Genomics, Proteomics & Bioinformatics, 2024, **22(6)**, qzae084
https://doi.org/10.1093/gpbjnl/qzae084
Advance access publication: 25 November 2024
Original Research

iMFP-LG: Identify Novel Multi-functional Peptides Using Protein Language Models and Graph-based Deep Learning

Jiawei Luo [1,#], Kejuan Zhao [2,#], Junjie Chen [1,#,*], Caihua Yang [1], Fuchuan Qu [1], Yumeng Liu [3], Xiaopeng Jin [3], Ke Yan [4], Yang Zhang [2], Bin Liu [4,5]

[1]School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen 518055, China
[2]School of Science, Harbin Institute of Technology, Shenzhen 518055, China
[3]College of Big Data and Internet, Shenzhen Technology University, Shenzhen 518055, China
[4]School of Computer Science and Technology, Beijing Institute of Technology, Beijing 10081, China
[5]Advanced Research Institute of Multidisciplinary Science, Beijing Institute of Technology, Beijing 10081, China

*Corresponding author: junjiechen@hit.edu.cn (Chen J).

#Equal contribution.

Handling Editor: Jing Tang

Abstract

Functional peptides are short amino acid fragments that have a wide range of beneficial functions for living organisms. The majority of previous studies have focused on mono-functional peptides, but an increasing number of multi-functional peptides have been discovered. Although there have been enormous experimental efforts to assay multi-functional peptides, only a small portion of millions of known peptides has been explored. The development of effective and accurate techniques for identifying multi-functional peptides can facilitate their discovery and mechanistic understanding. In this study, we presented iMFP-LG, a method for multi-functional peptide identification based on protein language models (pLMs) and graph attention networks (GATs). Our comparative analyses demonstrated that iMFP-LG outperformed the state-of-the-art methods in identifying both multi-functional bioactive peptides and multi-functional therapeutic peptides. The interpretability of iMFP-LG was also illustrated by visualizing attention patterns in pLMs and GATs. Regarding the outstanding performance of iMFP-LG on the identification of multi-functional peptides, we employed iMFP-LG to screen novel peptides with both anti-microbial and anti-cancer functions from millions of known peptides in the UniRef90 database. As a result, eight candidate peptides were identified, among which one candidate was validated to process both anti-bacterial and anti-cancer properties through molecular structure alignment and biological experiments. We anticipate that iMFP-LG can assist in the discovery of multi-functional peptides and contribute to the advancement of peptide drug design.

Key words: Multi-functional peptide discovery; Protein language model; Graph attention network; Therapeutic peptide screening; Deep learning.

Introduction

Functional peptides are short amino acid fragments, usually ≤ 50 amino acids, which play an important role in the regulation of a variety of biological functions, such as regulating hormones, neurotransmitters, and growth factors [1,2]. Because of their excellent selectivity, effectiveness, comparative safety, and good tolerability in biological activities, functional peptides have attracted tremendous attention in medicine [3–5]. Up to now, peptides with a wide range of functions have been discovered [6,7], including anti-microbial peptides (AMPs) and anti-cancer peptides (ACPs). Notably, an increasing number of peptides have been demonstrated to have multiple functions. For instance, some AMPs show lethal effects on cancer cells [8]. Effective and accurate techniques for identifying multi-functional peptides can facilitate their discovery and mechanistic understanding. Unfortunately, biological experiments for studying peptide functions are time-consuming and expensive in both labor and materials.

Currently, computational approaches have achieved remarkable success in the discovery of peptide functions [5,9–12]. With the development of machine learning techniques, the methodologies of peptide discovery have gone through three stages: conventional machine learning-based methods, deep learning-based methods, and pre-trained protein language model (pLM)-based methods. Conventional machine learning-based methods (*e.g.*, AVPpred [13], PredAPP [14], AIPpred [15], THPep [16], and AntiCP 2.0 [17]) identify peptides by employing Support Vector Machine (SVM) and Random Forest (RF) algorithms based on various feature engineering techniques, including Position-Specific Scoring Matrix (PSSM) [18], physicochemical properties [19], and pseudo amino acid composition (PseAAC) [20]. Despite their effectiveness, conventional machine learning-based methods are often hard to generalize to other peptide datasets. In contrast to the hand-crafted features in conventional machine learning-based methods, deep learning-based methods (*e.g.*, DeepACP [21], Su et al. [22], Veltri et al. [23], and Ma et al. [24]) automatically extract features by employing various deep learning architectures, including Convolutional Neural Network (CNN), Long Short Term Memory (LSTM), and their combination, for distinguishing functional peptides. Some researchers have combined various feature engineering techniques with deep learning methods to construct powerful predictors (*e.g.*, iACP-DRLF [25], ACP-MHCNN [26], ITP-Pred [27], MLACP 2.0 [28], and ACP-2DCNN [29]). However, due to the small datasets of functional peptides, these supervised deep learning-based methods encounter the challenge of learning robust peptide representation. Recently, pre-trained language models have emerged as a novel

Received: 10 July 2023; **Revised:** 25 October 2024; **Accepted:** 21 November 2024.

powerful paradigm in the field of natural language processing (NLP) [30,31]. They are typically initially trained on extensive datasets in a self-supervised manner and are subsequently leveraged for a myriad of downstream tasks. pLMs have also been proposed and applied to peptide identification [32], such as anti-bacterial peptides (ABPs) [33], AMPs [24], signal peptides [34], anti-hypertensive peptides (AHPs) [35], and bitter peptides [36]. However, these studies focus on mono-functional peptide prediction.

In contrast to the discovery of mono-functional peptides, multi-functional peptide identification is a multi-label classification task, which assigns a set of relevant functional labels to a peptide simultaneously. Multi-label classification is more complex due to the challenges in capturing hidden connections of labels and resolving data imbalances. Several studies have made a difficult endeavor in the discovery of multi-functional peptides. Tang et al. [37] and Li et al. [38] identified multi-functional peptides by using a multi-label deep learning method, which combines CNN and recurrent neural network (RNN) to extract peptide features and assigns function labels separately. PrMFTP [39] improved the performance of multi-functional therapeutic peptide identification by employing a multi-scale CNN and an attention-based bidirectional long short-term memory (BiLSTM). Since the number of multi-functional peptides is fewer than that of mono-functional peptides, the training datasets are extremely imbalanced. ETFC [40] utilized a text CNN combined with a multi-label focal dice loss function to solve the inherent imbalance problem in the multi-functional peptide prediction. Although some training tricks have effectively mitigated the effect of imbalanced data, existing methods still lack generalization to learn comprehensive peptide representations and have high false positive rates. In addition, they predicted all function labels independently without considering their correlations. Pre-trained pLMs and graph attention networks (GATs) offer solutions to these problems. The pLMs pre-trained on millions of protein sequences [41] can capture long-range dependencies of amino acid residues [42–44]. GATs have the ability to capture complex relationships by computing attention coefficients between different objects [45]. The advantages of both pLMs and GATs can provide helpful insights for the discovery of multi-functional peptides.

In this study, we developed a method, iMFP-LG, for discovering multi-functional peptides based on pLMs and GATs. To the best of our knowledge, iMFP-LG is the first approach that considers the associations between function labels to identify multi-functional peptides by converting the multi-label problem to the graph node classification. The computational results showed that iMFP-LG outperformed the state-of-the-art methods on both multi-functional bioactive peptide (MFBP) and multi-functional therapeutic peptide (MFTP) datasets. iMFP-LG is also interpretable by visualizing the distribution of peptide representations, motif patterns obtained by the pLM, and function relationships captured by the GAT. We employed the iMFP-LG model to establish a robust peptide discovery pipeline. Through this pipeline, eight novel candidate multi-functional peptides were screened out with high confidence from the UniRef90 database, which is an extensive collection of millions of known peptides. Further biological experiments showed that one of the candidates had remarkable bioactivities in terms of anti-microbial and anti-cancer functions. These results demonstrate the capability of iMFP-LG for the discovery of novel multi-functional peptides.

Results and discussion

Overview of the proposed method iMFP-LG

The architecture of iMFP-LG consists of two modules: a peptide representation module and a graph classification module (**Figure 1**). Within the peptide representation module, the pLM is responsible for acquiring high-quality peptide representations through a multi-head self-attention mechanism. In the graph classification module, the GAT is employed to capture the interrelationships among various function labels. The nodes in the graph are function labels and the edges are the correlations between function labels. The node features are first generated by the pLM according to the input peptide sequences, while edge weights are learned by the GAT. All node features are further updated in the GAT to integrate complex relationships between peptide function labels according to the corresponding edge weights. At last, the multi-functions of peptides are determined by a set of node binary classifiers based on the updated node features. In addition, adversarial training is used to improve model robustness and generalization ability. To discover novel multi-functional peptides, we established a robust pipeline based on the trained iMFP-LG model. The architecture of iMFP-LG and the adversarial training are introduced as follows.

Graph node classification framework improves discovery of multi-functional peptides

The discovery of multi-functional peptides is a multi-label classification task. To capture the complex relationships among multi-functions, we proposed a graph node classification framework. In this section, we evaluated the effectiveness of graph node classification framework on several widely used sequence feature extraction methods, including four sequence composition features (CFs) extracted by iFeatureOmega [46], *i.e.*, amino acid composition (AAC), pseudo-amino acid composition (PAAC) [20], Distance-Pairs (DP) [47], and Composition-Transition-Distribution (CTDD) [48], three deep learning-based methods (*i.e.*, CNN, RNN, and CNN+BiLSTM), and a pLM-based method (*i.e.*, TAPE [49]).

We first evaluated the performance of abovementioned feature extraction methods with or without GAT on the MFBP (**Figure 2A**) and MFTP (Figure 2B) datasets in terms of precision, coverage, accuracy, and absolute true. We observed that all methods without GAT were surrounded by the corresponding methods with GAT in radar maps, suggesting that the methods with GAT outperform the methods without GAT in terms of precision, coverage, accuracy, and absolute true. These results demonstrate that the proposed graph node classification framework can enhance the identification of multi-functional peptides.

We also compared the performance of all competing methods on the MFBP (Figure 2C) and MFTP (Figure 2D) datasets. The pLM without GAT outperformed other features with or without GAT. And the performance of pLM was further improved by GAT, achieving the best performance with a precision of 0.777, coverage of 0.785, accuracy of 0.776, absolute true of 0.767, and absolute false of 0.082 on the MFBP dataset and a precision of 0.721, coverage of 0.722, accuracy of 0.679, absolute true of 0.605, and absolute false

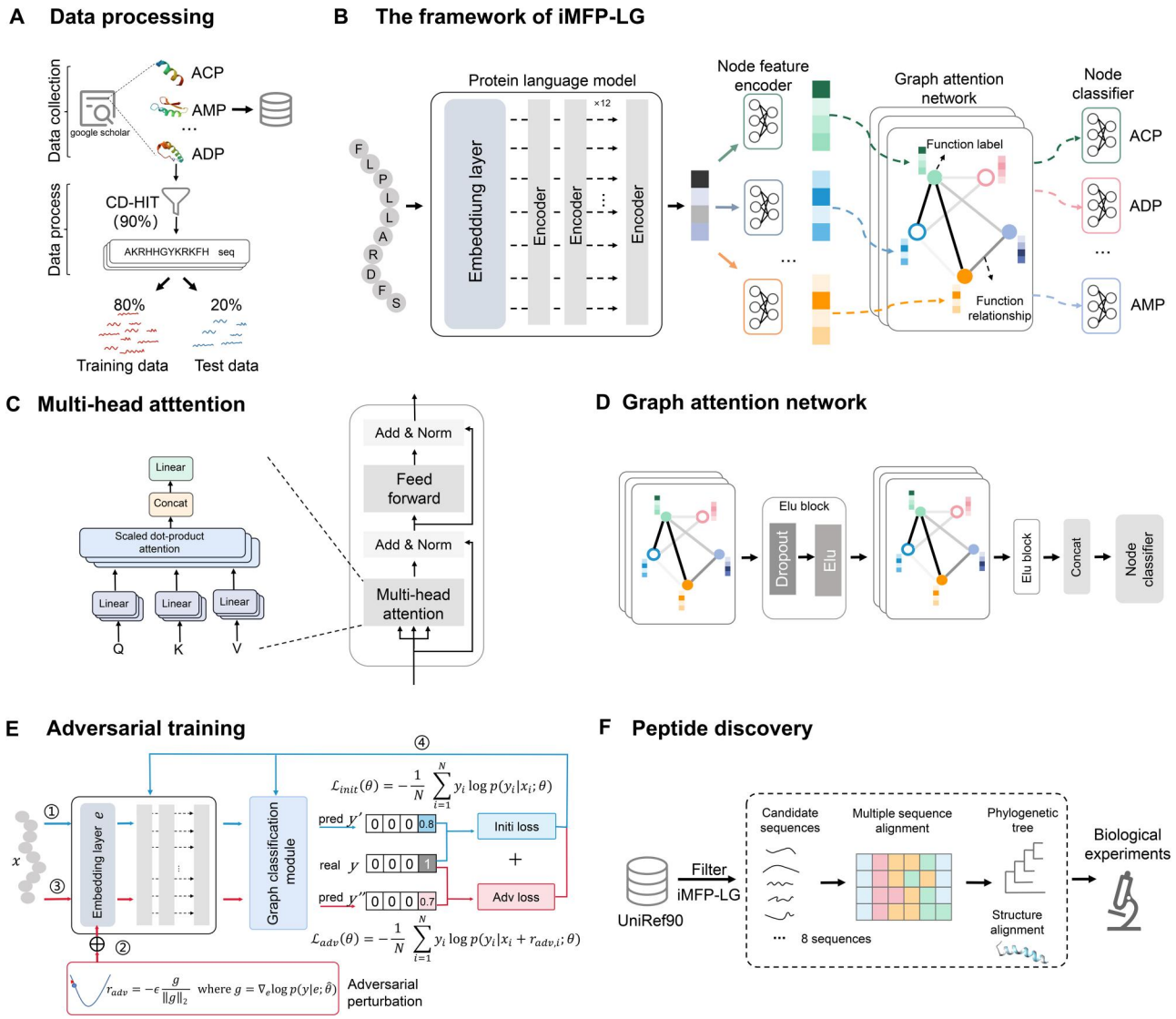


Figure 1 Illustration of the proposed iMFP-LG
A. Collection and processing of multi-functional peptide datasets. **B**. The framework of iMFP-LG, including a peptide representation module and a graph classification module. The peptide representation module employs a protein language model to extract high-quality peptide representations from primary sequences. The graph classification module is composed of a node feature encoder, a graph attention network, and a node classifier to learn the correlation of function labels. **C**. The multi-head attention mechanism of an encoder layer in the protein language model. **D**. Two stacked graph layers in the graph attention network. **E**. The procedure of adversarial training. **F**. The pipeline of peptide discovery by iMFP-LG.

of 0.032 on the MFTP dataset. All results can be found in Tables S1 and S2. These results indicate that pLM can extract more comprehensive and high-quality features from peptide sequences than traditional features and other deep learning-based methods.

Thus, we constructed iMFP-LG by incorporating pLM as a feature extraction module and GAT as an identification module. To improve the generalization capability and avoid the over-fitting phenomenon, we also employed adversarial training to achieve better results in the final framework.

iMFP-LG outperforms the state-of-the-art methods

We compared our proposed method, iMFP-LG, with several state-of-the-art methods, including four conventional machine learning-based methods (CLR [50], RAKEL [51], RBRL [52], and MLDF [53]) and four deep learning-based methods (MPMABP [38], MLBP [37], PrMFTP [39], and ETFC [40]). MPMABP and MLBP employed CNNs and RNNs for identifying multi-functional peptides. In addition to CNNs and RNNs, PrMFTP employed a multi-head self-attention module to further optimize the extracted features for predicting multi-functional therapeutic peptides. ETFC mitigated data imbalance using the focal dice loss function. The focal dice loss mitigates class imbalance by down-weighting the contribution of well-classified examples, allowing the model to focus more on hard-to-classify instances, which are often underrepresented. Note that due to the randomness of dividing the entire datasets into training and test datasets, we could not access their original training datasets. Besides, the compared methods didn't provide the sensitivity and specificity analyses to each peptide category. Therefore, we reproduced the four deep learning-based methods in the same training dataset with our proposed method. All the reproduced results were comparable in performance to their

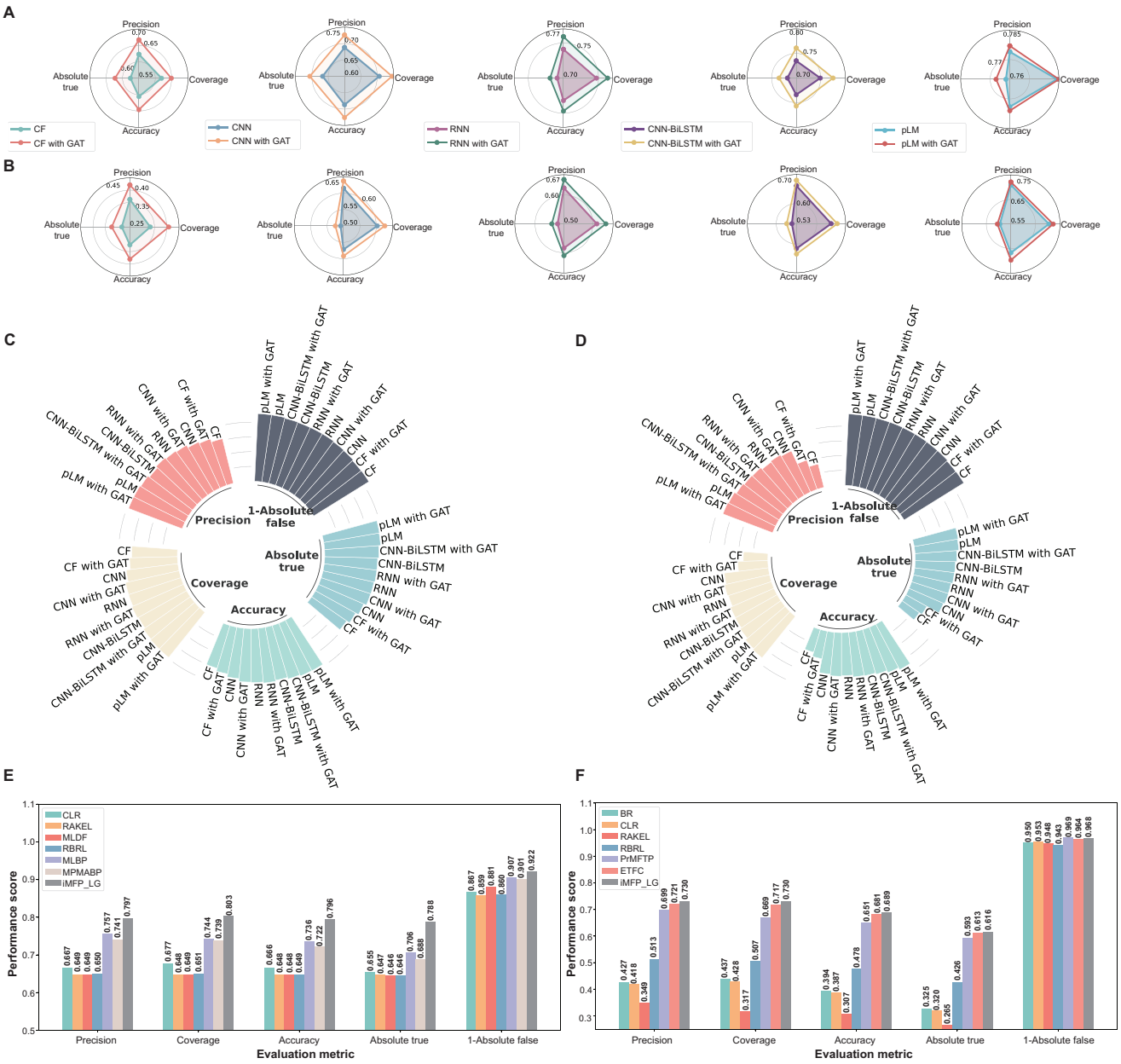


Figure 2 Performance comparison of different methods for multi-functional peptide identification

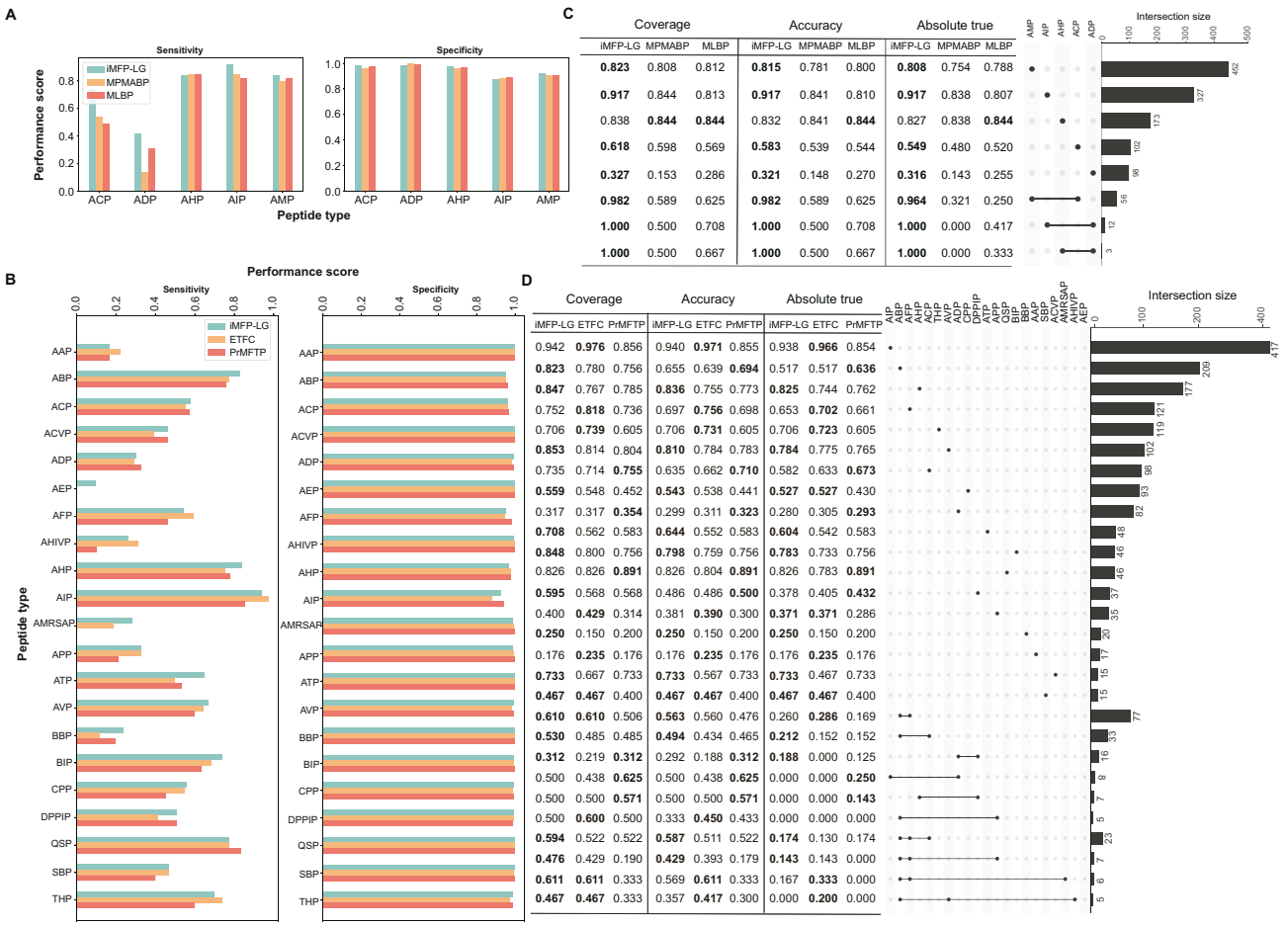
A. and **B**. Effect of GAT on different feature extraction methods on the MFBP (A) and MFTP (B) datasets. **C**. and **D**. Performance comparison of different feature extraction methods on the MFBP (C) and MFTP (D) datasets. **E**. and **F**. Performance comparison of our proposed method iMFP-LG and the state-of-the-art methods on the MFBP (E) and MFTP (F) datasets. GAT, graph attention network; pLM, protein language model; CF, composition feature; MFBP, multi-functional bioactive peptide; MFTP, multi-functional therapeutic peptide; CNN, convolutional neural network; RNN, recurrent neural network; BiLSTM, bidirectional long short-term memory.

reported results. For those methods that were not publicly available, we used only the results reported in the literature.

According to the performance comparison on the MFBP (Figure 2E) and MFTP (Figure 2F) datasets, iMFP-LG outperformed the state-of-the-art methods on both datasets in terms of all metrics, except for a comparable performance in terms of absolute false in the MFTP dataset. When compared on the MFBP dataset, iMFP-LG achieved a precision of 0.797, coverage of 0.803, accuracy of 0.796, absolute true of 0.788, and absolute false of 0.078, greatly outperforming the state-of-the-art method MLBP by 4.0%, 5.9%, 6.0%, 8.2%, and 1.5%, respectively. When compared on the MFTP dataset, iMFP-LG achieved a precision of 0.730, coverage of 0.730, accuracy of 0.689, and absolute true of 0.616, outperforming the state-of-the-art ETFC model by 0.9%, 1.3%, 0.8%, and 0.3%, respectively. All the methods achieved comparable performance in terms of absolute false. These results can be found in Tables S3 and S4. Overall, iMFP-LG comprehensively outperforms all multi-functional peptide prediction methods.

iMFP-LG is more sensitive to peptide categories with small size and multi-functions

As there are 5 types of bioactive peptide functions in the MFBP dataset and 21 types of therapeutic peptide functions



C	Coverage			Accuracy			Absolute true		
	iMFP-LG	MPMABP	MLBP	iMFP-LG	MPMABP	MLBP	iMFP-LG	MPMABP	MLBP
	0.823	0.808	0.812	**0.815**	0.781	0.800	**0.808**	0.754	0.788
	0.917	0.844	0.813	**0.917**	0.841	0.810	**0.917**	0.838	0.807
	0.838	**0.844**	**0.844**	0.832	0.841	**0.844**	0.827	0.838	**0.844**
	0.618	0.598	0.569	**0.583**	0.539	0.544	**0.549**	0.480	0.520
	0.327	0.153	0.286	**0.321**	0.148	0.270	**0.316**	0.143	0.255
	0.982	0.589	0.625	**0.982**	0.589	0.625	**0.964**	0.321	0.250
	1.000	0.500	0.708	**1.000**	0.500	0.708	**1.000**	0.000	0.417
	1.000	0.500	0.667	**1.000**	0.500	0.667	**1.000**	0.000	0.333

D	Coverage			Accuracy			Absolute true		
	iMFP-LG	ETFC	PrMFTP	iMFP-LG	ETFC	PrMFTP	iMFP-LG	ETFC	PrMFTP
	0.942	**0.976**	0.856	0.940	**0.971**	0.855	0.938	**0.966**	0.854
	0.823	0.780	0.756	0.655	0.639	**0.694**	0.517	0.517	**0.636**
	0.847	0.767	0.785	**0.836**	0.755	0.773	**0.825**	0.744	0.762
	0.752	**0.818**	0.736	0.697	**0.756**	0.698	0.653	**0.702**	0.661
	0.706	**0.739**	0.605	0.706	**0.731**	0.605	0.706	**0.723**	0.605
	0.853	0.814	0.804	**0.810**	0.784	0.783	**0.784**	0.775	0.765
	0.735	0.714	**0.755**	0.635	0.662	**0.710**	0.582	0.633	**0.673**
	0.559	0.548	0.452	**0.543**	0.538	0.441	**0.527**	**0.527**	0.430
	0.317	0.317	**0.354**	0.299	0.311	**0.323**	0.280	0.305	**0.293**
	0.708	0.562	0.583	**0.644**	0.552	0.583	**0.604**	0.542	0.583
	0.848	0.800	0.756	**0.798**	0.759	0.756	**0.783**	0.733	0.756
	0.826	0.826	**0.891**	0.826	0.804	**0.891**	0.826	0.783	**0.891**
	0.595	0.568	0.568	0.486	0.486	**0.500**	0.378	0.405	**0.432**
	0.400	**0.429**	0.314	0.381	**0.390**	0.300	**0.371**	**0.371**	0.286
	0.250	0.150	0.200	**0.250**	0.150	0.200	**0.250**	0.150	0.200
	0.176	**0.235**	0.176	0.176	**0.235**	0.176	0.176	**0.235**	0.176
	0.733	0.667	0.733	**0.733**	0.567	0.733	**0.733**	0.467	0.733
	0.467	**0.467**	0.400	**0.467**	**0.467**	0.400	**0.467**	**0.467**	0.400
	0.610	**0.610**	0.506	**0.563**	0.560	0.476	0.260	**0.286**	0.169
	0.530	0.485	0.485	**0.494**	0.434	0.465	**0.212**	0.152	0.152
	0.312	0.219	**0.312**	0.292	0.188	**0.312**	**0.188**	0.000	0.125
	0.500	0.438	**0.625**	0.500	0.438	**0.625**	0.000	0.000	**0.250**
	0.500	0.500	**0.571**	0.500	0.500	**0.571**	0.000	0.000	**0.143**
	0.500	**0.600**	0.500	0.333	**0.450**	0.433	0.000	0.000	0.000
	0.594	0.522	0.522	**0.587**	0.511	0.522	**0.174**	0.130	0.174
	0.476	0.429	0.190	**0.429**	0.393	0.179	**0.143**	0.143	0.000
	0.611	**0.611**	0.333	0.569	**0.611**	0.333	0.167	**0.333**	0.000
	0.467	**0.467**	0.333	0.357	**0.417**	0.300	0.000	**0.200**	0.000

Figure 3 Performance comparison of different methods in each peptide category
A. and **B**. Sensitivity and specificity of different competing methods on the MFBP (A) and MFTP (B) datasets. **C**. and **D**. Performance of different competing methods on mono-functional and multi-functional peptide prediction on the MFBP (C) and MFTP (D) datasets. The upset plot on the right shows the size of peptide categories in test datasets. For MFTP, we only show the peptide categories with size ≥ 5 and multi-functions ≤ 3. The table on the left shows the performance in terms of coverage, accuracy, and absolute true for corresponding peptide categories in the upset plot. The values in bold indicate the best performance across the compared methods. ACP, anti-cancer peptide; ADP, anti-diabetic peptide; AHP, anti-hypertensive peptide; AIP, anti-inflammatory peptide; AMP, anti-microbial peptide; AAP, anti-angiogenic peptide; ABP, anti-bacterial peptide; ACVP, anti-coronavirus peptide; AEP, anti-endotoxin peptide; AFP, anti-fungal peptide; AHIVP, anti-HIV peptide; AMRSAP, anti-MRSA peptide; APP, anti-parasitic peptide; ATP, anti-tubercular peptide; AVP, anti-viral peptide; BBP, blood-brain barrier peptide; CPP, cell-penetrating peptide; SBP, surface binding peptide; THP, tumor homing peptide.

in the MFTP dataset, we further investigated the performance of competing methods in each peptide category.

We compared the sensitivity and specificity of iMFP-LG with other state-of-the-art methods on all kinds of peptide functions. The sensitivity and specificity were calculated by considering peptides with a function as positive samples and other peptides without that function as negative samples. The comparison results on the MFBP (**Figure 3A**) and MFTP (Figure 3B) datasets showed that all competing methods had high specificities on both datasets, indicating that these models have high specificities in single function classification. However, their sensitivities are low and differ largely. On the MFBP dataset in terms of sensitivity, iMFP-LG achieved the best results in the prediction of ACP, anti-diabetic peptide (ADP), anti-inflammatory peptide (AIP), and AMP. Specifically, it had a much higher sensitivity than the other two competing methods in two small categories ACP and ADP. On the MFTP dataset in terms of sensitivity, iMFP-LG also achieved better results than PrMFTP and ETFC across almost all peptide functions. Interestingly, PrMFTP failed to distinguish the peptides in three small categories anti-endotoxin peptide (AEP), anti-HIV peptide (AHIVP), and anti-MRSA peptide (AMRSAP), and ETFC failed to predict the AEP. Nonetheless, iMFP-LG greatly improved the prediction performance. These results demonstrate that iMFP-LG is more sensitive to peptide categories of small size by taking advantage of pLM to learn high-quality peptide representations.

We then evaluated the performance of iMFP-LG in identifying both mono-functional and multi-functional peptides on the MFBP (Figure 3C) and MFTP (Figure 3D) datasets. On the MFBP dataset, iMFP-LG outperformed the competing methods across all peptide categories except for AHP. Notably, for multi-functional peptides, iMFP-LG achieved an absolute true of 1.00 in both ADP_AIP and ADP_AHP functions, and an absolute true of 0.964 in the ACP_AMP functions. On the MFTP dataset, iMFP-LG also achieved better performance than the competing methods in most peptide categories, especially in the multi-functional peptides. These results demonstrate that iMFP-LG is more sensitive to peptide categories with multi-functions by taking advantage of GAT to capture the complex functional relationships.

Table 1 The performance of iMFP-LG and its variants on the MFBP dataset

Model	Precision ↑	Coverage ↑	Accuracy ↑	Absolute true ↑	Absolute false ↓
iMFP-LG	**0.797**	**0.803**	**0.796**	**0.788**	**0.078**
w/o ad[a]	0.777	0.785	0.776	0.767	0.082
w/o GAT[b]	0.785	0.791	0.784	0.776	0.080
w/o pretrain[c]	0.754	0.769	0.752	0.733	0.095

Note: The highest values are highlighted in bold. ↑ means that a larger value is better on this metric; ↓ means that a smaller value is better on this metric; [a] w/o ad is a variant in which the adversarial training is not used during training process; [b] w/o GAT is a variant without GAT; [c] w/o pretrain is a variant in which the protein language model is re-initialized randomly instead of using pre-trained weights. GAT, graph attention network; MFBP, multi-functional bioactive peptide.

Table 2 The performance of iMFP-LG and its variants on the MFTP dataset

Model	Precision ↑	Coverage ↑	Accuracy ↑	Absolute true ↑	Absolute false ↓
iMFP-LG	**0.730**	**0.730**	**0.689**	**0.616**	**0.032**
w/o ad[a]	0.721	0.722	0.679	0.605	0.032
w/o GAT[b]	0.709	0.705	0.667	0.598	0.033
w/o pretrain[c]	0.658	0.657	0.618	0.547	0.036

Note: The highest values are highlighted in bold. ↑ means that a larger value is better on this metric; ↓ means that a smaller value is better on this metric; [a] w/o ad is a variant in which the adversarial training is not used during training process; [b] w/o GAT is a variant without GAT; [c] w/o pretrain is a variant in which the protein language model is re-initialized randomly instead of using pre-trained weights. MFTP, multi-functional therapeutic peptide.

Model ablation study

We then explored the contribution of each module to our proposed method using ablation experiments on two datasets. There are three important modules in iMFP-LG: adversarial training, GAT, and pre-trained pLM. We built several variants of iMFP-LG with and without these modules.

Tables 1 and 2 show the performance of iMFP-LG and its variants on the MFBP and MFTP datasets. All variants were consistent with the experimental settings except for the learning rate in the ablation study. Because pLMs with random initialization are difficult to train at small learning rates, we set the learning rate to 1E–5 in the "w/o pretrain" model and 5E–5 for all others. As can be seen, removing any module from the proposed model reduces its performance. On the MFBP dataset, the performance of the "w/o pretrain" model was the most deteriorated with an accuracy of 0.752 and absolute true of 0.733. The "w/o ad" model showed a reduction in performance after removing the adversarial training with an accuracy of 0.776 and absolute true of 0.767. Similarly, the performance of the "w/o GAT" model was dropped with an accuracy of 0.784 and absolute true of 0.776. On the MFTP dataset, the performance of the "w/o pretrain" model decreased drastically, and its accuracy and absolute true decreased by 7.1% and 6.9%, respectively. The performance of both "w/o GAT" and "w/o ad" models also decreased obviously in terms of all metrics. Ablation studies reveal that all three modules have critical contributions to improving the performance of multi-functional peptide prediction, especially the GAT and pLM modules.

Interpretability of iMFP-LG

The intuition behind iMFP-LG is to extract key sequence patterns using pLM and capture intricate multi-functional relationships via GAT. Thus, we can unveil its decision process of how to assign multi-functional labels to peptides by visualizing the distribution of peptide representations, sequence motifs, and multi-function relationships.

All peptide representations were extracted by the pLM from the MFBP and MFTP test datasets and visualized using *t*-distributed Stochastic Neighborhood Embedding (*t*-SNE) [54]. We compared the distribution of peptide representations extracted by the pre-trained and fine-tuned pLMs from the MFBP and MFTP datasets, respectively (**Figure 4A–D**). Although the pre-trained model initially identified different multi-functional peptides, the fine-tuned model exhibited more distinct ability to cluster peptides with the same functions into categories. Interestingly, the peptide clusters with multiple functions lay between corresponding clusters of mono-functional peptides. For example, in the distribution of multi-functional bioactive peptides (Figure 4B), the peptide cluster with ADP_AIP function lay between the peptide cluster with ADP function and the peptide cluster with AIP function. And the peptide cluster with ACP_AMP function was also close to the cluster with ACP function and the cluster with AMP function. A similar phenomenon was also observed in the distribution of multi-functional therapeutic peptides (Figure 4D), such as the distribution of the peptide clusters with ABP, ACP, and ABP_ACP functions as well as the peptide clusters with ADP, dipeptidyl peptidase IV peptide (DPPIP), and ADP_DPPIP functions. These results suggest that iMFP-LG has the capability to map peptides into a robust representation space, demonstrating spatial interpretability with the continuity property: two similar multi-functional peptides should not be projected as two distant points in the representation space. This advancement provides an opportunity to explore the phylogenetic relationships of multi-functional peptides in relation to mono-functional peptides.

To extract sequence patterns from the attention mechanism, we calculated the importance of each amino acid in a peptide sequence. Each sequence is transformed into an embedding by 12 attention heads in each of 12 layers, resulting in a total of 144 attention heads in the pLM. The weight $\beta_{i,j}$ in an attention head matrix indicates how much information from the amino acid in position j should be used when computing the representation of amino acid in position i. We summed the attention of all amino acids to amino acid j as its effect $d_j = \sum_{i=1}^{n} \beta_{i,j}$ to the entire peptide sequence. A higher d_j value suggests that amino acid j is more important in the peptide sequence. Figure 4E shows three AMP cases where the

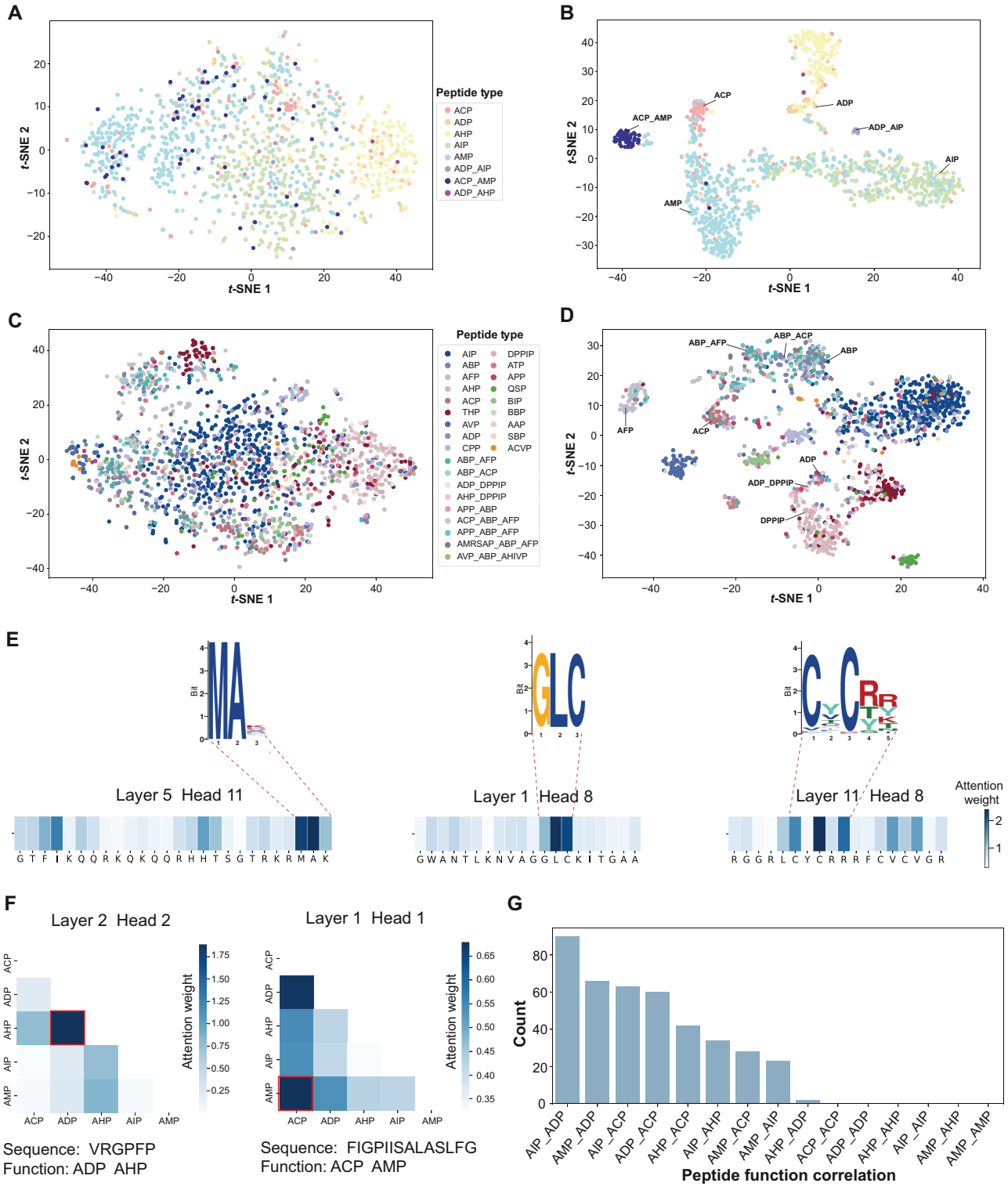


Figure 4 Illustration of the interpretability of iMFP-LG

t-SNE visualization of the distribution of peptide representations obtained by pre-trained and fine-tuned pLMs on the MFBP dataset (**A** and **B**) and the MFTP dataset (**C** and **D**), respectively. **E**. Three AMP cases where the sequence patterns captured by pLM are matched with the motifs discovered by STREME. **F**. Two multi-functional peptide cases in which the learned graph node relationships are consistent with their true function labels. **G**. Peptide function correlation between the maximum values within each GAT attention matrix. *t*-SNE, *t*-distributed Stochastic Neighborhood Embedding.

sequence patterns captured by attention mechanism were matched with the motifs discovered by STREME. The parameters of STREME are shown in Table S7. We also used the attention visualization tool bertviz [55] to reveal the detail of the attention pattern (Figure S1). These results suggest that iMFP-LG can identify functional regions of peptides.

To interpret the intricate correlations captured by GAT, we visualized the node connections across two layers and six heads. The node connection $r_{i,j} = \gamma_{i,j} + \gamma_{j,i}$ is calculated from the graph attention matrix γ, where each element $\gamma_{i,j}$ describes the importance of edge from node i to node j. Figure 4F shows two multi-functional peptide cases, where the node connections with the highest values matched with their true function labels. To find more reliable functional correlations in GAT, we also calculated the pairwise correlations of peptide functions by counting the occurrences of two function labels that have the maximum node connection values in all graph attention matrices. We assessed the captured node correlations of AIP_ADP training samples (Figure 4G). The functions AIP and ADP have the highest counts, indicating that they have the strongest correlation. These results indicate that iMFP-LG can learn the correlations among different peptide function labels.

Discovery of novel multi-functional peptides

Regarding the outstanding performance of iMFP-LG in identifying multi-functional peptides, especially on the MFBP dataset where it achieves an absolute true of 0.964 in ACP_AMP function, we employed iMFP-LG to screen novel candidate peptides with both ACP and AMP functions from millions of known peptides in the UniRef90 database.

Considering that UniRef90 contains 166,459,614 protein sequences, we filtered out any sequences longer than 40 amino acids or short than 4 amino acids to specifically focus on peptides. This resulted in a dataset of 1,077,593 peptide sequences, which were then fed into iMFP-LG. To discover novel multi-functional peptides, we re-trained iMFP-LG on the entire MFBP dataset with the same hyperparameters. To obtain candidate peptides with high confidence, the classification threshold was set to 0.95 for both ACP and AMP functions. After removing duplicated peptides that appeared in the MFBP dataset, we ultimately achieved 8 candidate peptides (**Figure 5A**). The functions of these peptides were further verified by searching their homologous sequences (Figure 5B, Figure S2A) through multiple sequence alignment [56], using candidate sequences as queries and peptides with ACP or AMP function in the MFBP dataset as targets. For the candidates that had ≥ 4 homologous sequences, we successfully constructed their phylogenetic trees, demonstrating that these candidates have an evolutionary relationship with known ACPs and AMPs. We then predicted their structures using ESM-Fold [57], and performed a structural alignment with their homologous sequences. The peptide UniRef90_P82904 exhibited exceptional alignment results (Figure 5C), while the peptides UniRef90_P83653 and UniRef90_B9W4V2 had no similar structures to their homologous sequences (Figure S2B and C).

To further assess the functions of these three candidate peptides, we subsequently conducted biological experiments. A positive control (UniRef90_P56917) with both AMP and ACP functions [58] was randomly selected. Peptides at a concentration of 500 μM were added to the cell culture for 24 h, followed by the 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide (MTT) cytotoxicity assay. Three independent replicates were conducted. The results showed that the peptide UniRef90_P82904 had an excellent anti-bacterial effect against both *Escherichia coli* and *Staphylococcus aureus* (Figure 5D and E). The other two peptides, UniRef90_P83653 and UniRef90_B9W4V2, exhibited no anti-bacterial activity against either *E. coli* or *S. aureus* (Figure S2D and E). To assess the anti-cancer activity of these peptides, cell viability tests were performed on three kinds of human tumor cell lines, including bladder cancer (T24), cervical cancer (HeLa), and liver cancer (HepG2). We found that the peptide UniRef90_P82904 had the strongest anti-cancer activity against all three tested cell lines (Figure 5F), and the peptide UniRef90_P83653 demonstrated strong anti-cancer activity against HeLa cells but had weak anti-cancer activity against T24 and HepG2 cells (Figure S2F). The peptide UniRef90_B9W4V2 exclusively exhibited weak anti-cancer activity against T24 cells, and had no effect on the other two tested cell lines (Figure S2F). Subsequently, the peptide UniRef90_P82904 was chosen for a dose-dependence analysis in a 24-h assay to evaluate its cytotoxic effect against three tumor cell lines using the standard MTT assay. As shown in Figure 5G–I, the cell viability decreased to as low as 4.7%, 7.7%, and 4.3% at 250 μg/ml against HeLa, T24, and HepG2 cancer cells, demonstrating the promising capability for ablation of these three cancer cells. The outcomes of biological experiments are consistent with the computational screening results of iMFP-LG, indicating that iMFP-LG has a strong potential to discover novel multi-functional peptides.

Conclusion

In this study, we proposed a method iMFP-LG for discovering multi-functional peptides. iMFP-LG converts multi-label predictions to graph node classifications based on pLM and GAT. Comparison results on the MFBP and MFTP datasets showed that iMFP-LG outperformed the state-of-the-art methods, especially for small and multi-functional peptide categories. iMFP-LG is also interpretable by visualizing the patterns captured by the attention mechanisms of pLM and GAT. Subsequently, a peptide discovery pipeline was established based on iMFP-LG to screen for novel multi-functional peptides. Eight candidate peptides with both anti-microbial and anti-cancer functions were discovered from the UniRef90 database. Further biological experiments demonstrated the promising anti-cancer and anti-bacterial activities of the candidate peptides, indicating that iMFP-LG has a strong potential to discover novel multi-functional peptides.

In future studies, we plan to integrate function-related features, structure information, and physic-chemical properties to strengthen the capability of graph nodes. Graph networks can be developed to delve further into the connections between various functional features [10]. Although deep learning-based predictions have achieved significant success, experimental validation of the predicted functions of peptides remains necessary for future study. In our opinion, other multi-label bioinformatics tasks [59] can effectively use our technique as an extension.

Materials and methods

Datasets

We evaluated the performance of our proposed method on two widely used multi-functional peptide datasets, MFBP [37] and MFTP [39]. Both were collected from the literature by searching specific keywords in Google Scholar. Notably, all peptide functions are based on the experimental results

A

No.	UniRef90_ID	Sequence	Length (AA)
1	UniRef90_P82904	FLPAIVGAAGQFLPKIFCAISKKC	24
2	UniRef90_B9W4V2	MRKEFHNVLSSDQLLTDKRPARDYNRK	27
3	UniRef90_P83653	CIKNGNGCQPNGSQGNCCSGYCHKQPGWVAGYCRRK	36
4	UniRef90_P81463	FVPYNPPRPGQSKPFPSFPGHGPFNPKIQWPYPLPNPGH	39
5	UniRef90_A0A493R6M6	CLGVGSCVDFAGCGYAVVCFW	21
6	UniRef90_P85078	CLGVGSCNDFAGCGYAIVCFW	21
7	UniRef90_P81865	GFFALIPKIISSPLFKTLLSAVGSALSSSGEQE	33
8	UniRef90_C0HKN4	GFGSLLGKALKIGTNLL	17

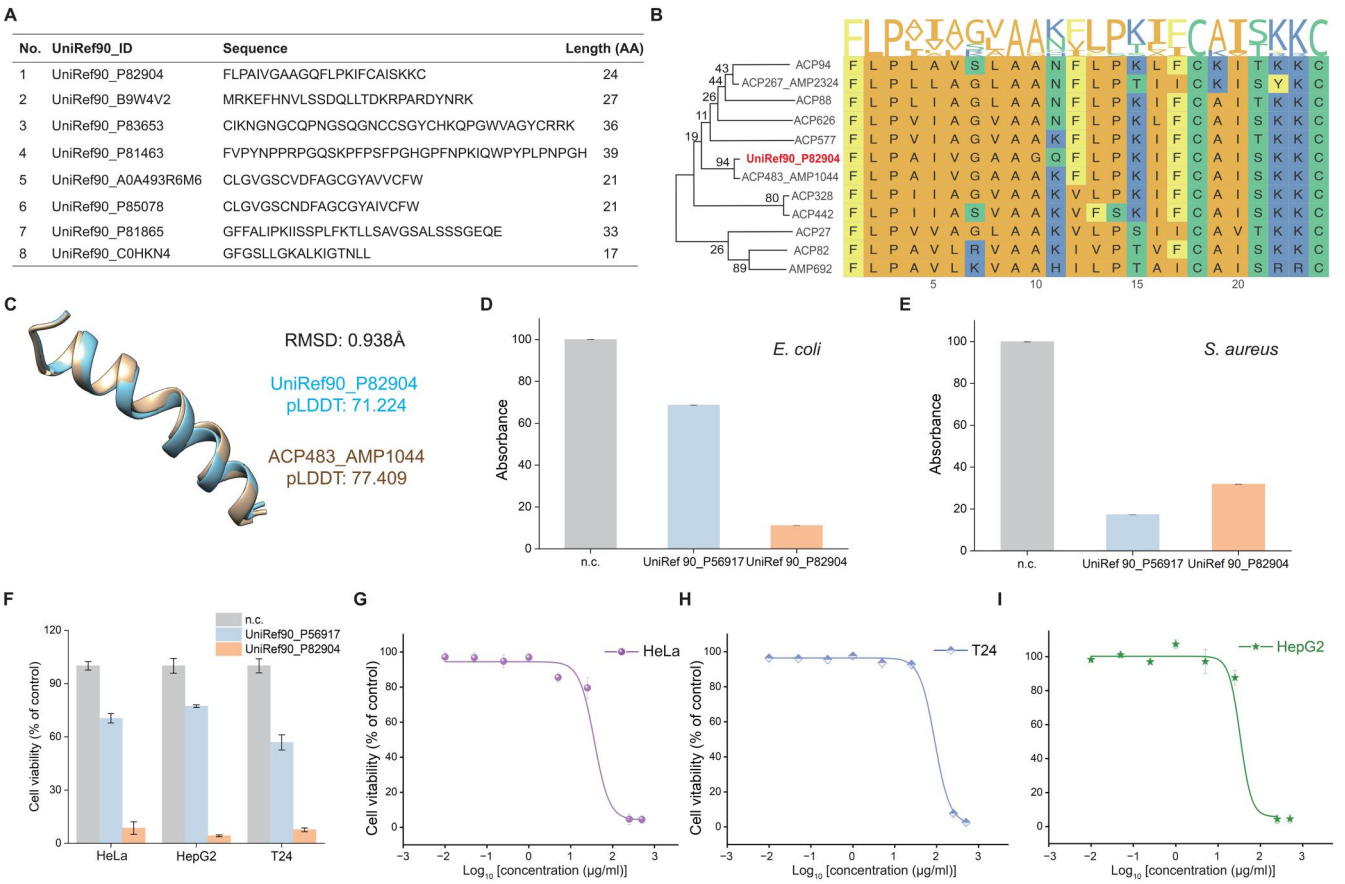


Figure 5 Multi-functional peptides discovered by iMFP-LG from UniRef90
A. Candidate peptides with both anti-cancer and anti-microbial functions screened from UniRef90 by iMFP-LG. **B**. Multiple sequence alignment and phylogenetic tree of the candidate peptide UniRef90_P82904. **C**. Structure alignment of the candidate peptide UniRef90_P82904 with its homologous sequence. **D**. and **E**. Bacterial inhibition effects of peptides UniRef90_P56917 and UniRef90_P82904 on *E. coli* (D) and *S. aureus* (E). **F**. Cytotoxic effects of peptides UniRef90_P56917 and UniRef90_P82904 on HeLa, HepG2, and T24 tumor cells. **G**.–**I**. Dose-dependent cytotoxic effects of the candidate peptide UniRef90_P82904 on HeLa (G), T24 (H), and HepG2 (I) tumor cells. AA, amino acid; n.c., negative control.

available to date. It is highly possible that some peptides have potential functions that have not yet been unveiled. We randomly sampled 80% of the data as training data, and the remaining 20% was the test data.

The MFBP dataset was collected by searching the keyword 'bioactive peptides' in Google Scholar in June 2020. It contains 5986 bioactive peptides with 5 different functional attributes, including ACP, ADP, AHP, AIP, and AMP. For each functional peptide category, CD-HIT [60] was applied to remove sequences with similarity greater than 90% in order to avoid redundancy and homology bias. The majority of bioactive peptides only have one type of activity, a few peptides have two types of activity jointly, and no peptide has more than two types of activities at the same time. The distribution of peptide categories is shown in Figure S3.

The MFTP dataset was collected by searching the keyword 'therapeutic peptides' in Google Scholar in July 2021. The collected data were preprocessed using following criteria: (1) sequences containing non-standard amino acids were discard; (2) peptides longer than 50 amino acids or shorter than 5 amino acids were removed [61], since more than 97% of functional peptides are shorter than 50 amino acids according to the distribution of sequence lengths of AMPs in the APD3 dataset [62]; and (3) peptides with fewer than 40 samples were deleted. After data processing, there were 9874 therapeutic peptides with 21 different functional attributes, including anti-angiogenic peptide (AAP), ABP, ACP, anti-coronavirus peptide (ACVP), ADP, AEP, anti-fungal peptide (AFP), AHIVP, AHP, AIP, AMRSAP, anti-parasitic peptide (APP), anti-tubercular peptide (ATP), anti-viral peptide (AVP), blood-brain barrier peptide (BBP), biofilm-inhibitory peptide (BIP), cell-penetrating peptide (CPP), DPPIP, quorum-sensing peptide (QSP), surface binding peptide (SBP), and tumor homing peptide (THP). The distribution of peptides for each type is shown in Figure S4.

Peptide representation module

Here, the pre-trained pLM TAPE [49] was employed to generate peptide representations. TAPE is a bidirectional encoder representation from a transformer-based language model [31] constructed with a 12-layer Transformer encoder and 12 attention heads in each layer. The core of attention mechanism can be formulated as Equations 1 and 2:

$$Q = XW^Q,\ K = XW^K,\ V = XW^V \tag{1}$$

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

where $X \in R^{L\times d_m}$ is an embedding of a peptide sequence. The sequence embedding X is transformed to a query matrix $Q \in R^{L\times d_k}$, a key matrix $K \in R^{L\times d_k}$, and a value

matrix $V \in R^{L \times d_k}$ by linear transformation, where W^Q, W^K, $W^V \in R^{d_m \times d_k}$ are the learnable parameters of the attention layer. The attention scores are the normalized dot product of the key and query vectors in Equation 2. The multi-head attention is an extension of the attention mechanism to catch rich information from multiple projections, and it can be formulated as Equations 3 and 4:

$$head_i = Attention\left(XW_i^Q, XW_i^K, XW_i^V\right) \tag{3}$$

$$MultiheadAttention(Q,K,V) = Concat(head_i, \ldots, head_h)W^O \tag{4}$$

where $W_i^Q, W_i^K, W_i^V \in R^{d_m \times d_k}$ are the learnable parameters in $head_i$. The output of the multi-head attention layer is obtained by concatenating the outputs of all attention heads, followed by a linear transformation using $W^O \in R^{hd_v \times d_m}$.

TAPE was pre-trained by masked-token prediction on the Pfam [63] corpus, which contains more than 31 million protein domains. The aim of masked-token prediction is to predict randomly masked amino acids based on other amino acids in the protein. Thus, TAPE can model the general relationships between amino acid residues. In this study, the parameters in TAPE were trainable and updated during the training process of the whole framework. Subsequently, the pLM was fine-tuned to identify multi-functional peptides in the training datasets. We used the output of the pooler layer in TAPE as peptide representations with dimension of 768.

Graph classification module

The graph classification module is used to transform the multi-label classification problem into a graph node classification problem. It consists of three parts: a node feature encoder, a GAT, and a node classifier.

The node feature encoder constructs a linear transform layer for each graph node to convert the peptide representation to the corresponding node representation, which can be formulated as Equation 5:

$$h_i = dropout(W_i \cdot p) \tag{5}$$

where $h_i \in R^{d_{pLM}}$ represents the feature of node i, $p \in R^{d_{pLM}}$ is the peptide representation, and $W_i \in R^{d_{pLM} \times d_{pLM}}$ is the trainable parameter of the linear transformation layer. Specifically, a peptide representation with the dimension of 768 was converted into 5 node representations for the MFBP dataset and 21 node representations for the MFTP dataset. All node representations are dimension of 768.

The GAT is a multi-head attention-based graph neural network for fine-tuning node representation by learning relationships among peptide functions. The graph nodes represent the different peptide functions and the edges represent the associations between two functions. The node representation is updated by the GAT as Equation 6:

$$h_i' = \sum_{j \in N(i)} \alpha_{ij} W h_j \tag{6}$$

where h_i' represents the updated node representation, $W \in R^{d \times d_{pLM}}$ is the transformation matrix, α represents the attention matrix, and α_{ij} indicates the importance of the node j to node i. It can be calculated as Equation 7:

$$\alpha_{ij} = \frac{exp\left(LeakyReLU\left(\alpha^T[Wh_i \| Wh_j]\right)\right)}{\sum_{k \in N(i)} exp\left(LeakyReLU\left(\alpha^T[Wh_i \| Wh_k]\right)\right)} \tag{7}$$

where $\alpha^T \in R^{2d}$ is the parameter matrix of the attention mechanism, and $\|$ denotes the concatenation operation. Multi-head attention [64] is used in GAT. The final node features are obtained by concatenating K independent attention heads, which can be formulated as Equation 8:

$$h_i' = \|_{k=1}^K \sigma\left(\sum_{j \in N(i)} \alpha_{ij}^k W^k h_j\right) \tag{8}$$

where K is the number of attention heads, σ represents the activation functions, $\|$ denotes the concatenation operation, α_{ij}^k and W^k are attention coefficients and weight matrix in k-th attention mechanism. In this study, we created 2-layer fully connected graphs among function labels, with all edge weights initialized to 1 and each layer having 6 attention heads.

The node classifier is a set of binary predictors used for the final prediction of function labels. The updated node representations from GAT are fed into the corresponding binary predictors to predict whether the peptide has the corresponding function or not. In this study, each binary predictor has a hidden layer and an output layer with sigmoid activation function, which can be formulated as Equation 9:

$$res_i = sigmoid\left(W_i \cdot h_i'\right) \tag{9}$$

where $W_i \in R^{1 \times d_{pLM}}$ is learnable in the hidden layer. In practice, we concatenated the outcomes of each classification to calculate the loss using the binary cross-entropy loss function.

Adversarial training

In order to improve the protein language representations' capability and avoid the overfitting phenomenon, we employed an adversarial training strategy called Fast Gradient Method (FGM) [65,66] during the training process.

FGM introduces an adversarial perturbation to the embeddings of amino acids according to the updated gradients. The adversarial perturbation r_{adv} can be defined as Equations 10 and 11.

$$r_{adv} = -\epsilon g / \|g\|_2 \tag{10}$$

$$g = \nabla_e \log p\left(y|e;\hat{\theta}\right) \tag{11}$$

where e is the embedding of peptide sequences; y represents true function labels; $p(y|e;\theta)$ represents the conditional probability of y given e; θ represents the parameter of the model, and $\hat{\theta}$ is a constant set to the current parameters of the model; and ϵ represents the shared norm constraint of adversarial loss.

The goal of adversarial training is to minimize the original loss without perturbation and the adversarial loss with the adversarial perturbation (Figure 1E). The procedure of these two goals can be formulated as Equations 12 and 13, respectively.

$$L_{init}(\theta) = -\frac{1}{N}\sum_{i=1}^{N} y_i \log p(y_i|x_i;\theta) \quad (12)$$

$$L_{adv}(\theta) = -\frac{1}{N}\sum_{i=1}^{N} y_i \log p\left(y_i|x_i + r_{adv,i};\theta\right) \quad (13)$$

where N is the number of batch size, x_i, y_i, and $r_{adv,i}$ are the input peptide sequence, label, and perturbation on the i-th sample in a batch, respectively, and θ denotes the model's parameters.

Evaluation metrics

In order to make a full and fair comparison with the state-of-the-art methods, we adopt five evaluation metrics, including precision, coverage, accuracy, absolute true, and absolute false. These metrics are defined as Equations 14–18:

$$Precision = \frac{1}{N}\sum_{i=1}^{N}\frac{\|L_i \cap L_i^*\|}{\|L_i^*\|} \quad (14)$$

$$Coverage = \frac{1}{N}\sum_{i=1}^{N}\frac{\|L_i \cap L_i^*\|}{\|L_i\|} \quad (15)$$

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N}\frac{\|L_i \cap L_i^*\|}{\|L_i \cup L_i^*\|} \quad (16)$$

$$Absolute\ true = \frac{1}{N}\sum_{i=1}^{N}\triangle\left(L_i, L_i^*\right) \quad (17)$$

$$Absolute\ false = \frac{1}{N}\sum_{i=1}^{N}\frac{\|L_i \cup L_i^*\| - \|L_i \cap L_i^*\|}{M} \quad (18)$$

where N represents the total number of peptide sequences in the dataset, M denotes the number of labels, $\cap$ and $\cup$ are the intersect and union operations in the set theory, $\|S\|$ represents the size of a set S, L_i denotes the true label subset of the i-th peptide sample, L_i^* represents the predict label subset of i-th sample by the classifier, and $\triangle(L_i, L_i^*)$ can be formulated as Equation 19:

$$\triangle(L_i, L_i^*) = \begin{cases} 0, & \text{if } L_i \text{ is identical to } L_i^* \\ 1, & \text{other} \end{cases} \quad (19)$$

We also employed the sensitivity and specificity metrics [37,67] to further compare the performance on each peptide function, which can be formulated as Equations 20 and 21:

$$Sensitivity = \frac{TP}{TP + FN} \quad (20)$$

$$Specificity = \frac{TN}{TN + FP} \quad (21)$$

where the numbers of true positives, true negatives, false positives, and false negatives are denoted by TP, TN, FP, and FN, respectively. When calculating the sensitivity and specificity of a specific functional peptide, that class of peptide is considered a positive sample and other peptides without that function are considered negative samples.

Implementation details

Our proposed model was implemented using PyTorch1.12 in a computing server equipped with an Intel(R) Xeon(R) Gold 6248R CPU @ 3.00 GHz and an NVIDIA A100 GPU. The proposed model was trained in 100 epochs with batch size 32 and AdamW optimizer [68]. Since the pLM was already pre-trained, we set a small learning rate of 5E–5 to fine-tune it. The learning rates of the GAT were 1E–3 and 5E–4 for the MFBP and MFTP datasets respectively. We constructed fully-connected graphs with 5 and 21 nodes for the MFBP and MFTP datasets, respectively, and both initialized with edge weights of 1. The number of attention heads in GAT was 6 and the dimension of node features in each attention head was 128. The final node features in GAT had a dimension of 768, the same as the representation obtained from the language model. The norm constraint ϵ of adversarial training was set to 0.5. In order to reduce the effects caused by the random initialization of the deep learning framework and to maintain a consistent setting with the compared methods [37,39], all models were trained 10 times repeatedly, and the prediction results were averaged as the final prediction for testing samples.

Peptide synthesis and biological experiments

Peptide synthesis

The peptides used in this study were synthesized via solid-phase peptide synthesis (Beijing Liuhe Bada Gene Technology, China), and their precise molecular weights were determined using mass spectrometry. The purity of all peptides was assessed by high-performance liquid chromatography, and all samples showing a purity greater than 90%.

Bacterial inhibition experiment

An *S. aureus* strain was streaked on Luriae-Bertani (LB) agar medium and incubated at 37°C overnight. An individual colony was picked into LB culture medium and shaken at 120 r/min at 37°C overnight. The LB bacterial suspension was diluted to the predetermined starting concentration [optical density at 600 nm (OD600) = 0.1] and then further diluted 1000 times for the inhibition test. Freeze-dried peptide powder was thawed and dissolved in double-distilled water to 50 mM. Three experimental groups were set up to test peptide anti-bacterial activity: (1) blank control group, 50 μl of LB solution; (2) bacterial control group, 25 μl of LB solution and 25 μl of bacterial solution; and (3) peptide group, 23 μl of LB solution, 25 μl of bacterial solution, and 2 μl of peptide solution (500 μM). All experiments were performed on 96-well plates with each single well containing 50 μl of final volume. After culture at 37°C for 12 h, the absorbance value of each well was determined by using a microplate reader at OD600. All experiments were performed with three independent replicates.

Tumor inhibition experiment

Tumor inhibition effects of peptides were determined using MTT cytotoxicity assay for T24, HeLa, and HepG2 cell lines. Exponentially growing cells were seeded in 96-well microtiter plates at a density of approximately 5×10^3 cells per well. After a 24-h incubation at 37°C with 5% CO_2, the medium was replaced with fresh medium, and peptides at final concentration of 500 μg/ml each were added, followed by 24-h incubation. Cell viability was monitored by the addition of 5 mg/ml MTT solution, and the absorbance value was measured at OD570 after incubation for 4 h. Each peptide experiment was performed with three replicates. The optical density of wells containing cells cultured without peptides was assumed to represent 100% cell viability.

Code availability

The associated code is available at GitHub (https://github.com/chen-bioinfo/iMFP-LG). The code has also been submitted to BioCode at the National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation (BioCode: BT007494), which is publicly accessible at https://ngdc.cncb.ac.cn/biocode/tools/BT007494.

CRediT author statement

Jiawei Luo: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Visualization, Writing – original draft. **Kejuan Zhao:** Resources, Visualization, Formal analysis, Writing – original draft. **Junjie Chen:** Conceptualization, Formal analysis, Methodology, Funding acquisition, Supervision, Writing – original draft, Writing – review & editing. **Caihua Yang:** Methodology, Resources. **Fuchuan Qu:** Formal analysis. **Yumeng Liu:** Writing – review & editing. **Xiaopeng Jin:** Writing – review & editing. **Ke Yan:** Writing – review & editing. **Yang Zhang:** Supervision, Writing – review & editing. **Bin Liu:** Writing – review & editing. All authors have read and approved the final manuscript.

Supplementary material

Supplementary material is available at *Genomics, Proteomics & Bioinformatics* online (https://doi.org/10.1093/gpbjnl/qzae084).

Competing interests

The authors have declared no competing interests.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant Nos. 62102118, 62302316, and 62302317), the Shenzhen Science and Technology Program (Grant No. JCYJ20230807094318038), the Shenzhen Colleges and Universities Stable Support Program (Grant Nos. GXWD20220811170504001 and 20220715183602001), and the Natural Science Foundation of Top Talent of SZTU (Grant No. GDRC202319), China. We also acknowledge Joanna Siebert for proofreading to improve the overall clarity and readability of the text.

ORCID

0009-0001-9965-296X (Jiawei Luo)
0009-0001-6842-1309 (Kejuan Zhao)
0000-0002-0483-303X (Junjie Chen)
0009-0002-9766-6652 (Caihua Yang)
0009-0001-4461-1378 (Fuchuan Qu)
0009-0009-0888-6575 (Yumeng Liu)
0000-0001-9481-3309 (Xiaopeng Jin)
0000-0001-6314-0762 (Ke Yan)
0000-0002-3503-5161 (Yang Zhang)
0000-0002-8520-8374 (Bin Liu)

References

[1] Chu Y, Zhang Y, Wang Q, Zhang L, Wang X, Wang Y, et al. A transformer-based model to predict peptide–HLA class I binding and optimize mutated peptides for vaccine design. Nat Mach Intell 2022;4:300–11.
[2] Sánchez A, Vázquez A. Bioactive peptides: a review. Food Qual Saf 2017;1:29–46.
[3] Fosgerau K, Hoffmann T. Peptide therapeutics: current status and future directions. Drug Discov Today 2015;20:122–8.
[4] Mei S, Li F, Leier A, Marquez-Lago TT, Giam K, Croft NP, et al. A comprehensive review and performance evaluation of bioinformatics tools for HLA class I peptide-binding prediction. Brief Bioinform 2020;21:1119–35.
[5] Yan K, Lv H, Guo Y, Chen Y, Wu H, Liu B. TPpred-ATMV: therapeutic peptide prediction by adaptive multi-view tensor learning model. Bioinformatics 2022;38:2712–8.
[6] Usmani SS, Bedi G, Samuel JS, Singh S, Kalra S, Kumar P, et al. THPdb: database of FDA-approved peptide and protein therapeutics. PLoS One 2017;12:e0181748.
[7] Zhang Y, Zhu G, Li K, Li F, Huang L, Duan M, et al. HLAB: learning the BiLSTM features from the ProtBert-encoded proteins for the class I HLA-peptide binding prediction. Brief Bioinform 2022;23:bbac173.
[8] Hoskin DW, Ramamoorthy A. Studies on anticancer activities of antimicrobial peptides. Biochim Biophys Acta Biomembr 2008;1778:357–75.
[9] Du Z, Comer J, Li Y. Bioinformatics approaches to discovering food-derived bioactive peptides: reviews and perspectives. Trends Analyt Chem 2023;162:117051.
[10] Yan K, Lv H, Guo Y, Peng W, Liu B. sAMPpred-GAT: prediction of antimicrobial peptide by graph attention network and predicted peptide structure. Bioinformatics 2023;39:btac715.
[11] Wardah W, Dehzangi A, Taherzadeh G, Rashid MA, Khan MG, Tsunoda T, et al. Predicting protein-peptide binding sites with a deep convolutional neural network. J Theor Biol 2020;496:110278.
[12] Romero-Molina S, Ruiz-Blanco YB, Mieres-Perez J, Harms M, Münch J, Ehrmann M, et al. PPI-Affinity: a web tool for the prediction and optimization of protein–peptide and protein–protein binding affinity. J Proteome Res 2022;21:1829–41.
[13] Thakur N, Qureshi A, Kumar M. AVPpred: collection and prediction of highly effective antiviral peptides. Nucleic Acids Res 2012;40:W199–204.
[14] Zhang W, Xia E, Dai R, Tang W, Bin Y, Xia J. PredAPP: predicting anti-parasitic peptides with undersampling and ensemble approaches. Interdiscip Sci 2022;14:258–68.
[15] Manavalan B, Shin TH, Kim MO, Lee G. AIPpred: sequence-based prediction of anti-inflammatory peptides using random forest. Front Pharmacol 2018;9:348997.
[16] Shoombuatong W, Schaduangrat N, Pratiwi R, Nantasenamat C. THPep: a machine learning-based approach for predicting tumor homing peptides. Comput Biol Chem 2019;80:441–51.
[17] Agrawal P, Bhagat D, Mahalwal M, Sharma N, Raghava GP. AntiCP 2.0: an updated model for predicting anticancer peptides. Brief Bioinform 2021;22:bbaa153.
[18] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–402.
[19] Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M. AAindex: amino acid index database, progress report 2008. Nucleic Acids Res 2008;36:D202–5.
[20] Shen HB, Chou KC. PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition. Anal Biochem 2008;373:386–8.
[21] Yu L, Jing R, Liu F, Luo J, Li Y. DeepACP: a novel computational approach for accurate identification of anticancer peptides by deep learning algorithm. Mol Ther Nucleic Acids 2020;22:862–70.

[22] Su X, Xu J, Yin Y, Quan X, Zhang H. Antimicrobial peptide identification using multi-scale convolutional network. BMC Bioinformatics 2019;20:730.
[23] Veltri D, Kamath U, Shehu A. Deep learning improves antimicrobial peptide recognition. Bioinformatics 2018;34:2740–7.
[24] Ma Y, Guo Z, Xia B, Zhang Y, Liu X, Yu Y, et al. Identification of antimicrobial peptides from the human gut microbiome using deep learning. Nat Biotechnol 2022;40:921–31.
[25] Lv Z, Cui F, Zou Q, Zhang L, Xu L. Anticancer peptides prediction with deep representation learning features. Brief Bioinform 2021;22:bbab008.
[26] Ahmed S, Muhammod R, Khan ZH, Adilina S, Sharma A, Shatabda S, et al. ACP-MHCNN: an accurate multi-headed deep-convolutional neural network to predict anticancer peptides. Sci Rep 2021;11:23676.
[27] Cai L, Wang L, Fu X, Xia C, Zeng X, Zou Q. ITP-Pred: an interpretable method for predicting, therapeutic peptides with fused features low-dimension representation. Brief Bioinform 2021; 22:bbaa367.
[28] Park HW, Pitti T, Madhavan T, Jeon YJ, Manavalan B. MLACP 2.0: an updated machine learning tool for anticancer peptide prediction. Comput Struct Biotechnol J 2022;20:4473–80.
[29] Ghulam A, Ali F, Sikander R, Ahmad A, Ahmed A, Patil S. ACP-2DCNN: deep learning-based model for improving prediction of anticancer peptides using two-dimensional convolutional neural network. Chemometr Intell Lab Syst 2022;226:104589.
[30] Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training, 2018. https://gwern.net/doc/www/s3-us-west-2.amazonaws.com/d73fdc5ffa8627bce44dcda2fc012da638ffb158.pdf.
[31] Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics 2019:4171–86.
[32] Du Z, Ding X, Xu Y, Li Y. UniDL4BioPep: a universal deep learning architecture for binary classification in peptide bioactivity. Brief Bioinform 2023;24:bbad135.
[33] Zhang Y, Lin J, Zhao L, Zeng X, Liu X. A novel antibacterial peptide recognition algorithm based on BERT. Brief Bioinform 2021;22:bbab200.
[34] Teufel F, Almagro Armenteros JJ, Johansen AR, Gíslason MH, Pihl SI, Tsirigos KD, et al. SignalP 6.0 predicts all five types of signal peptides using protein language models. Nat Biotechnol 2022;40:1023–5.
[35] Du Z, Ding X, Hsu W, Munir A, Xu Y, Li Y. pLM4ACE: a protein language model based predictor for antihypertensive peptide screening. Food Chem 2024;431:137162.
[36] Charoenkwan P, Nantasenamat C, Hasan MM, Manavalan B, Shoombuatong W. BERT4Bitter: a bidirectional encoder representations from transformers (BERT)-based model for improving the prediction of bitter peptides. Bioinformatics 2021; 37:2556–62.
[37] Tang W, Dai R, Yan W, Zhang W, Bin Y, Xia E, et al. Identifying multi-functional bioactive peptide functions using multi-label deep learning. Brief Bioinform 2022;23:bbab414.
[38] Li Y, Li X, Liu Y, Yao Y, Huang G. MPMABP: a CNN and Bi-LSTM-based method for predicting multi-activities of bioactive peptides. Pharmaceuticals (Basel) 2022;15:707.
[39] Yan W, Tang W, Wang L, Bin Y, Xia J. PrMFTP: multi-functional therapeutic peptides prediction based on multi-head self-attention mechanism and class weight optimization. PLoS Comput Biol 2022;18:e1010511.
[40] Fan H, Yan W, Wang L, Liu J, Bin Y, Xia J. Deep learning-based multi-functional therapeutic peptides prediction with a multi-label focal dice loss function. Bioinformatics 2023;39:btad334.
[41] Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci U S A 2021; 118:e2016239118.
[42] Madani A, Krause B, Greene ER, Subramanian S, Mohr BP, Holton JM, et al. Large language models generate functional protein sequences across diverse families. Nat Biotechnol 2023; 41:1099–106.
[43] Wang R, Jin J, Zou Q, Nakai K, Wei L. Predicting protein–peptide binding residues via interpretable deep learning. Bioinformatics 2022;38:3351–60.
[44] Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, et al. Prottrans: toward understanding the language of life through self-supervised learning. IEEE Trans Pattern Anal Mach Intell 2021;44:7112–27.
[45] Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y. Graph attention networks. The 6th International Conference on Learning Representations 2017.
[46] Chen Z, Liu X, Zhao P, Li C, Wang Y, Li F, et al. iFeatureOmega: an integrative platform for engineering, visualization and analysis of features from molecular sequences, structural and ligand data sets. Nucleic Acids Res 2022;50:W434–47.
[47] Liu B, Xu J, Lan X, Xu R, Zhou J, Wang X, et al. iDNA-Prot|dis: identifying DNA-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition. PLoS One 2014; 9:e106691.
[48] Govindan G, Nair AS. Composition, Transition and Distribution (CTD)—a dynamic feature for predictions based on hierarchical structure of cellular sorting. 2011 Annual IEEE India Conference 2011:1–6.
[49] Rao R, Bhattacharya N, Thomas N, Duan Y, Chen P, Canny J, et al. Evaluating protein transfer learning with TAPE. Proceedings of the 32nd International Conference on Neural Information Processing Systems 2019.
[50] Fürnkranz J, Hüllermeier E, Loza Mencía E, Brinker K. Multilabel classification via calibrated label ranking. Mach Learn 2008;73:133–53.
[51] Tsoumakas G, Vlahavas I. Random *k*-labelsets: an ensemble method for multilabel classification. The 18th European Conference on Machine Learning 2007:406–17.
[52] Wu G, Zheng R, Tian Y, Liu D. Joint ranking SVM and binary relevance with robust low-rank learning for multi-label classification. Neural Netw 2020;122:24–39.
[53] Yang L, Wu XZ, Jiang Y, Zhou ZH. Multi-label learning with deep forest. The 24th European Conference on Artificial Intelligence 2019.
[54] van der Maaten L, Hinton G. Visualizing data using *t*-SNE. J Mach Learn Res 2008;9:2579–605.
[55] Vig J. BertViz: a tool for visualizing multihead self-attention in the BERT model. ICLR workshop: debugging machine learning models 2019.
[56] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–10.
[57] Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 2023;379:1123–30.
[58] Simmaco M, Mignogna G, Canofeni S, Miele R, Mangoni ML, Barra D. Temporins, antimicrobial peptides from the European red frog *Rana temporaria*. Eur J Biochem 1996; 242:788–92.
[59] Wu Z, Guo M, Jin X, Chen J, Liu B. CFAGO: cross-fusion of network and attributes based on attention mechanism for protein function prediction. Bioinformatics 2023;39:btad123.
[60] Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 2012; 28:3150–2.
[61] Cui Z, Wang SG, He Y, Chen ZH, Zhang Q. DeepTPpred: a deep learning approach with matrix factorization for predicting

therapeutic peptides by integrating length information. IEEE J Biomed Health Inform 2023;27:4611–22.
[62] Wang G, Li X, Wang Z. APD3: the antimicrobial peptide database as a tool for research and education. Nucleic Acids Res 2016;44:D1087–93.
[63] El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families database in 2019. Nucleic Acids Res 2019;47:D427–32.
[64] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Proceedings of the 30th International Conference on Neural Information Processing Systems 2017.
[65] Jin J, Yu Y, Wang R, Zeng X, Pang C, Jiang Y, et al. iDNA-ABF: multi-scale deep biological language learning model for the interpretable prediction of DNA methylations. Genome Biol 2022; 23:219.
[66] Miyato T, Dai AM, Goodfellow I. Adversarial training methods for semi-supervised text classification. The 5th International Conference on Learning Representations 2017.
[67] Li J, Zhao T, Zhang Y, Zhang K, Shi L, Chen Y, et al. Performance evaluation of pathogenicity-computation methods for missense variants. Nucleic Acids Res 2018;46:7793–804.
[68] Loshchilov I, Hutter F. Decoupled weight decay regularization. The 7th International Conference on Learning Representations 2019.

Genomics, Proteomics & Bioinformatics, 2024, 22, 1–14
https://doi.org/10.1093/gpbjnl/qzae084
Original Research