide the training and test sets. The results of PLPTP under this stricter data partitioning remain outstanding, and the relevant evaluation results can be found in the supplementary Table 1.

UMAP visualization and analysis of model attention

In this study, we utilized UMAP (Uniform Manifold Approximation and Projection)[41,42] to visualize the performance of the model on the test set. UMAP is a dimensionality reduction technique that can embed high-dimensional data into a lower-dimensional space, making it easier to observe the structure and patterns of the data.

As shown in Figure 6(A–B), Figure 6(A) depicts the raw data after encoding but prior to training. At this stage, the features have not yet been optimized by the model, resulting in a scattered and disordered distribution of data points in high-dimensional space. The points from different classes are intermingled and difficult to distinguish, fully reflecting the high heterogeneity of the original data. To further investigate the model's classification effectiveness, we applied the same UMAP dimensionality reduction to the features after model processing. Figure 6(B) illustrates the features extracted from the layer immediately preceding the classification step in the DNN. At this point, the UMAP visualization reveals a striking transformation in the data structure: the data points are distinctly clustered into two well-separated groups. This clear separation demonstrates that the model, through training, successfully extracts and differentiates the key features of samples from different classes, highlighting its exceptional classification capability.

In addition to the UMAP analysis, we conducted a detailed examination of the model's attention mechanisms and compared it with the motif analysis obtained from the data. Attention mechanisms provide insights into which parts of the input data the model focuses on during the classification process. By analyzing these attention maps, we could identify how the model emphasizes different regions of the data and how

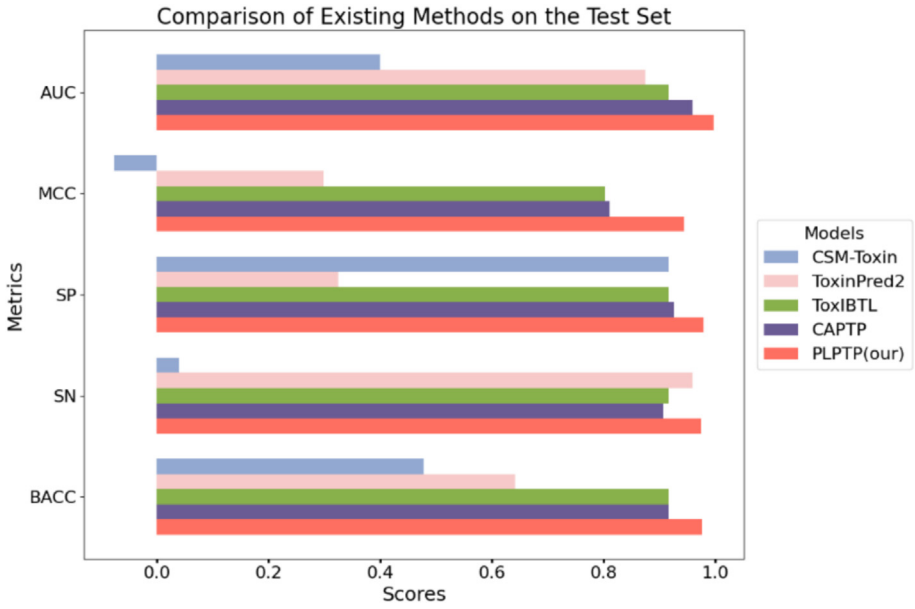

Figure 5. The performances of our method and existing methods on the Test set.

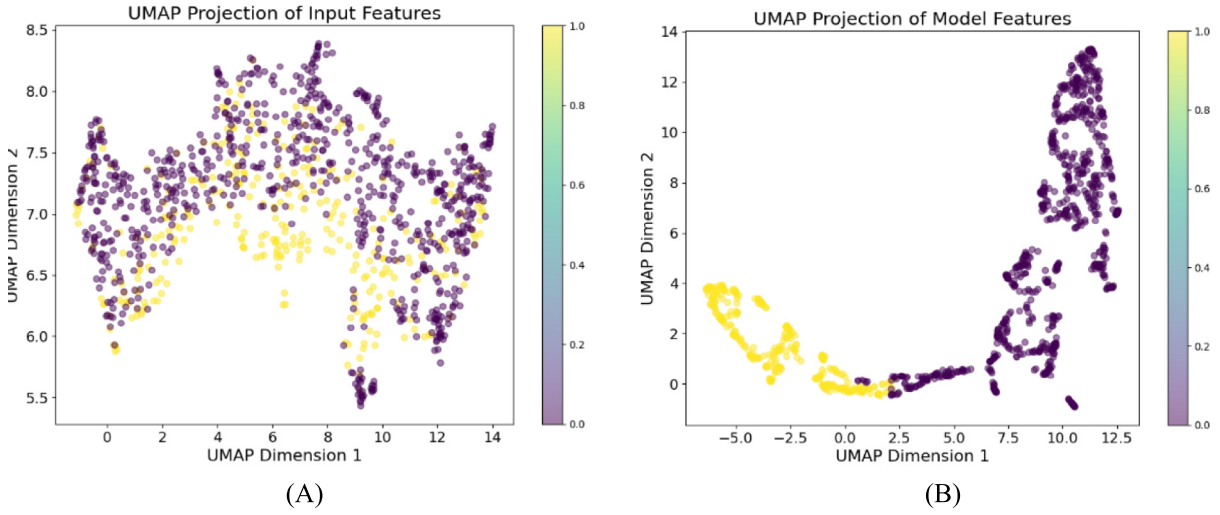

Figure 6. (A) The UMAP visualization of the input data prior to model inference. (B) The UMAP visualization of the data after model inference.

these regions impact the classification results. The motif analysis[43,44] further complemented our understanding by identifying recurring patterns or motifs from the data, which help to explain the features the attention mechanism focuses on.

To enhance the reliability and transparency of the model, we performed motif analysis using the traditional tool STREME,[45] with the results displayed on the right side of Figure 7. The left side of Figure 7 shows the sequence attention score heatmap obtained from the ESM2 model, where the brightness of the color reflects the score level, with higher scores corresponding to brighter colors. By comparing the left and right sides of Figure 7, we can observe that the sequence fragments captured by the model effectively map to the motifs identified by STREME on the right side, further confirming that the model is able to recognize and capture critical sequence information. Our motif identification results are highly consistent with those detected by the MEME Suite tool STREME, validating the model's effectiveness in recognizing conserved sequence features. Additionally, both P-values and E-values are used to assess the significance of the match between the sequence and motif. The *P*-value indicates the degree of deviation from randomness, while the *E*-value represents the expected frequency of the match occurring in random data. The smaller the values, the more significant the match.

Through both the UMAP visualization and the detailed attention and motif analysis, we gain

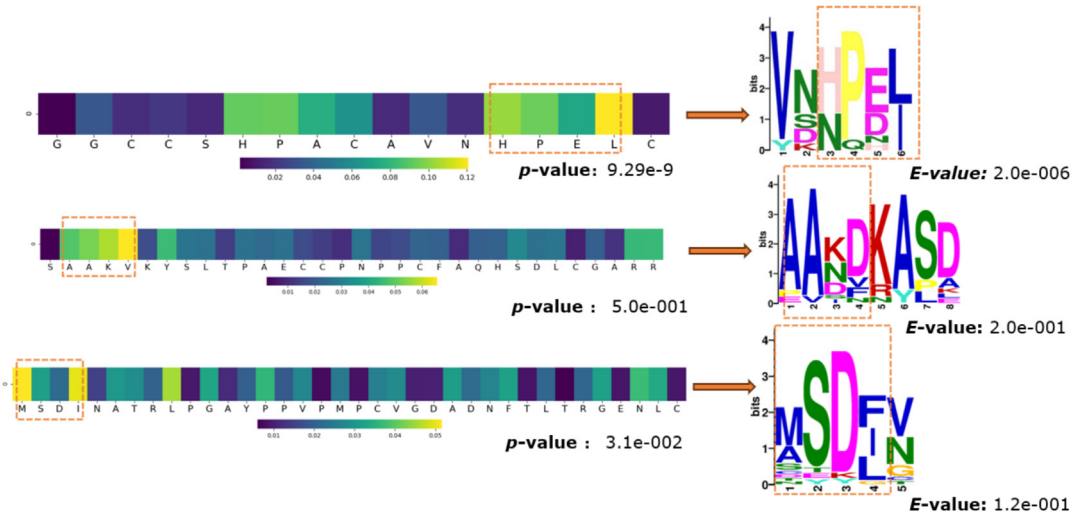

Figure 7. The correspondence between motif analysis and the attention heatmap in the model.

comprehensive insights into the model's ability to classify and differentiate between samples. The UMAP plots illustrate the improvement in the model's classification capabilities, while the attention and motif analyses provide a deeper understanding of the model's internal processes and feature extraction mechanisms. Together, these analyses validate the performance improvements and offer valuable perspectives on the model's operational dynamics.

Web Server Module

To facilitate peptide toxicity prediction for researchers and users, we developed a user-friendly online web server that allows users to perform quick predictive analysis using the model trained in this study. By integrating the trained model, users can directly upload sequences and obtain prediction results without needing complex programming or hardware support.

Users only need to upload the target peptide sequence in FASTA or text format to the platform, and the server will automatically run preprocessing steps and predict the peptide's toxicity through the model. The backend server uses ESM2 to extract high-level semantic features of the peptide sequence, then BiLSTM captures bidirectional dependencies within the sequence. The entire process is transparent to the user, without requiring any deep understanding of machine learning or deep learning models. The PLPTP web server is available at https://www.bioai-lab.com/PLPTP.

Discussion and Limitations

In this study, we employed a model integrating ESM2, BiLSTM, and DNN to address the issue of peptide toxicity prediction. ESM2 is responsible for extracting deep features from peptide sequences, enhancing the sequence representation using evolutionary information, which enables the model to capture the potential functional characteristics of peptide chains and provides high-quality input for the subsequent components. BiLSTM processes the peptide sequences bidirectionally, capturing the temporal dependencies between amino acids, allowing the model to better understand the contextual information within the sequences, especially the influence of distant amino acids, thus enhancing the feature representation. DNN then integrates the features from ESM2 and BiLSTM, utilizing its nonlinear mapping ability to help the model recognize complex feature relationships, ultimately improving the model's ability to predict peptide toxicity.

Although the model demonstrates significant performance improvements, there are still several limitations and potential areas for enhancement. One of the primary limitations is the model's complexity and computational demands. The combination of ESM2, BiLSTM, and DNN results in a complex architecture that requires substantial computational resources and time for training. This complexity may lead to longer training periods and increased resource consumption, limiting the model's scalability, particularly when applied to larger datasets or real-time scenarios.

Additionally, given the model's complexity, the risk of overfitting remains a concern. Despite employing techniques such as dropout and regularization, the model's complexity still presents a risk of overfitting, especially if the dataset is not sufficiently large or diverse. Without proper management, the model may perform well on training data but struggle with generalization to unseen samples.

The issue of data imbalance continues to be a critical challenge. While Focal Loss has been employed to address this problem, the model's performance can still be affected by the inherent imbalance in the dataset, particularly with lower accuracy in predicting underrepresented classes.

To address these limitations, future research could explore several avenues for improvement. Simplifying the model architecture or employing more efficient training algorithms could reduce the computational burden and improve scalability. Additionally, incorporating more advanced regularization techniques or ensemble methods could help improve the model's generalization ability and reduce the risk of overfitting. To further tackle the issue of data imbalance, methods such as data augmentation or synthetic sample generation could be investigated to better represent minority classes and enhance the overall model performance.

Conclusion

This study presents a model based on the ESM2 combined with BiLSTM and DNN architecture applied to peptide toxicity prediction tasks. By integrating the powerful representation capabilities of ESM2, the efficient sequence information capture of BiLSTM, and the deep feature extraction of DNN, the model demonstrates outstanding performance across multiple evaluation metrics, achieving significant improvements in balanced accuracy, AUC, SN, SP, and MCC.

Furthermore, comparisons with classic machine learning and deep learning models further validate the model's superiority in complex sequence pattern recognition tasks. Combined with the interpretability analysis of the attention mechanism, the model not only performs efficient peptide toxicity prediction but also accurately identifies and focuses on key motif regions within sequences, providing biologists with a tool for the in-depth understanding of the relationship

between protein sequences and toxicity. Overall, the PLPTP model offers a precise, robust, and interpretable solution for peptide toxicity prediction, with significant research and practical value.

Availability and implementation

The data and methods are available at https://github.com/birdsmart/PLPTP. For the convenience of the research community, a web server has been established at https://www.bioai-lab.com/PLPTP.

CRediT authorship contribution statement

Shun Gao: Writing – original draft, Conceptualization. **Yanna Jia:** Writing – review & editing, Visualization. **Feifei Cui:** Funding acquisition, Conceptualization. **Junlin Xu:** Methodology, Formal analysis. **Yajie Meng:** Writing – review & editing, Validation. **Leyi Wei:** Writing – review & editing, Investigation. **Qingchen Zhang:** Writing – review & editing, Formal analysis. **Quan Zou:** Writing – review & editing, Methodology. **Zilong Zhang:** Writing – review & editing, Validation.

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.

Acknowledgment

The work is supported by the National Natural Science Foundation of China (Nos. 62450002, 62262015) and Science and Technology special fund of Hainan Province (ZDYF2024GXJS018).

Appendix A. Supplementary material

Supplementary material to this article can be found online at https://doi.org/10.1016/j.jmb.2025.169115.

Received 10 January 2025;
Accepted 25 March 2025;
Available online 28 March 2025

Keywords:
Peptide toxicity prediction;
ESM2;
BiLSTM

References

1. Barman, P., Joshi, S., Sharma, S., Preet, S., Sharma, S., Saini, A., (2023). Strategic approaches to improvise peptide drugs as next generation therapeutics. *Int. J. Pept. Res. Ther.* **29** (4), 61.
2. Anand, U., Bandyopadhyay, A., Jha, N.K., Pérez de la Lastra, J.M., Dey, A., (2023). Translational aspect in peptide drug discovery and development: an emerging therapeutic candidate. *Biofactors* **49** (2), 251–269.
3. Liu, R., Fu, X., Yan, S., Zhang, Z., Cui, F., (2023). AIPPT: predicts anti-inflammatory peptides using the most characteristic subset of bases and sequences by stacking ensemble learning strategies. *In: 2023 IEEE international conference on bioinformatics and biomedicine (BIBM): 5–8 Dec., 2023*, pp. 23–29.
4. Mehta, K., Vyas, P., Mujawar, S., Hazam, P.K., Vyas, A., (2023). Design and in-silico screening of short antimicrobial peptides (AMPs) as anti-tubercular agents targeting INHA. *Curr. Bioinform.* **18** (9), 715–736.
5. Agostini, F., Cirillo, D., Livi, C.M., Delli Ponti, R., Tartaglia, G.G., (2014). cc SOL omics: a webserver for solubility prediction of endogenous and heterologous expression in *Escherichia coli*. *Bioinformatics* **30** (20), 2975–2977.
6. Aguilar-Zapata, D., Petraitiene, R., Petraitis, V., (2015). Echinocandins: the expanding antifungal armamentarium. *Clin. Infect. Dis.* **61** (suppl_6), S604–S611.
7. Padhi, A., Sengupta, M., Sengupta, S., Roehm, K.H., Sonawane, A., (2014). Antimicrobial peptides and proteins in mycobacterial therapy: current status and future prospects. *Tuberculosis* **94** (4), 363–373.
8. Souery, W.N., Bishop, C.J., (2018). Clinically advancing and promising polymer-based therapeutics. *Acta Biomater.* **67**, 1–20.
9. Muttenthaler, M., King, G.F., Adams, D.J., Alewood, P.F., (2021). Trends in peptide drug discovery. *Nature Rev. Drug Discov.* **20** (4), 309–325.
10. Henninot, A., Collins, J.C., Nuss, J.M., (2018). The current state of peptide drug discovery: back to the future?. *J. Med. Chem.* **61** (4), 1382–1414.
11. Fosgerau, K., Hoffmann, T., (2015). Peptide therapeutics: current status and future directions. *Drug Discov. Today* **20** (1), 122–128.
12. Fu, X., Duan, H., Zang, X., Liu, C., Li, X., Zhang, Q., Zhang, Z., Zou, Q., Cui, F., (2024). Hyb_SEnc: an antituberculosis peptide predictor based on a hybrid feature vector and stacked ensemble learning. *IEEE/ACM Trans. Comput. Biol. Bioinf.*, 1–17.
13. Yan, C., Geng, A., Pan, Z., Zhang, Z., Cui, F., (2024). MultiFeatVotPIP: a voting-based ensemble learning framework for predicting proinflammatory peptides. *Brief. Bioinform.* **25** (6)
14. Rezende, S.B., Lima, L.R., Macedo, M.L.R., Franco, O.L., Cardoso, M.H., (2023). Advances in peptide/protein structure prediction tools and their relevance for structural biology in the last decade. *Curr. Bioinform.* **18** (7), 559–575.
15. Pognan, F., Beilmann, M., Boonen, H., Czich, A., Dear, G., Hewitt, P., Mow, T., Oinonen, T., Roth, A., Steger-Hartmann, T., (2023). The evolving role of investigative toxicology in the pharmaceutical industry. *Nature Rev. Drug Discov.* **22** (4), 317–335.

16. Lee, M.F., Poh, C.L., (2023). Strategies to improve the physicochemical properties of peptide-based drugs. *Pharm. Res.* **40** (3), 617–632.
17. Sharma, K., Sharma, K.K., Sharma, A., Jain, R., (2023). Peptide-based drug discovery: current status and recent advances. *Drug Discov. Today* **28**, (2)103464
18. Singh, P.P., Gupta, V., Prakash, B., (2023). Recent advancement in functional properties and toxicity assessment of plant-derived bioactive peptides using bioinformatic approaches. *Crit. Rev. Food Sci. Nutr.* **63** (20), 4503–4521.
19. Rao, B., Han, B., Wei, L., Zhang, Z., Jiang, X., Manavalan, B., (2024). CFCN: an HLA-peptide prediction model based on taylor extension theory and multi-view learning. *Curr. Bioinform.* **19** (10), 977–990.
20. Ao, C., Jiao, S., Wang, Y., Yu, L., Zou, Q., (2022). Biological sequence classification: a review on data and general methods. *Research* **2022**, 0011.
21. Yan, K., Lv, H., Shao, J., Chen, S., Liu, B., (2024). TPpred-SC: multi-functional therapeutic peptide prediction based on multi-label supervised contrastive learning. *Sci. China-Inf. Sci.* **67**, (11)212105
22. Wei, L., Ye, X., Xue, Y., Sakurai, T., Wei, L., (2021). ATSE: a peptide toxicity predictor by exploiting structural and evolutionary information based on graph neural network and attention mechanism. *Brief. Bioinform.* **22**, (5)bbab041
23. Krizhevsky, A., Sutskever, I., Hinton, G.E., (2017). ImageNet classification with deep convolutional neural networks. *Commun. ACM* **60** (6), 84–90.
24. Li, R., Liu, H., Hu, Y., (2023). Financial performance evaluation and forecasting of enterprises by the combination of PCA and CNN deep learning. *Front. Comp. Sci.* **17**, (6)176345
25. Shen, Z., Yang, Y., She, Q., Wang, C., Ma, J., Lin, Z., (2023). Newton design: designing CNNs with the family of Newton's methods. *Sci. China-Inf. Sci.* **66**, (6) 162101
26. Zhao, H., Liang, Z., He, Y., Wang, C., Yang, M., (2023). Cy-CNN: cylinder convolution based rotation-invariant neural network for point cloud registration. *Sci. China-Inf. Sci.* **66**, (5)152102
27. Cho, K., Van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., (2020). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:14061078*.
28. Wei, L., Ye, X., Sakurai, T., Mu, Z., Wei, L., (2022). ToxIBTL: prediction of peptide toxicity based on information bottleneck and transfer learning. *Bioinformatics* **38** (6), 1514–1524.
29. Sharma, N., Naorem, L.D., Jain, S., Raghava, G.P., (2022). ToxinPred2: an improved method for predicting toxicity of proteins. *Brief. Bioinform.* **23**, (5)bbac174
30. Morozov, V., Rodrigues, C.H., Ascher, D.B., (2023). CSM-Toxin: a web-server for predicting protein toxicity. *Pharmaceutics* **15** (2), 431.
31. Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., et al., (2024). A survey on large language model based autonomous agents. *Front. Comp. Sci.* **18**, (6)186345
32. Jiao, S., Ye, X., Sakurai, T., Zou, Q., Liu, R., (2024). Integrated convolution and self-attention for improving peptide toxicity prediction. *Bioinformatics* **40**, (5)btae297
33. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379** (6637), 1123–1130.
34. Huang, Z., Xu, W., Yu, K., (2015). Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:150801991*.
35. UniProt Consortium T, (2018). UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **46** (5), 2699.
36. Lin, T., (2017). Focal loss for dense object detection. *arXiv preprint arXiv:170802002*.
37. Dou, L., Zhang, Z., Xu, L., Zou, Q., (2022). iKcr_CNN: a novel computational tool for imbalance classification of human nonhistone crotonylation sites based on convolutional neural networks with focal loss. *Comput. Struct. Biotechnol. J.* **20**, 3268–3279.
38. Bengio, Y., (2009). Learning deep architectures for AI. Now Publishers Inc..
39. Fu, X., Yang, C., Su, Y., Liu, C., Qiu, H., Yu, Y., Su, G., Zhang, Q., Wei, L., Cui, F., et al., (2024). Machine learning enables comprehensive prediction of the relative protein abundance of multiple proteins on the protein corona. *Research* **7**, 0487.
40. Wang, Y., Zhai, Y., Ding, Y., Zou, Q., (2024). SBSM-Pro: support bio-sequence machine for proteins. *Sci. China Inf. Sci.* **67**, (11)212106
41. Rumelhart, D.E., Hinton, G.E., Williams, R.J., (1986). Learning representations by back-propagating errors. *Nature* **323** (6088), 533–536.
42. McInnes, L., Healy, J., Melville, J., (2018). Umap: uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:180203426*.
43. Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., Noble, W.S., (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37** (suppl_2), W202–W208.
44. Zhou, Z., Xiao, C., Yin, J., She, J., Duan, H., Liu, C., Fu, X., Cui, F., Qi, Q., Zhang, Z., (2024). PSAC-6mA: 6mA site identifier using self-attention capsule network based on sequence-positioning. *Comput. Biol. Med.* **171**, 108129.
45. Bailey, T.L., (2021). STREME: accurate and versatile sequence motif discovery. *Bioinformatics* **37** (18), 2834–2840.