Electronic Research Archive

https://www.aimspress.com/journal/ERA

ERA, 32(4): 2921–2935.
DOI: 10.3934/era.2024133
Received: 03 February 2024
Revised: 09 March 2024
Accepted: 02 April 2024
Published: 16 April 2024

Research article

Predicting multifunctional peptides based on a multi-scale ResNet model combined with channel attention mechanisms

Jing Liu[1], Hongpu Zhao[1], Yu Zhang[2,3], Jin Liu[1] and Xiao Guan[2,3,]*

[1] College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China
[2] School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China
[3] National Grain Industry (Urban Grain and Oil Security) Technology Innovation Center, Shanghai 200093, China

* **Correspondence:** Email: gnxo@usst.edu.cn.

Abstract: Peptides are biomolecules composed of multiple amino acid residues connected by peptide bonds, which are widely involved in physiological and biochemical processes in organisms and exhibit diverse functions. In previous studies, the focus was primarily on single-functional peptides. However, research trends indicate that an increasing number of multifunctional peptides are being identified and discovered. To address this challenge, we proposed a deep learning method based on multi-scale ResNet as the backbone combined with a channel attention mechanism (called MSRC) for the identification of multifunctional peptides. Furthermore, the data imbalance problem was solved through the comprehensive use of online data augmentation and confidence-based weighted loss functions. Experimental results demonstrated that the proposed MSRC method achieved an accuracy of 0.688 with an absolute true rate of 0.619. Notably, in predicting minority class peptides such as AEP, AHIVP, and BBP, the MSRC model exhibited heightened sensitivity, showcasing its exceptional capability in addressing issues related to minority classes. By enhancing the precision in identifying and predicting multifunctional peptides, the MSRC method was poised to contribute significantly to advancements in drug discovery, disease treatment, and biotechnology.

Keywords: multifunctional peptides; multi-scale ResNet; channel attention; data augmentation; optimization loss

1. Introduction

Peptides are organic molecules composed of amino acids in a specific arrangement. Their length can range from just a few amino acid residues to multiple amino acid units connected in an ordered sequence [1]. This organic molecule possesses certain functionalities and is commonly employed in the treatment of various diseases, such as cancer, diabetes, cardiovascular diseases, and others [2]. In order to enhance the effectiveness of peptide therapy, accurate identification of peptide functionality has become particularly important. Over the past few decades, an increasing number of bioactive peptides with diverse functionalities have been identified for drug development [3]. However, in the post-genomic era, the ability to identify peptide functionality through experimental methods is no longer sufficient to meet the demands of large-scale bioactive peptide identification. In this context, the introduction of computational methods allows researchers to pre-screen the functionality of peptides [4]. Nowadays, the combination of advanced computational technology and experimental methods has accelerated the discovery of potential therapeutic agents such as antimicrobial peptides (AMPs) [5] and anticancer peptides (ACPs) [6].

Data-driven computational methods such as machine learning and deep learning are widely used to predict the function of peptides [7,8]. Among these, feature selection is a critical step in machine learning and data science, which involves identifying and selecting data features that are most relevant to the target variable to reduce the dimensionality of the data. This process helps in enhancing the performance of predictive models [9]. Support vector machines (SVM) and random forests (RF) [10–12] were used to identify anti-cancer peptide (ACP), anti-parasitic peptide (APP), and Anti-inflammatory Peptides (AIP). Due to the structural and functional diversity of peptides, machine learning models may have limited generalization ability when dealing with untrained peptides. The advancement of artificial intelligence technology has promoted the importance and superiority of deep learning methods in the field of bioinformatics [13,14]. ACP-DL, ACP-2DCNN, and DeepACP use deep learning architectures [15–17] such as convolutional neural networks (CNN), long short-term memory (LSTM), and their combinations to achieve functional differentiation. Although traditional machine learning and deep learning methods have demonstrated excellent performance on specific peptide datasets, the above-mentioned studies primarily focus on the prediction of single-functional peptides. However, it has been discovered that an increasing number of peptides exhibit multiple functionalities [18]. Tang [4] and colleagues developed a deep learning model that integrates CNN and GRU for identifying peptides with various biological activities. Fan [19] and associates introduced a deep learning approach based on TextCNN and a multi-label focal dice loss function for predicting the functionalities of multifunctional therapeutic peptides. Moreover, Lv [20] and team proposed a method based on the Transformer architecture, utilizing label embedding techniques to explicitly extract functionally relevant information for predicting the multifaceted functionalities of therapeutic peptides.

The identification of multifunctional peptides can be considered as a multi-label classification task, where a peptide may possess multiple associated functions, and the model needs to simultaneously assign this set of relevant functional labels to the peptide. Li et al. [21] employed a combination of convolutional neural network layers for extracting convolutional features from feature vectors and bidirectional long short-term memory networks to assign corresponding functional labels for each category. However, in multi-label classification tasks, the issue of imbalanced datasets is a prevalent phenomenon, adding complexity to the classification process. To address the problem of low predictive accuracy resulting from dataset imbalance, the classic Synthetic Minority Oversampling TEchnique (SMOTE) is an oversampling method. In bioinformatics, Lin et al. [22] applied SMOTE to balance skewed benchmark datasets. As a cost-sensitive approach, Yan et al. [23] adopted class-weight optimization to tackle the imbalance issue, achieving significant improvements in predictive performance. Nevertheless, there is room for further optimization in handling diverse imbalanced datasets with the mentioned methods.

To further enhance the predictive accuracy of the model, our study proposes a multi-scale ResNet combined with channel attention model, abbreviated as MSRC, summarized as follows: (i) In the multi-scale ResNet model, we introduce semantic embedding blocks, ResBlock structures, and channel attention (CA). The function of the embedding block is to convert the encoded polypeptide sequence into a feature matrix with multiple semantic information to better support the extraction of local feature information of the polypeptide by the convolutional neural network (CNN) in the ResBlock structure. In addition, the design of the channel attention (CA) mechanism enables the model to dynamically learn the importance of each channel in the convolution during the training process, allowing the model to focus more on task-critical feature channels. This integrated design allows the model to more comprehensively capture semantic information at different scales, thereby improving the expressive ability of features. (ii) To address the imbalance in multi-label datasets, we employ data augmentation methods and a dynamic weighted loss function based on confidence. Data augmentation involves augmenting the minority class peptide samples, aiding in balancing the class distribution. The dynamic weighted loss function, by considering the confidence of each sample, reduces the penalty for highly confident samples, thereby increasing the model's focus on crucial samples and improving predictive accuracy.

This comprehensive approach fully leverages the characteristics of the multi-scale ResNet and enhances the model's generalization through the use of the channel attention mechanism. Experimental validation demonstrates the outstanding performance of our model in multi-label prediction tasks, providing an effective solution for the application of deep learning in the analysis of peptide sequences.

2. Materials and methods

2.1. Datasets

The dataset used in this study is identical to the PrMFTP [23] dataset, which was compiled by Yan et al. in 2021. It comprises a total of 9841 peptide sequences categorized into 21 classes. These classes include Anti-Angiogenic Peptides (AAP), Anti-Bacterial Peptides (ABP), Anti-Cancer Peptides (ACP), Anti-Coronavirus Peptides (ACVP), Anti-Diabetic Peptides (ADP), Anti-Endotoxin Peptides (AEP), Anti-Fungal Peptides (AFP), Anti-HIV Peptides (AHIVP), Anti-Hypertensive Peptides (AHP), Anti-Inflammatory Peptides (AIP), Anti-MRSA Peptides (AMRSAP), Anti-Parasitic Peptides (APP), Anti-Tuberculosis Peptides (ATP), Anti-Viral Peptides (AVP), Blood-Brain Barrier Peptides (BBP), Biofilm Inhibitory Peptides (BIP), Cell-Penetrating Peptides (CPP), Dipeptidyl Peptidase IV Peptides (DPPIP), Quorum Sensing Peptides (QSP), Surface Binding Peptides (SBP), and Tumor Homing Peptides (THP). The number of samples in each category varies, detailed in Table 1. Prior to utilizing these peptide datasets, further filtration is conducted to remove sequences shorter than 5 amino acids or longer than 50 amino acids, as longer peptides are typically more toxic and less stable, while very short peptides exhibit poor sequence activity [24]. Sequences containing non-standard amino acids are also excluded. The datasets are then split into training and testing sets following an 80:20 ratio.

Table 1. Benchmark dataset.

Type	Number	Type	Number
AAP	133	APP	279
ABP	2145	ATP	242
ACP	1043	AVP	711
ACVP	126	BBP	117
ADP	509	BIP	333
AEP	58	CPP	459
AFP	1352	DPPIP	313
AHIVP	101	QSP	220
AHP	917	SBP	104
AIP	2049	THP	651
AMRSAP	168	**Total**	**9841**

2.2. Method framework

Figure 1 illustrates the overall framework for multi-label peptide prediction (MFTP) using a multi-scale ResNet network, divided into two main components. Part I encompasses data mapping and data augmentation. During the data mapping phase, based on the alphabetical order of the 20 natural amino acids (A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, and Y), these amino acids are encoded as natural numbers (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20). Given the potential significant variation in peptide sequence lengths, directly using these sequences of differing lengths as input could result in incorrect data handling by the model. Therefore, peptide sequences shorter than 50 are padded with zeros to ensure that each peptide sequence has the same dimension when entered into the model. Next, data augmentation is performed on the encoded minority peptides to increase the diversity of the sample. Second, we import the encoded data into the MSRC model for effective feature extraction.

Part II describes the structure of the ResNet model, including the embedding layer, ResBlock module, CNN convolution module, CA module, and fully connected layer. Figure 1(A) delineates the overall data processing flow.

Initially, the polypeptide sequences are processed using an embedding layer [25] to transform the normalized 158 × 50 feature vectors into dimensions of 158 × 50 × 256. Here, 158 represents the number of polypeptides in each training batch, 50 denotes the vectors representing the numerical representation of the polypeptide sequences before embedding, and 256 signifies the transformation into high-dimensional semantic matrices through embedding. In this process, the aim is to convert the original discrete sequence into a dense semantic representation to better capture the relationships and characteristics between different amino acids. In order to adapt to the input requirements of the ResNet layer, especially considering that the channel dimensions of data in convolution operations usually need to be placed at specific positions, we performed dimensionality replacement in the layer normalization (Layer Normalization) stage. Second, the peptide data is introduced into 6 ResBlocks [26], each of which includes two 3 × 3, 5 × 5, and 7 × 7 convolutional layers (convs1 and 2), two batch normalization layers (bns1 and 2), and one layer normalization (norm). ReLU activation function and a channel attention (CA) are shown in Figure 1(B).

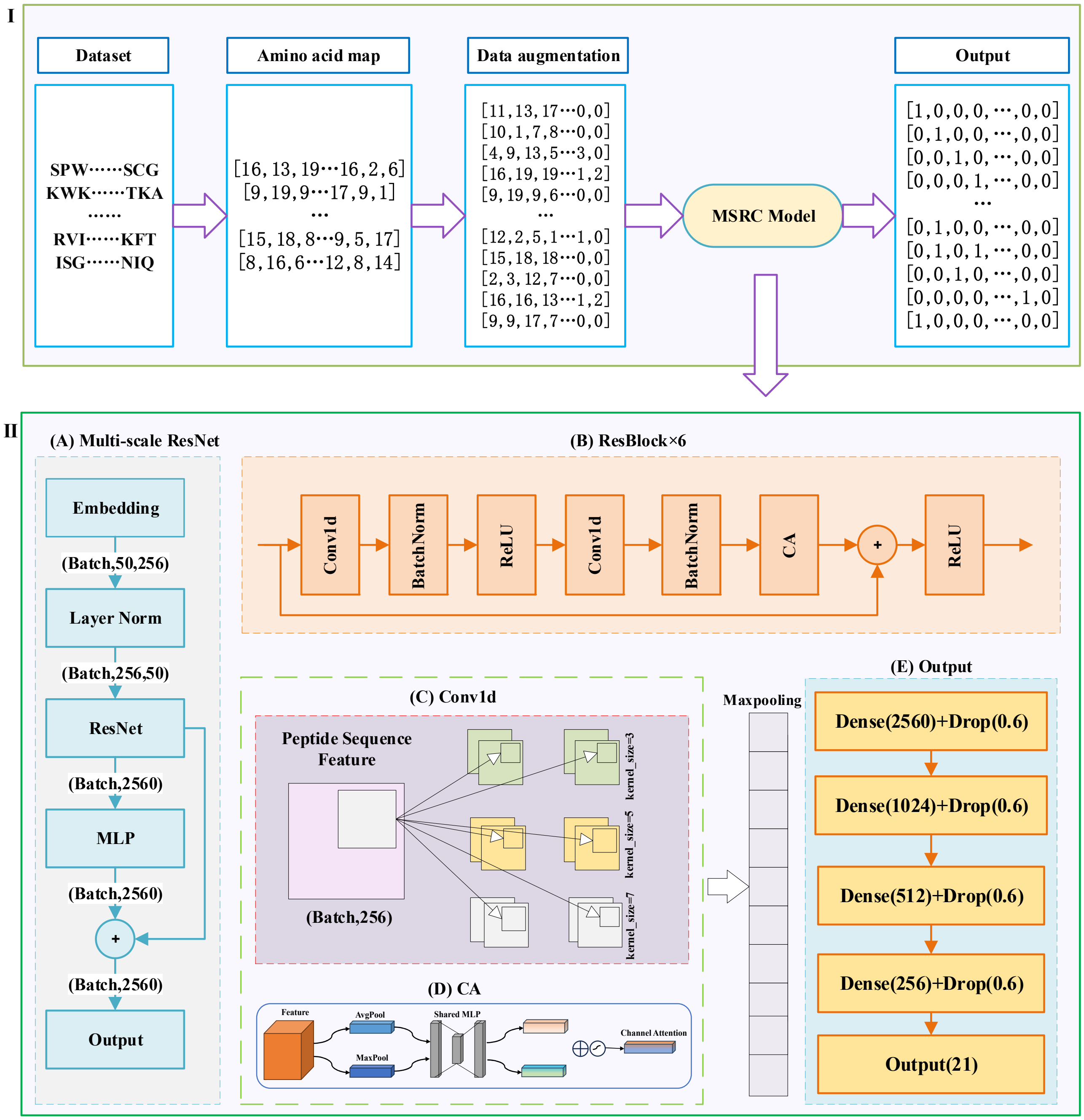


Figure 1. Multi-scale ResNet model architecture diagram.

The two convolutional layers in the ResBlock have kernel sizes of 3, 5, and 7 as shown in Figure 1(C), aiming to extract local features of the polypeptide sequence. The incorporation of the channel attention (CA) mechanism [27] within each ResBlock enables the network to focus on the most information-rich features across the channel dimension. The implementation of the CA mechanism begins with the spatial compression of the peptide feature matrix, transforming it into a one-dimensional vector. This vector is then multiplied element-wise with the original peptide feature matrix to emphasize the original features through weighting, a process that is visually depicted in Figure 1(D). During the spatial compression, two pooling strategies are employed: Average pooling and max pooling. The calculation of channel attention weights can be represented by the following formula (1):

$$A(X) = \sigma(W_2 \cdot ReLU(W_1 \cdot Pool(X)) + W_2 \cdot ReLU(W_1 \cdot Pool(X))). \quad (1)$$

In the formula, W_1 and W_2 represent the weight matrices, and σ denotes the sigmoid activation function. Finally, the channel attention weight vector is calculated by the sigmoid function to dynamically weight each channel, thereby highlighting the feature channels that are crucial for peptide function prediction.

Figure 1(E) illustrates the classification layer, where the vector obtained after max pooling undergoes dimensionality reduction through fully connected layers. Subsequently, the sigmoid function is applied to compute predicted probability scores for various functions associated with the peptide sequence. Overall, Part E processes the features after max pooling through a series of operations, making full use of combined structures such as ResBlock, MLP, CA, and MaxPooling. This synergistic integration endows the entire model with robust capabilities for feature learning and representation.

2.3. Data augmentation and optimization loss function

Prior to embedding the dataset into the model, the imbalance rates of the 21 peptide categories in the training set are computed using formula (2). The imbalance rate is calculated through the following formula:

$$IR_{c_i} = \frac{\sum_{i \neq j} C_i}{(N-1) \cdot C_j}, \quad (2)$$

where N represents the number of categories, C_i and C_j denote the number of samples in the i-th and j-th classes, respectively. Subsequently, classes with an imbalance rate exceeding 2 are identified as minority classes based on a predefined threshold [28]. Next, potential minority regions near the original minority samples are found, and these regions are used to determine pure sub-regions that do not contain majority samples. These pure sub-regions are regarded as possible minority group areas, and finally these pure areas are filled in by synthesizing new samples from the features between the minority class samples.

For the results after data augmentation, we further optimized the loss function. In the context of multi-label problems, where the probabilities of each node are independent of each other, binary cross-entropy is employed as the loss function [29], as shown in formula (3). In order to force the model to pay attention to and optimize the information of the minority class data, this study dynamically adjusts the loss function during the training process based on the confidence of correct classification for each sample (P_{sn}), resulting in the modified loss function shown in formula (4).

$$L_0(Y_{sn}, \hat{Y}_{sn}) = -Y_{sn}\, lg\, \hat{Y}_{sn} - (1 - Y_{sn})\, lg(1 - \hat{Y}_{sn}), \quad (3)$$

$$L_1(Y, \hat{Y}) = \frac{1}{S} \sum_{s=1}^{S} \sum_{n=1}^{N} \sigma[\gamma(1 - 2P_{sn})]\, L_0(Y_{sn}, \hat{Y}_{sn}), \quad (4)$$

where Y_{sn} is the true label of the nth type peptide, $\hat{Y}_{sn}$ is the probability value of predicting the nth type peptide, where $\sigma[\gamma(1-2P_{sn})]$ is the weighting coefficient, σ is the sigmoid function, γ is a constant that smoothens the loss function, and P_{sn} represents the confidence of the classification. It indicates the model's probability of predicting sample s as category n.

2.4. Evaluation metrics

In order to facilitate a comprehensive and fair comparison with other methods, we employed five evaluation metrics [4], including accuracy, coverage, precision, absolute true, and absolute false. These metrics are defined as follows: where N represents the total number of peptide sequences in the dataset, M denotes the number of labels, ∩ and ∪ are set-theoretic intersection and union operations, respectively. L_i represents the subset of true labels for the *i*-th peptide sample, and L_i^* denotes the predicted label subset by the classifier for the i-th sample.

$$Precision = \frac{1}{N}\sum_{i=1}^{N}\frac{\|L_i \cap L_i^*\|}{\|L_i^*\|}, \tag{5}$$

$$Coverage = \frac{1}{N}\sum_{i=1}^{N}\frac{\|L_i \cap L_i^*\|}{\|L_i\|}, \tag{6}$$

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N}\frac{\|L_i \cap L_i^*\|}{\|L_i \cup L_i^*\|}, \tag{7}$$

$$Absolute\ true = \frac{1}{N}\sum_{i=1}^{N}\Delta(L_i, L_i^*), \tag{8}$$

$$Absolute\ false = \frac{1}{N}\sum_{i=1}^{N}\frac{\|L_i \cup L_i^*\| - \|L_i \cap L_i^*\|}{M}, \tag{9}$$

$$\Delta(L_i, L_i^*) = \begin{cases} 1, if\ L_i^*\ is\ identical\ to\ L_i \\ 0, other \end{cases}. \tag{10}$$

3. Results and discussion

3.1. Comparison of multi-label models with common deep learning methods

This section compares models that do not address the issue of imbalanced datasets. In our study, we first conducted a reproducibility study of models described in the literature, including convolutional neural network (CNN), bidirectional long short-term memory network (BiLSTM), and a hybrid model combining CNN, BiLSTM, and multi-head self-attention (MHSA). As shown in Table 2, the hybrid model exhibited superior performance on the test set compared to individual models.

In the field of deep learning, ProtBERT [30] is a pre-trained model based on the Transformer architecture. It is specially designed for protein sequences and has strong representation capabilities when dealing with protein-related tasks. Therefore, we considered the protein pre-trained model ProtBERT and applied it to peptide function prediction. The ProtBERT model exhibited a significant improvement in performance on the test set compared to previous models, with key metrics such as an accuracy of 0.640 and absolute true of 0.578. In order to further improve the prediction effect, we combined ProtBERT with BiLSTM to form a ProtBERT+BiLSTM combination model. This combination model refers to the learning of BiLSTM features from proteins encoded by ProtBert proposed by Zhang et al. [31]. To further enhance predictive performance, we combined ProtBERT with BiLSTM, resulting in the ProtBERT+BiLSTM composite model. This combined model achieved an accuracy of 0.651 and absolute true of 0.579 on the test set. We also explored combining ProtBERT with CNN and BiLSTM. This model demonstrated an accuracy of 0.667 and absolute true of 0.587 on the test set, compared with the combination of CNN, BiLSTM and MHSA, the accuracy of the model increased by 7.2%, and absolute true increased by 5.3%.

In order to further improve the prediction accuracy, this study also explored the basic model with ResNet as the backbone. The table shows that the deep ResNet showed excellent performance. On this basis, we introduced MHSA and CA for comparison. Compared with MSHA, CA is more suitable for ResNet to learn more important channel information. The main indicators are accuracy = 0.680 and absolute true = 0.611. Compared with the combination model of CNN, BiLSTM and MHSA, the Accuracy is improved by 8.5%, and Absolute true is improved by 7.7%.

Based on the experimental results, it is evident that the combined models, as opposed to individual models, can more effectively extract and integrate features from polypeptide sequences, consequently enhancing prediction accuracy. Notably, the multi-scale ResNet and CA combined model exhibits superior performance across all metrics, highlighting its excellence in peptide function prediction tasks. Therefore, this study uses the multi-scale ResNet+CA model for multifunctional peptide prediction.

Table 2. Performance of multifunctional peptide model on test set.

Model	Precision	Coverage	Accuracy	Absolute true	Absolute false
CNN	0.459	0.413	0.411	0.372	0.041
CNN+BiLSTM	0.589	0.543	0.536	0.488	0.036
CNN+BiLSTM+MHSA	0.637	0.623	0.595	0.534	0.033
ProtBERT	0.680	0.689	0.640	0.578	0.037
ProtBERT+BiLSTM	0.701	0.700	0.651	0.579	0.036
ProtBERT+CNN+BiLSTM	0.703	0.708	0.667	0.587	0.036
Multi-scale ResNet	0.700	0.710	0.676	0.601	0.036
Multi-scale ResNet+MHSA	0.709	0.705	0.677	0.598	0.035
Multi-scale ResNet+CA	0.716	0.721	**0.680**	**0.611**	0.034

3.2. Comparative experiments on methods to deal with imbalanced data sets

To verify the effectiveness of the proposed method for handling dataset imbalance in this study, we applied four methods to the PrMFTP model and the MSRC model proposed in this study. In previous studies, two methods [28,29] have been proposed and applied to improve multi-label classification performance (here we refer to these methods as M1 and M2 respectively). Considering the success of M1 and M2, we adopted a combination of these two methods in this study to handle dataset imbalance, abbreviated as M3, for the sake of experimental fairness, We compared the accuracy and absolute true values of different methods between two base models on the same test subset, with each method repeated ten times on both models and the average value taken. As illustrated in Figures 2 and 3. According to the box plot results, the M3 method has the highest box median, both on the PrMFTP model and the MSRC model. Therefore, it shows that the M3 method has a better effect in dealing with the problem of dataset imbalance.

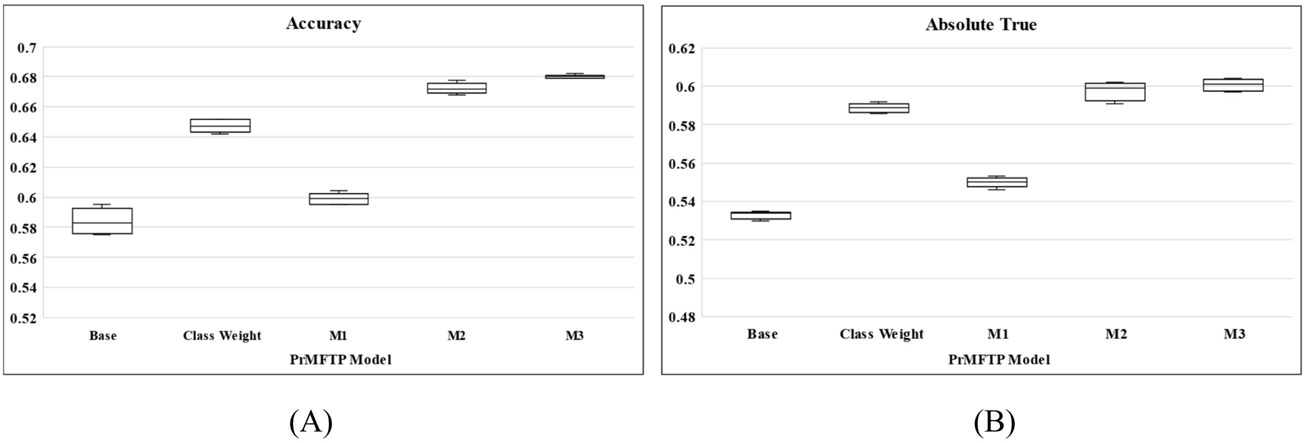


Figure 2. (A) and (B) are the accuracy and absolute true of four methods to deal with dataset imbalance on the PrMFTP model.

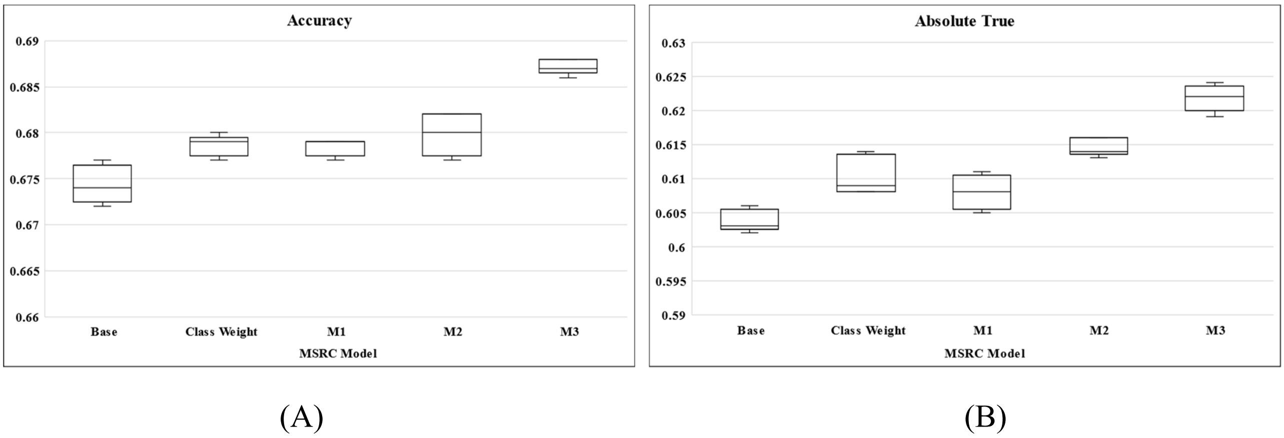


Figure 3. (A) and (B) are the accuracy and absolute true of four methods to deal with dataset imbalance on the MSRC model.

To further validate the superiority of the M3 method, the class weighting approach, as a cost-sensitive method, achieved satisfactory results in the PrMFTP model. However, for extremely imbalanced class issues, such as the dataset used in this study where ABP has 2469 entries while AEP only has 70, even if larger weight values are assigned to AEP before model training, it remains challenging for the model to learn useful information from the AEP samples. Table 3 describes that the M3 method has a better effect on the PrMFTP model than the class weight method, with the most important indicators of accuracy = 0.680 and absolute true = 0.597, which are 3.3% and 0.5% higher than the accuracy = 0.647 and absolute true = 0.592 of the class weight, respectively. Table 4 describes the performance of the M3 method on the MSRC model, which is also better than the class weight method, with accuracy = 0.688 and absolute true = 0.619. Compared to the class weight method, the accuracy = 0.680 and absolute true = 0.603 have increased by 0.8% and 1.6%, respectively.

In general, the M3 method used in this study has excellent performance in handling the problem of dataset imbalance and is significantly better than other methods. The M3 method demonstrates its excellent performance in adjusting sample imbalance between different categories, providing an effective solution to this common problem.

Table 3. Performance of different methods on PrMFTP model.

Method	Precision	Coverage	Accuracy	Absolute true	Absolute false
Base	0.637	0.623	0.595	0.534	0.033
Base+Class weight	0.696	0.664	0.647	0.592	0.030
Base+M1	0.646	0.630	0.604	0.551	0.032
Base+M2	0.720	0.719	0.678	0.594	0.031
Base+M3	0.721	0.721	**0.680**	**0.597**	0.032

Table 4. Performance of different methods on MSRC model.

Method	Precision	Coverage	Accuracy	Absolute true	Absolute false
Base	0.700	0.710	0.676	0.601	0.036
Base+Class weight	0.719	0.720	0.680	0.603	0.037
Base+M1	0.712	0.717	0.679	0.601	0.036
Base+M2	0.720	0.723	0.682	0.607	0.036
Base+M3	0.726	0.728	**0.688**	**0.619**	0.033

3.3. Performance comparison with existing methods

To further demonstrate the powerful capabilities of the multi-scale residual network combined with channel attention mechanism (MSRC) model, we randomly selected 80% of the data from the test set to form a testing subset and compared the MSRC model with several deep learning methods, including MPMABP [21], MLBP [4], and PrMFTP [23]. As shown in Table 5. In comparison, the MSRC model exhibited significant improvements across all metrics, with the primary metric, accuracy, reaching 0.688, and absolute true reaching 0.619. This indicates a significant advantage of the MSRC model in predicting multifunctional peptides. Additionally, with a precision of 0.726 and coverage of 0.728, the model not only accurately predicts the functions of peptides but also covers a broader range of functional categories in its predictions.

Table 5. Performance comparison of MSRC model and other methods.

Model	Precision	Coverage	Accuracy	Absolute true	Absolute false
MPMABP [21]	0481	0.451	0.435	0.378	0.039
MLBP [4]	0.551	0.493	0.489	0.450	0.036
PrMFTP [23]	0.699	0.669	0.651	0.593	**0.031**
MSRC	0.726	0.728	**0.688**	**0.619**	0.033

3.4. The MSRC model is more sensitive to a few types of peptides

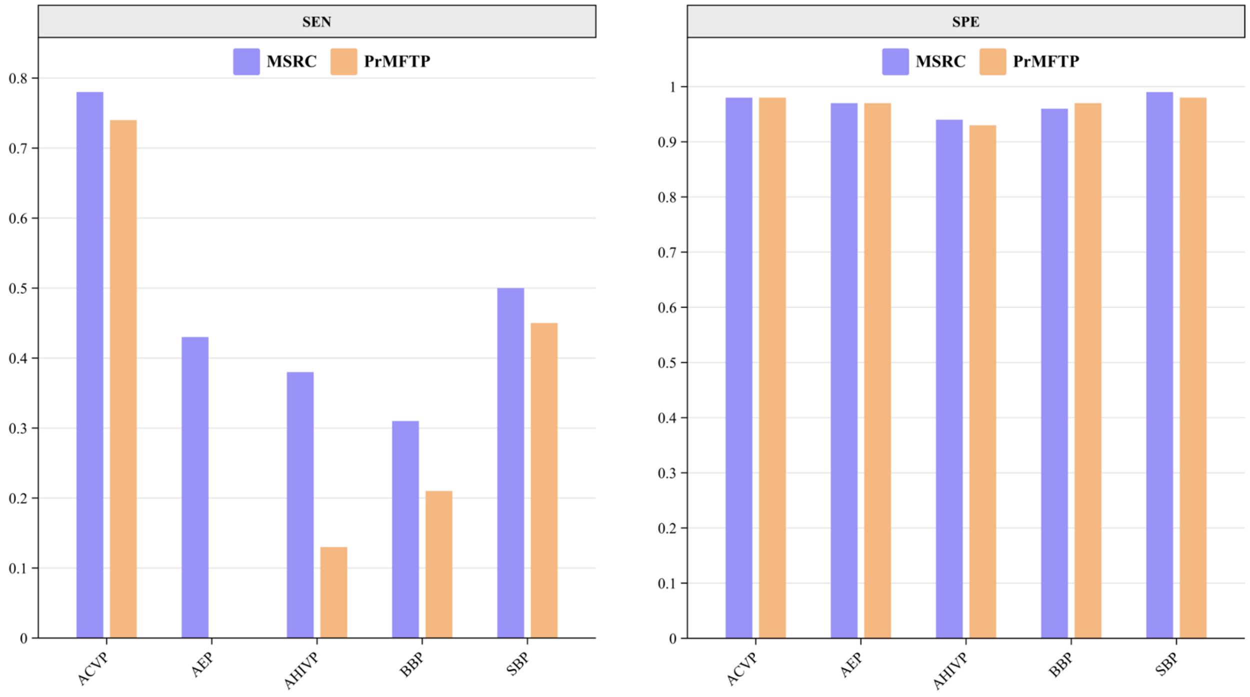
To verify the good predictive accuracy of the proposed model on minority peptide samples, we compared the sensitivity (SEN) and specificity (SPE) of the MSRC model and the PrMFTP model on specific peptide functions (ACVP, AEP, AHIVP, BBP, and SBP) in detail. The calculation formula is as follows:

$$SEN = \frac{TP}{TP+FN}, \tag{11}$$

$$SPE = \frac{TN}{TN+FP}. \tag{12}$$

TP represents the true positives, FN represents the false negatives, TN represents the true negatives, and FP represents the false positives. SEN and SPE are calculated by treating peptides with the specific function as positive samples and peptides without that function as negative samples.

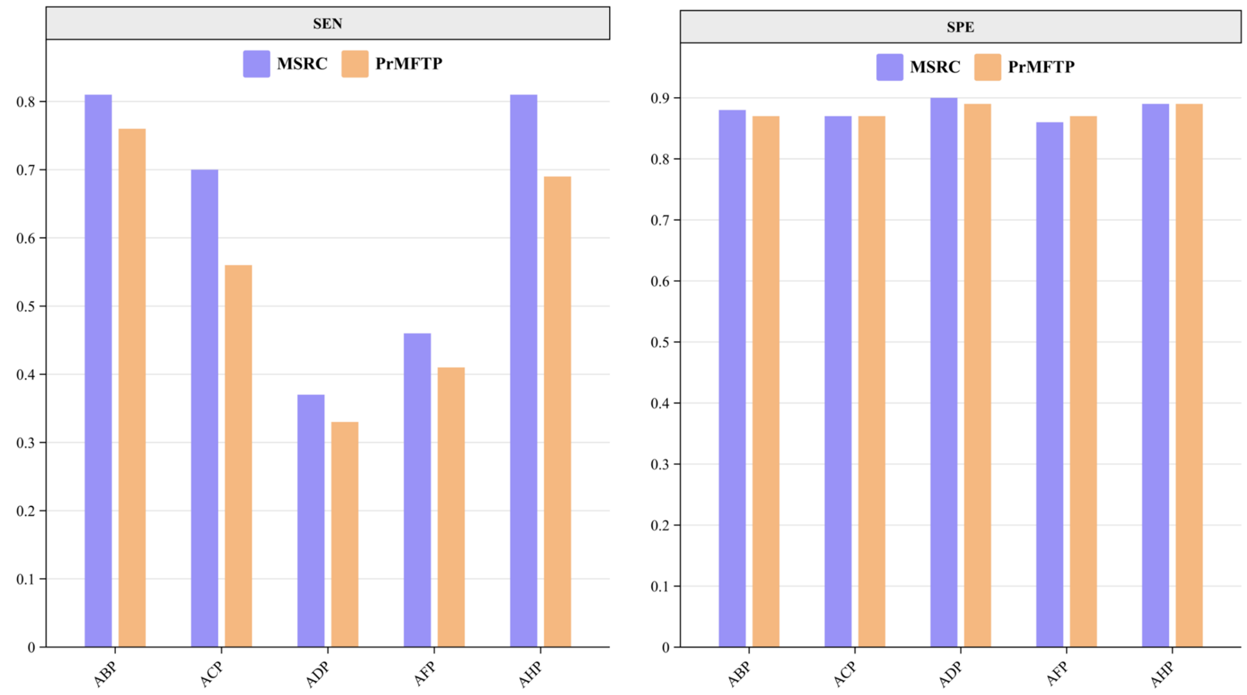
Figure 4 shows the high sensitivity of the MSRC method in predicting the functions of a few peptide classes, especially the AEP, AHIVP, BBP, and SBP classes. MSRC was almost identical to PrMFTP in predicting ACVP sensitivity, but achieved better results than PrMFTP in predicting peptide functions in the other four minority classes. It is worth noting that the PrMFTP model cannot effectively distinguish AEP, while the MSRC model significantly improves the prediction performance of AEP. This shows that MSRC performs well in predicting peptides with these specific functions, successfully capturing the key features of these peptides, giving it a significant advantage in the task. To ensure that the high sensitivity of the MSRC model for minority class samples is not a result of overfitting, we also conducted sensitivity tests for majority class samples with multiple functions, such as ABP, ACP, ADP, AHP, as shown in Figure 5 where the MSRC model also outperformed the PrMFTP method in the majority class peptides.



(A)

(B)

Figure 4. Sensitivity and specificity of minority peptide samples on PrMFTP and MSRC models.



(A) (B)

Figure 5. Sensitivity and specificity of majority peptide samples on PrMFTP and MSRC models.

In the field of deep learning models, especially in tasks involving minority class samples, achieving high sensitivity is crucial for the overall performance of the model. These results further demonstrate that MSRC, by effectively utilizing residual blocks, channel attention mechanisms, and appropriate strategies for handling dataset imbalance, successfully enables the model to learn information about minority class peptides. This makes it excel in tasks that require heightened sensitivity to small-sized peptide categories. The model's sensitivity to these relatively rare peptide functions contributes to improving its reliability in practical applications. In summary, through this comparative study, we emphasize the outstanding performance of the MSRC model in predicting minority class peptide functions and showcase its sensitivity to specific peptide functionalities.

4. Conclusions

Peptides commonly showcase therapeutic properties such as antibacterial, anticancer, antihypertensive, and more. They possess inherent qualities of being potential, safe, and natural organic substances. In this study, we introduce an innovative multi-functional peptide prediction model named MSRC, which is based on the ResNet architecture and incorporates strategies for data augmentation and optimized loss. Compared to existing multi-label methods, MSRC demonstrates satisfactory performance in predicting multi-functional peptides. Key components of the model include multi-scale ResBlocks and channel attention (CA). Through these components, the model extracts information from different weighted channels in the convolutional layers to enhance the perception and capture of complex features. Data augmentation and optimized loss further guide the model's attention towards information from minority class peptides. We conducted experiments on a peptide dataset containing 21 categories, and the results demonstrate that MSRC achieved a significant improvement in predicting multi-functional peptides compared to the PrMFTP method. Its accuracy increased by 3.7%, and the absolute improvement reached 2.6%.

Future research directions will focus on further optimizing the model structure and in-depth exploration of richer peptide sequence data to improve the accurate prediction of peptide function. Additionally, we plan to integrate structural information, function-related features, and physicochemical properties of peptides to enable the model to comprehensively learn sequence information.

Use of AI tools declaration

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

Acknowledgments

This work was supported by the National Key Research and Development Program of China (Nos. 2023YFF1103403 and 2021YFC2801001), the Domestic Science and Technology Cooperation Projects of Shanghai, China (No. 23015820700).

Conflict of interest

The authors declare there are no conflicts of interest.

Reference

1. C. Guntuboina, A. Das, P. Mollaei, S. Kim, A. B. Farimani, Peptidebert: A language model based on transformers for peptide property prediction, *J. Phys. Chem. Lett.*, **14** (2023), 10427–10434. https://doi.org/10.1021/acs.jpclett.3c02398
2. M. Muttenthaler, G. F. King, D. J. Adams, P. F. Alewood, Trends in peptide drug discovery, *Nat. Rev. Drug Discovery*, **20** (2021), 309–325. https://doi.org/10.1038/s41573-020-00135-8
3. E. B. M. Daliri, B. H. Lee, D. H. Oh, Current trends and perspectives of bioactive peptides, *Crit. Rev. Food Sci. Nutr.*, **58** (2018), 2273–2284. https://doi.org/10.1080/10408398.2017.1319795
4. W. Tang, R. Dai, W. Yan, W. Zhang, Y. Bin, E. Xia, et al., Identifying multi-functional bioactive peptide functions using multi-label deep learning, *Briefings Bioinf.*, **23** (2022), bbab414. https://doi.org/10.1093/bib/bbab414
5. Y. Ma, Z. Guo, B. Xia, Y. Zhang, X. Liu, Y. Yu, et al., Identification of antimicrobial peptides from the human gut microbiome using deep learning, *Nat. Biotechnol.*, **40** (2022), 921–931. https://doi.org/10.1038/s41587-022-01226-0
6. Y. Ma, X. Liu, X. Zhang, Y. Yu, Y. Li, M. Song, et al., Efficient mining of anticancer peptides from gut metagenome, *Adv. Sci.*, **10** (2023), 2300107. https://doi.org/10.1002/advs.202300107
7. J. Zhang, Z. Zhang, L. Pu, J. Tang, F. Guo, AIEpred: An ensemble predictive model of classifier chain to identify anti-inflammatory peptides, *IEEE/ACM Trans. Comput. Biol. Bioinf.*, **18** (2020), 1831–1840. https://doi.org/10.1109/TCBB.2020.2968419

8. F. F. Atanaki, S. Behrouzi, S. Ariaeenejad, A. Boroomand, K. Kavousi, BIPEP: Sequence-based prediction of biofilm inhibitory peptides using a combination of NMR and physicochemical descriptors, *ACS Omega*, **5** (2020), 7290–7297. https://doi.org/10.1021/acsomega.9b04119
9. K. Liu, Y. Fu, L. Wu, X. Li, C. Aggarwal, H. Xiong, Automated feature selection: A reinforcement learning perspective, *IEEE Trans. Knowl. Data Eng.*, **35** (2023), 2272–2284. https://doi.org/10.1109/TKDE.2021.3115477
10. P. Agrawal, D. Bhagat, M. Mahalwal, N. Sharma, G. P. S. Raghava, AntiCP 2.0: An updated model for predicting anticancer peptides, *Briefings Bioinf.*, **22** (2021), bbaa153. https://doi.org/10.1093/bib/bbaa153
11. W. Zhang, E. Xia, R. Dai, W. Tang, Y. Bin, J. Xia, PredAPP: Predicting anti-parasitic peptides with undersampling and ensemble approaches, *Interdiscip. Sci.: Comput. Life Sci.*, **14** (2022), 258–268. https://doi.org/10.1007/s12539-021-00484-x
12. B. Manavalan, T. H. Shin, M. O. Kim, G. Lee, AIPpred: Sequence-based prediction of anti-inflammatory peptides using random forest, *Front. Pharmacol.*, **9** (2018), 348997. https://doi.org/10.3389/fphar.2018.00276
13. Y. Han, D. Kim, Deep convolutional neural networks for pan-specific peptide-MHC class I binding prediction, *BMC Bioinf.*, **18** (2017), 585. https://doi.org/10.1186/s12859-017-1997-x
14. Y. Hu, Z. Wang, H. Hu, F. Wan, L. Chen, Y. Xiong, et al., ACME: Pan-specific peptide–MHC class I binding prediction through attention-based deep neural networks, *Bioinformatics*, **35** (2019), 4946–4954. https://doi.org/10.1093/bioinformatics/btz427
15. H. C. Yi, Z. H. You, X. Zhou, L. Cheng, X. Li, T. Jiang, et al., ACP-DL: A deep learning long short-term memory model to predict anticancer peptides using high-efficiency feature representation, *Mol. Ther. Nucleic Acids*, **17** (2019), 1–9. https://doi.org/10.1016/j.omtn.2019.04.025
16. A. Ghulam, F. Ali, R. Sikander, A. Ahmad, A. Ahmed, S. Patil, ACP-2DCNN: Deep learning-based model for improving prediction of anticancer peptides using two-dimensional convolutional neural network, *Chemom. Intell. Lab. Syst.*, **226** (2022), 104589. https://doi.org/10.1016/j.chemolab.2022.104589
17. L. Yu, R. Jing, F. Liu, J. Luo, Y. Li, DeepACP: A novel computational approach for accurate identification of anticancer peptides by deep learning algorithm, *Mol. Ther. Nucleic Acids*, **22** (2020), 862–870. https://doi.org/10.1016/j.omtn.2020.10.005
18. J. M. Conlon, M. Mechkarska, M. L. Lukic, P. R. Flatt, Potential therapeutic applications of multifunctional host-defense peptides from frog skin as anti-cancer, anti-viral, immunomodulatory, and anti-diabetic agents, *Peptides*, **57** (2014), 67–77. https://doi.org/10.1016/j.peptides.2014.04.019
19. H. Fan, W. Yan, L. Wang, J. Liu, Y. Bin, J. Xia, Deep learning-based multi-functional therapeutic peptides prediction with a multi-label focal dice loss function, *Bioinformatics*, **39** (2023), btad334. https://doi.org/10.1093/bioinformatics/btad334
20. H. Lv, K. Yan, B. Liu, TPpred-LE: Therapeutic peptide function prediction based on label embedding, *BMC Biol.*, **21** (2023), 238. https://doi.org/10.1186/s12915-023-01740-w
21. Y. Li, X. Li, Y. Liu, Y. Yao, G. Huang, MPMABP: A CNN and Bi-LSTM-Based method for predicting multi-activities of bioactive peptides, *Pharmaceuticals*, **15** (2022), 707. https://doi.org/10.3390/ph15060707

22. W. Lin, D. Xu, Imbalanced multi-label learning for identifying antimicrobial peptides and their functional types, *Bioinformatics*, **32** (2016), 3745–3752. https://doi.org/10.1093/bioinformatics/btw560
23. W. Yan, W. Tang, L. Wang, Y. Bin, J. Xia, PrMFTP: Multi-functional therapeutic peptides prediction based on multi-head self-attention mechanism and class weight optimization, *PLoS Comput. Biol.*, **18** (2022), e1010511. https://doi.org/10.1371/journal.pcbi.1010511
24. H. Kim, J. H. Jang, S. C. Kim, J. H. Cho, De novo generation of short antimicrobial peptides with enhanced stability and cell specificity, *J. Antimicrob. Chemother.*, **69** (2014), 121–132. https://doi.org/10.1093/jac/dkt322
25. E. Vušak, V. Kužina, A. Jović, A survey of word embedding algorithms for textual data information extraction, in *2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO)*, IEEE, (2021), 181–186. https://ieeexplore.ieee.org/document/9597076
26. F. Ge, Y. Zhang, J. Xu, A. Muhammad, J. Song, D. Yu, Prediction of disease-associated nsSNPs by integrating multi-scale ResNet models with deep feature fusion, *Briefings Bioinf.*, **23** (2022), bbab530. https://doi.org/10.1093/bib/bbab530
27. Z. Zhao, J. Gui, A. Yao, N. Q. K. Le, M. C. H. Chua, Improved prediction model of protein and peptide toxicity by integrating channel attention into a convolutional neural network and gated recurrent units, *ACS Omega*, **7** (2022), 40569–40577. https://doi.org/10.1021/acsomega.2c05881
28. T. Zhu, X. Liu, E. Zhu, Oversampling with reliably expanding minority class regions for imbalanced data learning, *IEEE Trans. Knowl. Data Eng.*, **35** (2023), 6167–6181. https://ieeexplore.ieee.org/document/9773030
29. D. Wang, H. Yu, G. Fan, Facial action unit recognition algorithm based on deep learning (in Chinese), *J. East China Univ. Sci. Technol. (Nat. Sci. Ed.)*, **46** (2020), 269–276. https://doi.org/10.14135/j.cnki.1006-3080.20190107003
30. A. Elnaggar, M. Heinzinger, C. Dallago, G. Rihawi, Y. Wang, L. Jones, et al., ProtTrans: Towards cracking the language of life's code through self-supervised deep learning and high performance computing, preprint, arXiv:2007.06225.
31. Y. Zhang, G. Zhu, K. Li, F. Li, L. Huang, M. Duan, et al., HLAB: Learning the BiLSTM features from the ProtBert-encoded proteins for the class I HLA-peptide binding prediction, *Briefings Bioinf.*, **23** (2022), bbac173. https://doi.org/10.1093/bib/bbac173