SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

November 2024, Vol. 67, Iss. 11, 212105:1–212105:12
https://doi.org/10.1007/s11432-024-4147-8

TPpred-SC: multi-functional therapeutic peptide prediction based on multi-label supervised contrastive learning

Ke YAN, Hongwu LV, Jiangyi SHAO, Shutao CHEN & Bin LIU*

School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081*, China*

Received 13 January 2024/Revised 15 April 2024/Accepted 18 May 2024/Published online 22 October 2024

Abstract Therapeutic peptides contribute significantly to human health and have the potential for personalized medicine. The prediction for the therapeutic peptides is beneficial and emerging for the discovery of drugs. Although several computational approaches have emerged to discern the functions of therapeutic peptides, predicting multi-functional therapeutic peptide types is challenging. In this research, a novel approach termed TPpred-SC has been introduced. This method leverages a pretrained protein language model alongside multi-label supervised contrastive learning to predict multi-functional therapeutic peptides. The framework incorporates sequential semantic information directly from large-scale protein sequences in TAPE. Then, TPpred-SC exploits multi-label supervised contrastive learning to enhance the representation of peptide sequences for imbalanced multi-label therapeutic peptide prediction. The experimental findings demonstrate that TPpred-SC achieves superior performance compared to existing related methods. To serve our work more efficiently, the web server of TPpred-SC can be accessed at http://bliulab.net/TPpred-SC.

Keywords therapeutic peptide prediction, multi-label classification, pretrained protein language model, multi-label supervised contrastive learning

1 Introduction

Therapeutic peptides are short amino acid sequences, that generally serve vital functions in human physiology, including modulating biological processes, regulating hormone levels, stimulating tissue regeneration, and combating infectious agents [1–3]. As powerful tools in the field of medicine, peptides offer immense benefits to human health and are promising for the treatment of various diseases [1, 4–7]. The therapeutic peptides are categorized based on functions, such as anti-viral, anti-cancer, anti-inflammatory, and anti-microbial, etc. [8]. With the development of protein synthesis technology in the post-genome era, therapeutic peptides with two or more functionalities have been discovered [9]. Therefore, predicting multi-functional therapeutic peptides (MTPs) is meaningful for treating diseases.

Over the past decades, computational methods have gained significant traction for predicting MTPs. These methods can be broadly categorized into two groups: (i) conventional machine learning-based methods and (ii) deep learning-based methods [10–12].

Conventional machine learning-based methods generally have two stages: sequence feature engineering and taxonomy-based predictors. In the first stage, several methods construct hand-crafted manual features based on different properties of the peptide sequence, such as the k-mer [13], Pse-AAC [14], etc. For the second stage, researchers utilized different primary predictors for therapeutic peptide identification, such as random forest (RF) [15], linear regression (LR) [16], and support vector machine (SVM) [17–20], etc. PEPred-Suite [8] extracts 99 feature descriptors and selects the most informative ones using redundancy and maximal relevance (mRMR) [21] to train an RF classifier for each peptide type. PPTPP [22] is another method that utilizes RF to recognize therapeutic peptides. In this approach, hundreds of physicochemical property-related descriptors are extracted, and feature selection is

* Corresponding author (email: bliu@bliulab.net)

required. TPpred-ATMV [23] utilizes multi-view tensor learning to learn the correlation between different properties feature encoding and predict eight function types of therapeutic peptides in a unified framework.

Deep learning-based methods have been widely used in bioinformatics [24–26]. Compared with conventional machine learning-based methods, deep learning methods construct high-latent features from sequences based on limited peptide knowledge [27]. In recent years, numerous methods based on different deep learning frameworks have been proposed for therapeutic peptide prediction [9, 10, 28–33], such as convolutional neural networks (CNN) [34], gated recurrent units (GRU) [35] and attention mechanism [36]. MLBP [10] and PrMFTP [9] are proposed for MTPs prediction, depending on one-hot encoding [37]. TPpred-LE [38] extracts a position-specific scoring matrix (PSSM) [39] and utilizes attention to capture relationship information. Moreover, contrastive learning models have been widely used in representation learning in both unsupervised or supervised learning. The primary goal of contrastive learning is to learn representations of peptides so that similar functional peptide sequences are mapped close together in a learned feature space. Conversely, dissimilar functional peptide sequences are mapped far apart.

In recent years, pretrained protein language models (pLM) have been widely used in bioinformatics to extract generalized sequence semantic features directly from sequences without using structural information [40–44]. The pLM models are generally based on Transformer architecture and perform self-supervised learning on large-scale unlabelled protein sequence corpora. Teufel et al. have shown the power of pLM on signal peptide prediction [45]. The AMPDeep [46], Deep-AFPpred [47], and LM-Pred [48] models utilized ProtTrans [41] to embed the peptide sequences for mono-functional therapeutic peptide prediction. BERTMHC [49] adopts TAPE [40] for predicting MHC-peptide class II interactions. BERT4Bitter utilizes the bidirectional encoder representation from transformers (BERT) model to predict bitter peptides [50].

Despite numerous previous methods, they still exhibit some limitations. (i) Conventional machine learning methods extract or select the hand-crafted features, which is time-consuming, inefficient, and requires prior professional knowledge. (ii) The MTPs dataset suffers from an imbalance issue, where the number of peptides associated with majority functions is significantly greater than those associated with minority functions. The existing deep learning-based methods make it challenging to construct discriminative feature representations of peptides from minority functions due to the limited scale of training data.

In this study, we developed TPpred-SC, a novel method for predicting MTPs focusing on representation learning. TPpred-SC utilizes the pretrained protein language model TAPE [25] to embed peptide sequences. We have implemented multi-label supervised contrastive learning [34, 35, 51] as a means to augment representation learning. Unlike traditional semi-supervised contrastive learning methods, our approach extensively capitalizes on label information, fostering a discriminative representation [34, 36]. This adaptation contrasts with conventional methodologies by comprehensively integrating labeled data, refining the learning process, and enhancing the quality of obtained representations. Furthermore, data augmentation strategies are used to improve the generalization ability of the proposed method. To the best of our knowledge, TPpred-SC is the first approach to apply supervised contrastive learning in MTP prediction. Experimental results demonstrate that TPpred-SC achieves superior performance compared to existing therapeutic peptide prediction methods. Additionally, a user-friendly webserver for TPpred-SC has been developed and is available at http://bliulab.net/TPpred-SC.

2 Materials and methods

2.1 Benchmark dataset

In this work, we adopt the benchmark dataset assembled by previous work [38], involving 15 distinct therapeutic peptide functions: AAP, ABP, ACP, AFP, AHTP, AIP, AMP, APP, AVP, CCC, CPP, DDV, PBP, QSP and TXP. This multi-functional therapeutic peptide benchmark dataset contains 10237 unique sequences, and each peptide was assigned at least one therapeutic function. Peptide sequences with over 90% similarity [28, 52–54] within each function subset were eliminated by leveraging CD-HIT [25, 39]. Additionally, it has been partitioned into a training dataset, validation dataset, and independent test dataset for the experiment. For more details, refer to [38]. The benchmark dataset can be accessed at:

http://bliulab.net/TPpred-SC/data.

2.2 Overview of the TPpred-SC

The flowchart of TPpred-SC is depicted in Figure 1. TPpred-SC comprises two training stages. (i) In the first stage, TPpred-SC learns discriminative representation by multi-label supervised contrastive learning using the training dataset. Initially, data augmentation approaches were employed to generate additional training samples and improve the generalization of TPpred-SC. Subsequently, the pLM method BERT pretrained through TAPE [40] was applied to extract sequential semantic information from the augmented sequences. The subsequent MLP module was constructed to map the BERT representation into the output space corresponding to the contrastive loss. Finally, multi-label supervised contrastive learning loss was applied to optimize the model. (ii) In the second stage, we fine-tuned the model by employing the loss related to multi-label classification over a few epochs. The parameters in the BERT module were initialized with the learned parameters from the first stage. A new MLP was reconstructed for classification purposes. Ultimately, TPpred-SC can accurately predict therapeutic peptide functions after completing these two training stages.

2.3 First stage: multi-label supervised contrastive learning

Contrastive learning is a robust technology suitable for feature representation learning tasks and has been widely used in unsupervised learning tasks [55]. In recent years, contrastive learning has been applied in supervised learning classification downstream tasks [56, 57]. In this section, we utilized the multi-label supervised contrastive learning model to learn discriminative features by reducing the distance between similar sequences and increasing the distances of the dissimilar ones apart from each other.

The multi-label contrastive learning of the TPpred-SC contains three modules: data augmentation approaches, the pretrained BERT language model, and the multi-label contrastive learning loss function module. Specifically, data augmentation approaches are utilized to generate more samples and enhance the generalization of the training process. Then, the augmented sequences are translated into digital feature representation using the pretrained protein language model TAPE, which incorporates the semantic information of sequences. Finally, we applied the contrastive learning loss function, which plays a significant role in representation learning. This technology is pivotal for reducing the distance in the feature space between samples sharing the same or similar labels, while simultaneously increasing the distance between samples with different labels.

(1) Data augmentation approaches. Data augmentation approaches have been widely used in contrastive learning to expand training samples, generate the related positive samples, and improve the generalization of the model [56, 58, 59]. We utilized two simple random replacement augmentation strategies. (i) Dictionary replacement. According to a replacement dictionary:{[A,V],[S,T],[F,Y],[K,R],[C,M],[D,E],[N,Q],[V,I]}, each amino acid in the sequence follows a probability p augmentation strategy of being replaced by another amino acid to simulate the amino acid mutation [60, 61]. (ii) Alanine replacement. Each amino acid in the sequence follows a probability p of being replaced by alanine (A) [60]. In this study, p is set to 0.1.

We define a mini-batch of the training set with N samples as $\{s_1, s_2, \ldots, s_N\}$, where s_i represents a peptide sequence. Each sequence underwent data augmentation twice [56, 59], with each process independently and randomly selected from the two approaches above. As a result, a batch training set is expanded into $\{s_1^{'}, s_2^{'}, \ldots, s_{2N}^{'}\}$, where s_i is augmented to $s_{2i-1}^{'}$ and $s_{2i}^{'}$.

(2) Pretrained protein language model. TPpred-SC utilizes a pretrained protein language model based on the BERT [62, 63] network presented by TAPE [40] to embed amino acid sequences. TAPE trained BERT on a dataset of over 31 million protein sequences using self-supervised learning to incorporate the rich amino acid contextual information. Two unique tokens $\langle CLS \rangle$ and $\langle SEP \rangle$ were added to the beginning and end of the raw sequences following the typical NLP BERT [62]. $\langle PAD \rangle$ tokens were added to pad short sequences to the maximum sequence length within the current training batch. Finally, each amino acid was encoded into a fixed vector of 768 dimensions. We generated the sequence-level representation by computing the average of all the residue-level representation vectors for a given sequence. As a result, an augmented peptide sequence, denoted by $s_i^{'}$, was represented as a 768-dimensional vector through the BERT layers.

$$h_i = \mathrm{BERT}(s_i^{'}), i \in [1, \ldots, 2N]. \tag{1}$$

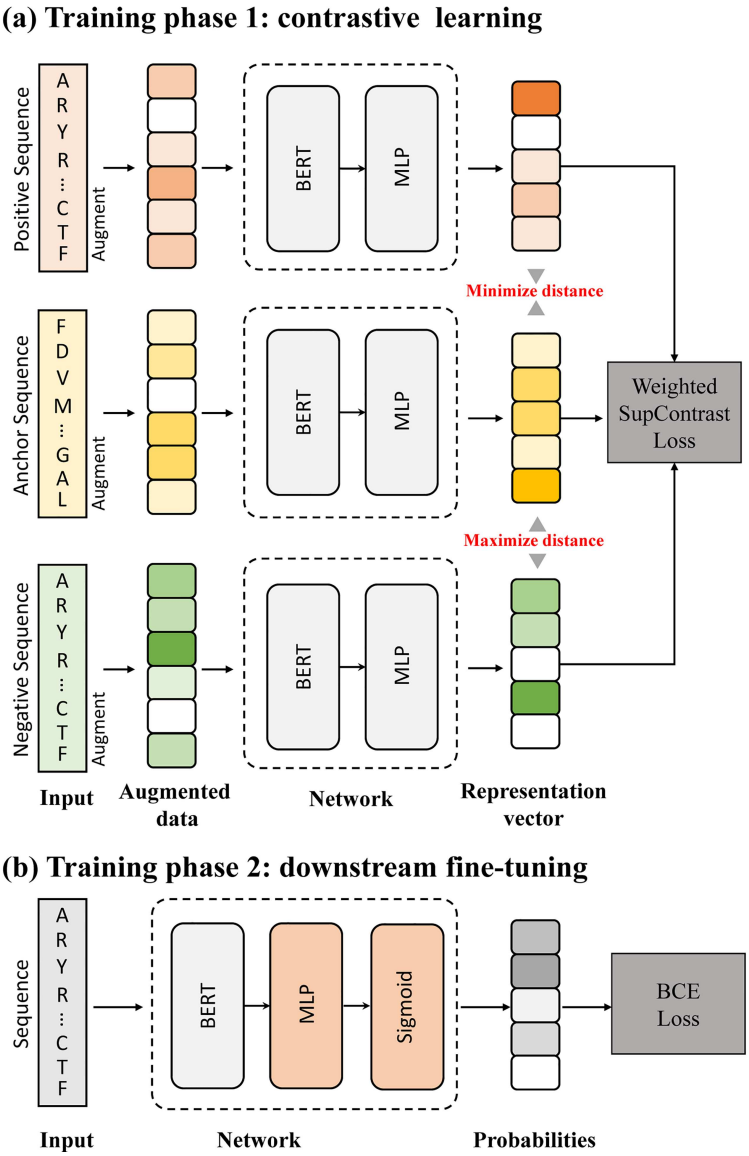

Figure 1 (Color online) The framework of TPpred-SC. (a) The contrastive learning phase. The sample is referred to as the anchor sequence. Positive samples are the sequences that have a similar label set to the anchor sample. Negative samples have a dissimilar label set compared to the anchor sample. They are augmented and input into the network to calculate the loss value. Note that the weights of the network are shared within all training inputs. (b) The downstream fine-tuning phase. The BERT in (b) is inherited from (a). Note that the MLP layers in (a) and (b) have distinct dimensions. The former is used for representation learning and the latter is used for classification.

A multilayer perceptron (MLP) [64] layer follows after BERT is applied to map the representations to the output space corresponding to the contrastive loss:

$$z_i = \mathrm{MLP}(h_i) = \mathbf{W}^{(2)}\mathrm{ReLU}(\mathbf{W}^{(1)}h_i), \tag{2}$$

where $\mathbf{W}^{(1)}$ and $\mathbf{W}^{(2)}$are projection matrices and ReLU( ) is the activation function.

(3) Multi-label contrastive learning loss. TPpred-SC categorizes training samples into three types. Sample i is considered the anchor sample in a training batch when calculating the associated loss. Additionally, samples with similar functions are regarded as positive, while the remaining ones are considered negative. In each training batch, every sample takes turns serving as the anchor.

The loss of TPpred-SC is based on the unsupervised contrastive learning method SimCLR [45]. The loss of SimCLR can be formulated as [59]:

$$\mathcal{L}^{\mathrm{Sim}} = \sum_{i \in I} \mathcal{L}_i^{\mathrm{Sim}} = -\sum_{i \in I} \log \frac{\exp(z_i \cdot z_{j(i)}/\tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a/\tau)}, \tag{3}$$

where $i \in I = \{1, 2, \ldots, 2N\}$ denotes the index of an arbitrary augmented sample within a training batch comprising N pairs of samples. Meanwhile, $j(i)$ signifies the index of the other augmented sample derived from the same source sample. $A(i) = I \backslash \{i\}$ denotes the index set excluding i. τ denotes a temperature parameter, influencing the control of penalty strength on the hard negative samples [65]. Specifically, contrastive learning loss with a small temperature tends to penalize the hardest negative samples more, whose representation vectors are similar to the concerned sample. In this work, τ is set to 0.07. The concerned sample with index i is the anchor sample. The sample with index $j(i)$ is the positive sample associated with i, and the others are viewed as negative samples. In other words, each anchor sample has only one positive sample but $(2N-1)$ negative samples in SimCLR.

SimCLR is generally used for pretrained on massive unlabelled training data. However, it neglects label information. SupCons [56] is proposed for supervised contrastive learning, integrating the labels into loss function based on SimCLR, allowing us to leverage label information effectively. It is formulated as [56]:

$$\mathcal{L}^{\mathrm{Sup}} = \sum_{i \in I} \mathcal{L}_i^{\mathrm{Sup}} = -\frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(z_i \cdot z_p/\tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a/\tau)}, \tag{4}$$

where $P(i) = \{p | p \in A(i) \bigwedge y_p = y_i\}$ represents the positive sample set where the elements have the same label as i. Compared to SimCLR, SupCons expands the positive set for each anchor sample according to its label. However, SupCons is designed for multi-class classification, where a sample has at most one label.

To apply the supervised contrastive learning to muti-label classification tasks, MultiSupCon [66] modified SupCons by weighting the loss according to the similarity of labels between different sample pairs [66]:

$$\mathcal{L}^{\mathrm{ml-Sup}} = \sum_{i \in I} \mathcal{L}_i^{\mathrm{ml-Sup}} = \sum_{i \in I} -\frac{1}{Q(i)} \sum_{p \in Q(i)} w_{i,p} \log \frac{\exp(z_i \cdot z_p/\tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a/\tau)}, \tag{5}$$

$$w_{i,j} = \frac{||y_i \cap y_j||}{||y_i \cup y_j||}, \tag{6}$$

where y_i denotes the binary label vector for sample i. $w_{i,j}$ is the Jaccard score [67] between samples i and j, which is used to measure the similarities of their labels. $Q(i) = \{p | w_{i,p} > \theta_w\}$ defines the positive sample set for anchor i, where the Jaccard score between i and p are larger than the threshold θ_w. In this work, θ_w is set to 0.1.

When minimizing $\mathcal{L}^{\mathrm{ml-Sup}}$, the samples with higher label similarity tend to be clustered closer in the feature space. Conversely, the dissimilarity samples with lower Jaccard scores tend to be pushed further from the anchor sample to obtain discriminative decision boundaries, as shown in Figure 2.

2.4 Second stage: downstream fine-tuning framework of the TPpred-SC

After training in the first stage of TPpred-SC, a discriminative and robust pretrained model is obtained. In the second stage of the TPpred-SC, we proposed a new MLP layer by adjusting the representation received from BERT for our multi-label classification, which projects the representations from BERT to the classification output space. Binary Cross Entropy (BCE) loss [68] is used as the training loss in this stage [68]:

$$\mathcal{L}^{\mathrm{BCE}} = \sum_{i=1}^{N} \mathcal{L}_i^{\mathrm{BCE}} = \sum_{i=1}^{N} \sum_{j=1}^{C} y_{ij} \cdot \log \hat{y}_{ij} + (1 - y_{ij}) \log(1 - \hat{y}_{ij}), \tag{7}$$

where C denotes the total number of functions. $y_{ij} \in \mathbb{R}^C$ is the ground truth label, and $\hat{y}_{ij} \in \mathbb{R}^C$ is the prediction probability of sample i corresponding to function j.

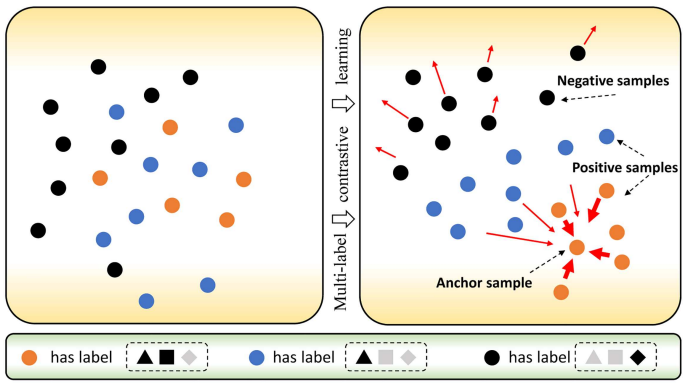

Figure 2 (Color online) The illustration of the multi-label contrastive learning model in the TPpred-SC. We randomly consider one of the orange samples as the anchor. There are three situations between two samples in terms of their label information: same, similar, and different. The orange samples are the same, and the Jaccard scores between them and the anchor sample are 1. The blue samples are noted as similar, of which the Jaccard scores between them and the anchor sample range from θ_w to 1. The different samples are outside the positive set $Q(i)$. The arrows indicate the optimization direction based on the anchor sample, with the thickness of the arrows representing the corresponding weights $w_{i,p}$. Note that the positive and negative samples are relative to the anchor sample. In other words, when a black sample is marked as the anchor, the blue and orange ones are regarded as negatives.

2.5 Model implementation

As depicted in Figure 1, TPpred-SC applies two training stages to fit the model. In the first stage, the multi-label supervised contrastive learning model is conducted to generate a discriminative and robust representation over 300 epochs. The initial learning rate is set to 0.01 and decays to 0.001 in the 100th epoch and 0.0001 in the 200th epoch, respectively. The size of the hidden layer of the MLP is 2048, and its output dimension is set to 128, and the SGD optimizer with momentum is 0.9 [56]. As described in [56], a larger batch size can lead to better performance in contrastive learning. Therefore, we employed a batch size of 128, which has reached the upper limit of our 3090 GPU training platform.

In the second stage, we constructed a new MLP layer with an output dimension equal to the number of categories 15, and fine-tuned the whole model with multi-label classification learning within 10 epochs. The learning rate is 0.0001 and the batch size is 64. AdamW is used to enhance the generalization of the model [69]. The final model is determined by selecting the model with the best accuracy on the validation dataset during the fine-tuning step.

2.6 Evaluation metrics

We leverage example-level accuracy ($\mathrm{ACC}_{\mathrm{example}}$) and label-level F1-score ($F1_{\mathrm{label}}$) to assess the overall performance of multi-functional therapeutic peptide prediction as described in [38]:

$$\mathrm{ACC}_{\mathrm{example}} = \frac{1}{N}\sum_{i=1}^{N}\frac{||y_i \cap \hat{y}_i||}{||y_i \cup \hat{y}_i||} \tag{8}$$

$$F1_{\mathrm{label}} = \frac{1}{C}\sum_{i=1}^{C} F1-\mathrm{measure}_i \tag{9}$$

where y_i represents the ground truth label set and $\hat{y}_i$ represents the predicted label set. C denotes the number of all function types. $(F1-\mathrm{measure}_i)$ is the harmonic mean of precision and recall for each function.

Furthermore, binary classification metrics, including the Area Under the ROC curve (AUC) [70, 71], $F1$-measure [72], and Matthews's correlation coefficient (MCC) [73–76], are also used to investigate the prediction ability in a one-vs-rest manner.

Table 1 The performance of our method based on four different training paradigms on the independent test dataset

Method	$\mathrm{ACC}_{\mathrm{example}}$	$F1_{\mathrm{label}}$
Scratch	0.295	0.068
Fine-tuning	0.567	0.431
SimCLR	0.510	0.287
The proposed method	**0.587**	**0.492**

Table 2 The ablation study about different combinations of data augmentation strategies for TPpred-SC on the independent test dataset

Model	Data augmentation		Performance	
	Dictionary	Alanine	$\mathrm{ACC}_{\mathrm{example}}$	$F1_{\mathrm{label}}$
This study	✔	✔	**0.587**	**0.492**
A	✘	✔	0.586	0.402
B	✔	✘	0.586	0.421
C	✘	✘	0.569	0.378

3 Results and discussion

3.1 Analysis of supervised contrastive learning method for MTPs prediction

In this section, we explore the performance of four different training paradigms for the same network, including Scratch, Fine-tuning, SimCLR, and the proposed method. The Scratch initialized the network randomly and trained it from scratch without any pretraining process. The Fine-tuning utilized the pretrained BERT model [40] and the network was fine-tuned on our benchmark training dataset. SimCLR [59], based on the unsupervised contrastive learning algorithm, and the proposed method, based on the supervised contrastive learning algorithm, were performed in the first training stage, respectively. Then, they were fine-tuned on the benchmark training set as the second training stage. The rest of the training frameworks were the same in the four methods to evaluate the performance equally.

The results are presented in Table 1, indicating that: (i) The Scratch model exhibits poorer performance than the other pretrained based models, underscoring the superiority of the sequential semantic information based on the pretrained BERT model. (ii) The SimCLR based model performs less effectively than the Fine-tuning based model and the proposed method, revealing that the unsupervised contrastive learning tends to have negative impacts on the MTPs function prediction task. In the first training stage, the loss function in the SimCLR based model considers another augmented sequence from the same raw sequence as the positive sample for an augmented sequence. Conversely, all the other sequences are viewed as negative samples, regardless of whether they share the same labels, ignoring label information. (iii) The proposed method outperforms the other training paradigms in terms of the $\mathrm{ACC}_{\mathrm{example}}$ and $F1_{\mathrm{label}}$. This is because the proposed method incorporates the multi-label information into the contrastive learning phase and utilizes the sequential semantic information based on the BERT model. Therefore, the proposed method incorporates more discriminative information and improves the prediction of MTPs.

3.2 Performance of data argumentation approaches

In this section, we confirm the effectiveness of different data augmentation approaches. Data augmentation approaches are generally used to create more training data to improve the generalization ability of the model. The results in Table 2 demonstrate that: (i) data augmentation approaches that randomly combine the dictionary replacement approach and alanine replacement approach achieve the highest performance; (ii) the model without using any data augmentation strategy (C row in Table 2) performs worse compared to the other models. Due to the data argumentation approaches providing several diverse variations of the peptides [60, 77], the proposed method achieves more robust and stable performance, demonstrating that the data augmentation strategies play an important role in MTPs prediction.

3.3 Performance comparison among different binary classification-based tools for therapeutic peptide function prediction

In this section, we analyze the effectiveness of TPpred-SC alongside other commonly utilized binary classification-based therapeutic peptide prediction methods, including PEPred-Suite [8], PPTPP [22], and

Table 3 Evaluating multiple methods in predicting eight therapeutic peptide functions using the independent test dataset

Function	Method	AUC	MCC	$F1_{\text{label}}$
AAP	PEPred-Suite[a)]	0.577	0.02	0.03
	PPTPP[a)]	0.604	0.037	0.033
	TPpred-ATMV[a)]	0.583	0.009	0.027
	TPpred-SC	**0.813**	**0.487**	**0.724**
ABP	PEPred-Suite[a)]	0.744	0.261	0.367
	PPTPP[a)]	0.732	0.261	0.365
	TPpred-ATMV[a)]	0.731	0.256	0.36
	TPpred-SC	**0.801**	**0.297**	**0.407**
ACP	PEPred-Suite[a)]	0.56	0.03	0.155
	PPTPP[a)]	0.625	0.049	0.162
	TPpred-ATMV[a)]	0.662	0.096	0.183
	TPpred-SC	**0.786**	**0.364**	**0.391**
AIP	PEPred-Suite[a)]	0.363	-0.19	0.18
	PPTPP[a)]	0.386	-0.06	0.168
	TPpred-ATMV[a)]	0.369	-0.25	0.196
	TPpred-SC	**0.938**	**0.618**	**0.675**
AVP	PEPred-Suite[a)]	0.382	-0.129	0.147
	PPTPP[a)]	0.404	-0.11	0.169
	TPpred-ATMV[a)]	0.394	-0.118	0.135
	TPpred-SC	**0.847**	**0.502**	**0.555**
CPP	PEPred-Suite[a)]	0.813	0.152	0.142
	PPTPP[a)]	0.814	0.14	0.139
	TPpred-ATMV[a)]	0.815	0.152	0.139
	TPpred-SC	**0.925**	**0.622**	**0.639**
PBP	PEPred-Suite[a)]	0.907	0.153	0.069
	PPTPP[a)]	0.829	0.119	0.07
	TPpred-ATMV[a)]	0.836	0.153	0.086
	TPpred-SC	**0.942**	**0.678**	**0.667**
QSP	PEPred-Suite[a)]	0.835	0.113	0.043
	PPTPP[a)]	0.815	0.08	0.033
	TPpred-ATMV[a)]	0.772	0.054	0.027
	TPpred-SC	**0.902**	**0.576**	**0.5**

a) The results are reported by [38].

TPpred-ATMV [23] in a "one-vs-rest" form. The results are presented in Table 3, showing that TPpred-SC surpasses the binary classification-based methods in almost all functions and metrics, demonstrating the stability and accuracy of our proposed method. TPpred-SC constructs a more discriminative model by considering all labels together based on the multi-label contrastive learning framework. Additionally, TPpred-SC incorporates the sequential semantic information based on the BERT model without using any expert knowledge. Consequently, TPpred-SC serves as a valuable tool for accurately predicting multiple therapeutic peptides.

3.4 Performance comparison between TPpred-SC and other MTPs predictors

To comprehensively evaluate the performance of TPpred-SC, we compared it with other MTPs predictors, including MLBP [10], PrMFTP [9], and TPpred-LE [57]. All of them are trained on the training dataset and evaluated on the independent test dataset. The results are depicted in Figure 3(a) indicate that TPpred-SC outperforms the others in terms of sample-level accuracy and label-level $F1$ score.

For analysis of performance variations across functions with varying degrees of data imbalance, [25] categorized the 15 functions into three function groups based on the number of samples: many-shot group (AMP, TXP, ABP, AIP, AVP), medium-shot group (ACP, AFP, DDV, CPP, CCC) and few-shot group (APP, AAP, AHTP, PBP, QSP). The medium-shot and few-shot groups are imbalanced because the positive samples are far fewer than the negative samples in a one-vs-rest manner.

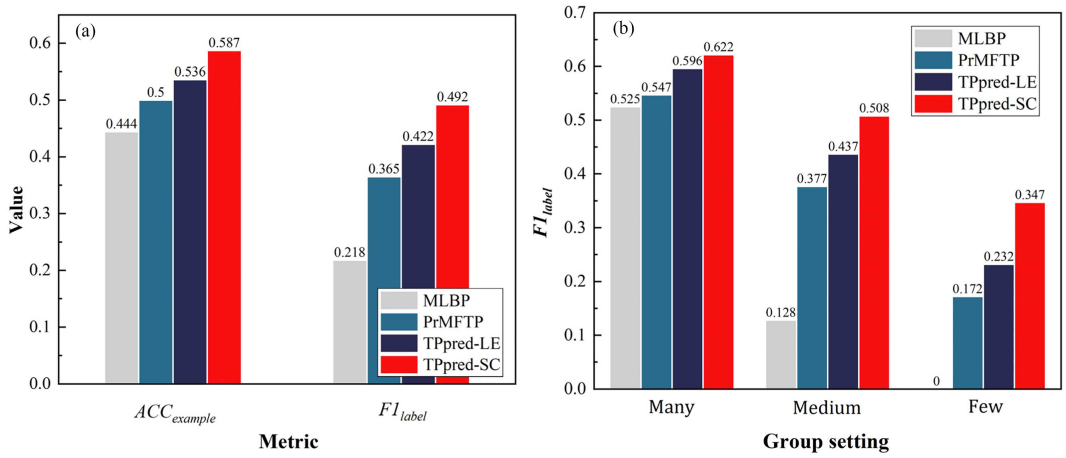

Figure 3 (Color online) The performance of TPpred-SC and other MTPs prediction methods on the independent dataset. (a) The overall performance of the three methods on the independent dataset. (b) The $F1_{\text{label}}$ scores of the three methods for each function group.

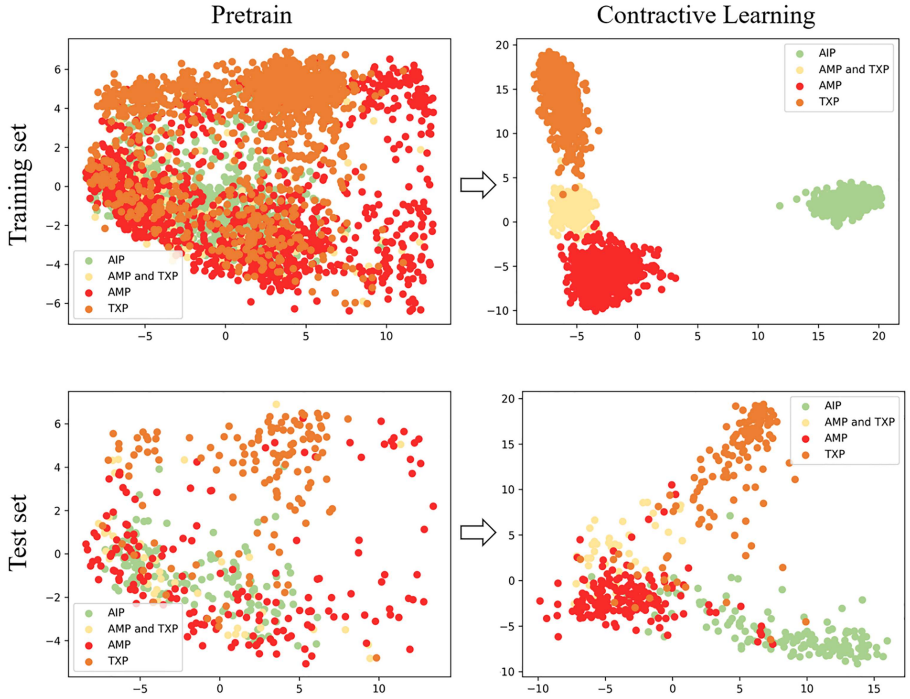

Figure 4 (Color online) The distribution of the features obtained by the BERT module. The first row represents the feature distribution of the training samples. The second row represents the feature distribution of the independent test samples. The first and the second columns represent the feature distribution of the samples extracted by the pretrained BERT without and with the multi-label contractive learning representation, respectively.

The performance of each function group is shown in Figure 3(b), from which we can conclude that: (i) Fuction prediction becomes increasingly challenging as the number of training samples decreases. (ii) TPpred-SC consistently outperforms previous methods in all function group settings, showcasing the benefits derived from the multi-label contrastive learning model. (iii) The performance of TPpred-SC on the few-shot group is clearly superior to that of other compared methods, benefiting from the pretrained BERT model. BERT has captured extensive sequential information from a vast number of unlabelled sequences so as to better represent the samples with few labels in this work [40]. Additionally, the utilization of data augmentation approaches contributes to outstanding performance by expanding the training set, especially for the peptides in the few-shot group [56, 60]. Therefore, TPpred-SC is able to identify MTPs more precisely.

3.5 Visualization of the power of supervised multi-label contrastive learning

To intuitively assess the effectiveness of the multi-label supervised contrastive learning, we utilized principal component analysis (PCA) [78] to interpret the feature representation ability obtained by the pretrained BERT model. We selected four functional therapeutic peptides, including AMP peptides, AIP peptides, TXP peptides, and the peptides with two functional categories AMP and TXP. The distribution of the specific peptides is illustrated in Figure 4, showing that: (i) the original pretrained BERT model has already clustered parts of the samples according to their labels, but the boundaries between different functions are almost mixed; (ii) the application of supervised contrastive learning enables better separation of function clusters from other samples. Notably, the "AMP" samples are closer to the "AMP and TXP" samples but far from "TXP" samples and "AIP" samples, which do not share any label overlap with the "AMP" samples. This observation indicates that supervised contrastive learning effectively discriminates by clustering samples from the same classes and bringing similar classes closer together, while simultaneously increasing the distance between different classes. Therefore, the multi-label contrastive learning enhances the representation, which drives performance improvement of the TPpred-SC model.

4 Conclusion

In this paper, we proposed a novel approach namely TPpred-SC to predict the functions of therapeutic peptides. We utilized the pretrained language model BERT as the backbone network to obtain the representation of contextual features. We leveraged the multi-label contrastive learning to enhance the representation of multi-functional therapeutic peptides to further utilize the classification module to make the final prediction. Experimental results show that TPpred-SC outperforms the existing therapeutic peptide function prediction methods in almost all metrics. Since the proposed multi-label contrastive learning model is a general framework for learning the representation of multi-functional therapeutic peptides, it will be applied to solve the peptide-protein interaction prediction, etc.

Acknowledgements We are greatly appreciative of the invaluable feedback provided by the anonymous reviewers, which has significantly enhanced the quality of this paper. This work was supported by National Natural Science Foundation of China (Grant Nos. 62325202, 62102030, U22A2039) and Beijing Natural Science Foundation (Grant No. L232067).

References

1 Fosgerau K, Hoffmann T. Peptide therapeutics: current status and future directions. Drug Discov Today, 2015, 20: 122–128
2 Lau J L, Dunn M K. Therapeutic peptides: historical perspectives, current development trends, and future directions. Bioorg Med Chem, 2018, 26: 2700–2707
3 Cai L, Wang L, Fu X, et al. Active semisupervised model for improving the identification of anticancer peptides. ACS Omega, 2021, 6: 23998–24008
4 Singh S, Chaudhary K, Dhanda S K, et al. SATPdb: a database of structurally annotated therapeutic peptides. Nucleic Acids Res, 2016, 44: D1119–D1126
5 Ao C, Jiao S, Wang Y, et al. Biological sequence classification: a review on data and general methods. Research, 2022, 2022: 0011
6 Cao C, Wang J, Kwok D, et al. webTWAS: a resource for disease candidate susceptibility genes identified by transcriptome-wide association study. Nucleic Acids Res, 2022, 50: D1123–D1130
7 Wei L, He W, Malik A, et al. Computational prediction and interpretation of cell-specific replication origin sites from multiple eukaryotes by exploiting stacking framework. Brief BioInf, 2021, 22: bbaa275
8 Wei L, Zhou C, Su R, et al. PEPred-Suite: improved and robust prediction of therapeutic peptides using adaptive feature representation learning. Bioinformatics, 2019, 35: 4272–4280
9 Yan W, Tang W, Wang L, et al. PrMFTP: multi-functional therapeutic peptides prediction based on multi-head self-attention mechanism and class weight optimization. PLoS Comput Biol, 2022, 18: e1010511
10 Tang W, Dai R, Yan W, et al. Identifying multi-functional bioactive peptide functions using multi-label deep learning. Brief BioInf, 2022, 23: bbab414
11 Wang Y, Zhai Y, Ding Y, et al. SBSM-pro: support bio-sequence machine for proteins. 2023. ArXiv:2308.10275
12 Zeng X, Wang F, Luo Y, et al. Deep generative molecular design reshapes drug discovery. Cell Reports Medicine, 2022, 3:1–13
13 Yan K, Lv H, Wen J, et al. TP-MV: therapeutic peptides prediction by multi-view learning. CBIO, 2022, 17: 174–183
14 Shen H B, Chou K C. PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition. Anal Biochem, 2008, 373: 386–388
15 Qi Y. Random forest for bioinformatics. Ensemble Machine Learning: Methods and Applications. Springer. 2012: 307–323
16 Aalen O O. A linear regression model for the analysis of life times. Stat Med, 1989, 8: 907–925
17 Hearst M A, Dumais S T, Osuna E, et al. Support vector machines. IEEE Intell Syst Their Appl, 1998, 13: 18–28
18 Ao C, Ye X, Sakurai T, et al. m5U-SVM: identification of RNA 5-methyluridine modification sites based on multi-view features of physicochemical features and distributed representation. BMC Biol, 2023, 21: 93
19 Wang Y, Zhai Y, Ding Y, et al. SBSM-pro: support bio-sequence machine for proteins. 2023. ArXiv:2308.10275
20 Li H L, Pang Y H, Liu B. BioSeq-BLM: a platform for analyzing DNA, RNA and protein sequences based on biological language models. Nucleic Acids Res, 2021, 49: e129

21 Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. J Bioinform Comput Biol, 2005, 03: 185–205
22 Zhang Y P, Zou Q, Luigi Martelli P. PPTPP: a novel therapeutic peptide prediction method using physicochemical property encoding and adaptive feature representation learning. Bioinformatics, 2020, 36: 3982–3987
23 Yan K, Lv H, Guo Y, et al. TPpred-ATMV: therapeutic peptide prediction by adaptive multi-view tensor learning model. Bioinformatics, 2022, 38: 2712–2718
24 Chen L, Yu L, Gao L, et al. Potent antibiotic design via guided search from antibacterial activity evaluations. Bioinformatics, 2023, 39: btad059
25 Yang H, Luo Y M, Ma C Y, et al. A gender specific risk assessment of coronary heart disease based on physical examination data. npj Digit Med, 2023, 6: 136
26 Zeng X, Xiang H, Yu L, et al. Accurate prediction of molecular properties and drug targets using a self-supervised image representation learning framework. Nat Mach Intell, 2022, 4: 1004–1016
27 Wei L, Ye X, Xue Y, et al. ATSE: a peptide toxicity predictor by exploiting structural and evolutionary information based on graph neural network and attention mechanism. Brief BioInf, 2021, 22: bbab041
28 Veltri D, Kamath U, Shehu A, et al. Deep learning improves antimicrobial peptide recognition. Bioinformatics, 2018, 34: 2740–2747
29 Wei L, Zhou C, Chen H, et al. ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. Bioinformatics, 2018, 34: 4007–4016
30 Yan K, Lv H, Guo Y, et al. sAMPpred-GAT: prediction of antimicrobial peptide by graph attention network and predicted peptide structure. Bioinformatics, 2023, 39: btac715
31 Yan K, Guo Y, Liu B, et al. PreTP-2L: identification of therapeutic peptides and their types using two-layer ensemble learning framework. Bioinformatics, 2023, 39: btad125
32 Yan J, Zhang B, Zhou M, et al. Multi-Branch-CNN: classification of ion channel interacting peptides using multi-branch convolutional neural network. Comput Biol Med, 2022, 147: 105717
33 Zhang J, Zhang Z, Pu L, et al. AIEpred: an ensemble predictive model of classifier chain to identify anti-inflammatory peptides. IEEE ACM Trans Comput Biol Bioinf, 2020, 18: 1831–1840
34 O'Shea K. An introduction to convolutional neural networks. 2015. ArXiv:1511.08458
35 Chung J, Gulcehre C, Cho K H, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling. 2014. ArXiv:1412.3555
36 Vaswani A. Attention is all you need. In: Proceedings of Conference on Neural Information Processing Systems, Long Beach, 2017. 1–11
37 Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space. 2013. ArXiv:1301.3781
38 Lv H, Yan K, Liu B. TPpred-LE: therapeutic peptide function prediction based on label embedding. BMC Biol, 2023, 21: 238
39 Altschul S. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res, 1997, 25: 3389–3402
40 Rao R, Bhattacharya N, Thomas N, et al. Evaluating protein transfer learning with TAPE. In: Proceedings of Conference on Neural Information Processing Systems, Vancouver, 2019. 32: 1–13
41 Elnaggar A, Heinzinger M, Dallago C, et al. ProtTrans: toward understanding the language of life through self-supervised learning. IEEE Trans Pattern Anal Mach Intell, 2021, 44: 7112–7127
42 Heinzinger M, Elnaggar A, Wang Y, et al. Modeling aspects of the language of life through transfer-learning protein sequences. BMC BioInf, 2019, 20: 1–7
43 Li Z, Jin J, Long W, et al. PLPMpro: enhancing promoter sequence prediction with prompt-learning based pre-trained language model. Comput Biol Med, 2023, 164: 107260
44 Jin J, Yu Y, Wang R, et al. iDNA-ABF: multi-scale deep biological language learning model for the interpretable prediction of DNA methylations. Genome Biol, 2022, 23: 219
45 Teufel F, Almagro Armenteros J J, Johansen A R, et al. SignalP 6.0 predicts all five types of signal peptides using protein language models. Nat Biotechnol, 2022, 40: 1023–1025
46 Salem M, Keshavarzi Arshadi A, Yuan J S. AMPDeep: hemolytic activity prediction of antimicrobial peptides using transfer learning. BMC BioInf, 2022, 23: 389
47 Sharma R, Shrivastava S, Kumar Singh S, et al. Deep-AFPpred: identifying novel antifungal peptides using pretrained embeddings from seq2vec with 1DCNN-BiLSTM. Brief BioInf, 2022, 23: bbab422
48 Dee W, Gromiha M. LMPred: predicting antimicrobial peptides using pre-trained language models and deep learning. BioInf Adv, 2022, 2: vbac021
49 Cheng J, Bendjama K, Rittner K, et al. BERTMHC: improved MHC-peptide class II interaction prediction with transformer and multiple instance learning. Bioinformatics, 2021, 37: 4172–4179
50 Charoenkwan P, Nantasenamat C, Hasan M M, et al. BERT4Bitter: a bidirectional encoder representations from transformers (BERT)-based model for improving the prediction of bitter peptides. Bioinformatics, 2021, 37: 2556–2562
51 Romero M, Nakano F K, Finke J, et al. Leveraging class hierarchy for detecting missing annotations on hierarchical multi-label classification. Comput Biol Med, 2023, 152: 106423
52 Khosravian M, Kazemi Faramarzi F, Mohammad Beigi M, et al. Predicting Antibacterial peptides by the concept of Chou's pseudo-amino acid composition and machine learning methods. Protein Peptide Lett, 2013, 20: 180–186
53 Burdukiewicz M, Sidorczuk K, Rafacz D, et al. Proteomic screening for prediction and design of antimicrobial peptides with ampgram. Int J Mol Sci, 2020, 21: 4310
54 Kavousi K, Bagheri M, Behrouzi S, et al. IAMPE: NMR-assisted computational prediction of antimicrobial peptides. J Chem Inf Model, 2020, 60: 4691–4701
55 Le-Khac P H, Healy G, Smeaton A F. Contrastive representation learning: a framework and review. IEEE Access, 2020, 8: 193907
56 Khosla P, Teterwak P, Wang C, et al. Supervised contrastive learning. In: Proceedings of Conference on Neural Information Processing Systems, Vancouver, 2020. 33: 18661–18673
57 Jaiswal A, Babu A R, Zadeh M Z, et al. A survey on contrastive self-supervised learning. Technologies, 2020, 9: 2
58 Tian Y, Sun C, Poole B, et al. What makes for good views for contrastive learning? In: Proceedings of Conference on Neural Information Processing Systems, Vancouver, 2020. 33: 6827–6839
59 Chen T, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations. In: Proceedings of International conference on machine learning, PMLR, 2020. 1597–1607

60 Shen H, Price L C, Bahadori T, et al. Improving generalizability of protein sequence models with data augmentations. 2021. BioRxiv: 2021.02
61 French S, Robson B. What is a conservative substitution? J Mol Evol, 1983, 19: 171–175
62 Devlin J. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018. ArXiv:1810.04805
63 Polanco C, Uversky V N, Huberman A, et al. Bioinformatics study of the DNA and RNA viruses infecting plants and bacteria that could potentially affect animals and humans. CBIO, 2023, 18: 170–191
64 Gardner M W, Dorling S R. Artificial neural networks (the multilayer perceptron)-a review of applications in the atmospheric sciences. Atmos Environ, 1998, 32: 2627–2636
65 Wang F, Liu H. Understanding the behaviour of contrastive loss. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Virtual, 2021. 2495–2504
66 Zaigrajew V, Zieba M. Contrastive learning for multi-label classification. In: Proceedings of Conference on Neural Information Processing Systems, New Orleans, 2022. 1–8
67 Murphy A H. The finley affair: a signal event in the history of forecast verification. Wea Forecasting, 1996, 11: 3–20
68 Jadon S. A survey of loss functions for semantic segmentation. In: Proceedings of IEEE conference on computational intelligence in bioinformatics and computational biology, Vina del Mar, 2020. 1–7
69 Loshchilov I. Decoupled weight decay regularization. 2017. ArXiv:1711.05101
70 Bradley A P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition, 1997, 30: 1145–1159
71 Wu J, Qu L, Yang G, et al. Diabetes induced factors prediction based on various improved machine learning methods. CBIO, 2022, 17: 254–262
72 Powers D M W. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. 2020. ArXiv:2010.16061
73 Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics, 2020, 21: 1–3
74 Zou X, Ren L, Cai P, et al. Accurately identifying hemagglutinin using sequence information and machine learning methods. Front Med, 2023, 10: 1281880
75 Zhu W, Yuan S S, Li J, et al. A first computational frame for recognizing heparin-binding protein. Diagnostics, 2023, 13: 2465
76 Li H, Liu B, Libbrecht M W. BioSeq-Diabolo: Biological sequence similarity analysis using Diabolo. PLoS Comput Biol, 2023, 19: e1011214
77 Shorten C, Khoshgoftaar T M. A survey on image data augmentation for deep learning. J Big Data, 2019, 6: 1–48
78 Abdi H, Williams L J. Principal component analysis. WIREs Comput Stats, 2010, 2: 433–459