PLOS Computational Biology

RESEARCH ARTICLE

Multi-ACPNet: A multi-scale sequence-structure feature fusion framework for anticancer peptide identification and functional prediction

Lu Meng[1,2*], **Lijun Zhou**[1]

1 College of Information Science and Engineering, Northeastern University, Shenyang, China, **2** The Foshan Graduate School of Innovation, Northeastern University, Shenyang, China

* menglu1982@gmail.com

OPEN ACCESS

Citation: Meng L, Zhou L (2026) Multi-ACPNet: A multi-scale sequence-structure feature fusion framework for anticancer peptide identification and functional prediction. PLoS Comput Biol 22(3): e1014053. https://doi.org/10.1371/journal.pcbi.1014053

Editor: Belal O. Al-najjar, Al-Ahliyya Amman University, JORDAN

Received: November 12, 2025

Accepted: February 23, 2026

Published: March 10, 2026

Data availability statement: The implementation of Multi-ACPNet is publicly available at https://github.com/zlj-zlj/Multi_ACPNet. A freely accessible prediction web server is available at http://www.isme.ln.cn:5001/ACPs_predict to facilitate ACP research.

Abstract

Anticancer peptides (ACPs) have emerged as promising therapeutic candidates for cancer treatment due to their high efficacy and low propensity for inducing drug resistance. However, existing ACP identification methods primarily rely on peptide sequence features while neglecting spatial structural characteristics. Moreover, few approaches can simultaneously predict the functional activity of ACPs. To address these limitations, this study proposes Multi-ACPNet, a novel dual-function predictor capable of both ACP identification and activity type classification. This model innovatively integrates sequence and structural features through a multi-stage framework. It employs a hybrid Bidirectional Long Short-Term Memory (BiLSTM) and causal convolutional network to capture both long-range dependencies and local sequence patterns, followed by a multi-scale Graph Convolutional Network (GCN) that dynamically fuses local and long-range structural dependencies using residual connections and adaptive weighting. Experimental results demonstrate that Multi-ACPNet achieves outstanding performance, with Accuracy of 0.8140, 0.9536, and 0.8770 on three benchmark datasets for ACP identification. For functional prediction, it attains an AUC of 0.9033, F1-score of 0.8472, and Hamming loss of 0.1303, significantly outperforming state-of-the-art predictors.

Author summary

Anticancer peptides (ACPs) have emerged as highly promising therapeutic candidates in cancer treatment due to their high efficacy and low propensity for drug resistance. However, existing prediction methods suffer from two major limitations: first, they predominantly rely on sequence information while neglecting three-dimensional structural features, leading to the loss of crucial spatial interaction information; second, their predictive capability is insufficient—most current models are limited to binary classification tasks for ACP identification and cannot

Funding: This study was supported by National Natural Science Foundation of China (62073061 to LM), Guangdong Basic and Applied Basic Research Foundation (2025A1515011602 to LM). In this study, the funders took the roles of project administrator, funding provider, study design, preparation of the manuscript. The funders had no role in data collection and analysis, decision to publish.

Competing interests: The authors have declared that no competing interests exist.

predict their targeting activities against specific cancer cell lines. To address these challenges, we developed Multi-ACPNet, an end-to-end framework that integrates both sequence and structural information for dual-task collaborative prediction. The model innovatively incorporates a multi-scale information fusion mechanism: at the sequence level, it combines Bidirectional Long Short-Term Memory (BiLSTM) with causal convolution to simultaneously capture long-range dependencies and local functional motifs; at the structural level, a multi-hop Graph Convolutional Network is constructed to dynamically learn multi-level spatial interactions within peptide molecules. This architecture not only significantly improves the accuracy of ACP identification but also enables efficient multi-label prediction of functional activities across seven cancer cell types. Furthermore, the model demonstrates excellent generalization capability, achieving competitive performance in extended tasks such as peptide toxicity prediction.

1. Introduction

Cancer remains one of the most severe threats to global human health. Current clinical interventions primarily include surgery, radiotherapy, chemotherapy, and targeted therapy. However, these conventional treatments are significantly limited by tumor heterogeneity, drug resistance, and severe side effects. In recent years, anticancer peptides (ACPs) have emerged as novel therapeutic agents, demonstrating superior efficacy, high selectivity, strong specificity, and excellent tumor penetration capabilities. These advantages allow ACPs to overcome many limitations of traditional therapies, positioning them as promising candidates for cancer treatment [1,2]. Nevertheless, only a limited number of ACPs have been identified, with even fewer approved for clinical use. To date, only 28 ACP-based drugs have been approved by the U.S. Food and Drug Administration (FDA) and the European Medicines Agency (EMA) [3]. Consequently, the development of novel and potent ACPs has become a critical research direction in cancer therapeutics.

ACPs are short peptide molecules that exhibit significant antitumor activity. Traditional ACP discovery relies on experimental screening, which is time-consuming, costly, and inefficient. With the rapid advancement of bioinformatics and artificial intelligence (AI), computational prediction methods have emerged as a powerful tool for accelerating ACP prescreening. These approaches streamline the labor-intensive discovery process [4,5] and significantly improve the efficiency of peptide drug candidate identification.

Current AI-based methods for ACP identification mainly consist of two key steps: feature encoding based on sequence information, followed by classifier training and selection. Traditional machine learning approaches still dominate this field [6]. Tyagi et al. [7] developed the first ACP predictor using Support Vector Machine (SVM) as the classifier. Agrawal et al. [8] extracted multiple input features and implemented six machine learning classifiers on two established datasets. Karim et al. [9] proposed ANNprob-ACPs, where they evaluated 33 machine learning models, selected the six

best-performing ones, and integrated their probability scores as input to an Artificial Neural Network (ANN) meta model, achieving significantly improved classification accuracy.

ACP-DL [10] pioneered the first deep learning framework for ACP identification by integrating binary profile features (BPF) and reduced-alphabet k-mer sparse matrices of peptide sequences, utilizing a Long Short-Term Memory (LSTM) network to automatically discriminate between ACPs and non-ACPs, achieving superior performance on benchmark datasets. Liu et al. [11] developed a parallel architecture combining Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) for learning sequence embeddings through ensemble learning. Zhang et al. [12] proposed an end-to-end ACP identification model utilizing both attention-augmented convolutional neural networks (AAConv) and standard CNNs for sequence feature extraction, with multi-head attention mechanisms further refining high-level features, achieving competitive classification performance. Notable contributions also include iACP-DRLF [13], ACP-OPE [14], ACP-CapsPred [15], CAPTURE [6], GRDF [16], AI4ACP [17], ACP_MS [18], ACP-ML [19], ACPred-BMF [20], ACPred-LAF [21], ME-ACP [22], ACP-CLB [23], and ACP-LSE [24], collectively advancing the field through diverse computational approaches.

Despite significant progress in ACP predictor development, several critical limitations persist. First, current models inadequately capture multi-scale feature interactions, potentially overlooking hierarchical relationships between local and global peptide characteristics. Second, although ACP tertiary structures are relatively simple, the specific residue arrangement may directly influence peptide function. Current predictors predominantly process sequence information alone, largely disregarding these essential structural features. Most notably, most methods are limited to binary classification (ACP or non-ACP), with few capable of predicting the functional types of ACPs. However, ACP research involves not only determining whether peptides have anticancer properties, but also predicting which cancer cell lines they can target, which is an equally important aspect.

To address these limitations, we propose Multi-ACPNet, a graph convolutional-based multidimensional collaborative framework that integrates sequence information, structural features, and multi-scale characteristics for comprehensive ACP prediction. Our primary contributions can be summarized as follows:

- To perform sequence analysis, we integrate Bidirectional Long Short-Term Memory (BiLSTM) and causal convolution to comprehensively capture and model both global contextual information and local functional motifs in peptide sequences.
- We further design a multi-scale residual dynamic Graph Convolutional Network (GCN) that computes adjacency matrices with varying hop distances to adaptively capture critical structural features from both the full graph and key subgraph at varying spatial scales.
- Our network is capable of dual-functional prediction. Using high-quality benchmark datasets, we first constructed and optimized a binary classifier, and further developed a multilabel classifier to accurately predict peptide sequences' specific bioactivities against different cancer cell lines, providing critical decision support for personalized therapy.
- Beyond core anticancer function prediction, we have also successfully validated our model's generalization capability on peptide toxicity prediction tasks, where it demonstrates competitive and robust performance, thereby expanding its utility for comprehensive safety assessment in therapeutic peptide discovery.

2. Methods

2.1. Datasets

For the first task (predicting whether a peptide is an ACP), we employ three datasets: the ACP-Mixed-80 [21], AntiCP 2.0_Main and AntiCP 2.0_Alternate [8]. The ACP-Mixed-80 dataset was constructed by integrating multiple benchmark datasets [8,25–28], followed by rigorous preprocessing including label verification, label correction or removal, duplicate elimination, and separation of positive and negative samples. Sequences with > 80% similarity were further removed using CD-HIT [29]. The negative samples in the last two datasets consisted of antimicrobial peptides (AMPs) lacking

anticancer properties and random peptides generated using SwissProt proteins, respectively. All three datasets were pre-partitioned into training and independent test sets (see Table A in S1 Text for detailed statistics).

For the second task involving multi-label functional prediction, we use the dataset from ACP-MLC [30], which was collected from CancerPPD [31]. The dataset was rigorously processed by retaining only functional categories with over 40 entries, removing duplicate sequences, non-linear peptides, sequences containing non-standard amino acids, and peptides > 100 amino acids in length. Additionally, sequences with over 90% similarity were filtered using CD-HIT. The final dataset comprises seven tissue-specific cancer types: Colon, Breast, Cervix, Skin, Lung, Prostate, and Blood. (detailed in Table B in S1 Text).

2.2. Overview of the model

Our Multi-ACPNet (Fig 1A) integrates peptide sequence and structural information through three key steps: (1) Feature encoding, where we comprehensively encode peptide sequences using ESM C [32] embeddings, BPF, positional encoding, and AAindex [33] features; (2) Sequence-based representation learning, implemented via a Sequence Multi-Scale Network (detailed in Fig 1B) that combines a BiLSTM and parallel CNN module to extract sequence patterns from ESM C embeddings; and (3) Structure-based geometric learning, which represents peptides as graph structures with fused features as node attributes, followed by structural analysis through a Graph Multi-Scale Network (as shown in Fig 1C). The processed features are then aggregated via global pooling and fed into fully connected (FC) layers for prediction.

2.3. Feature encoding

To encode peptide sequences as numerical vectors and enhance the model's learning capability through multi-perspective feature representation, we select four different encoding methods: ESM C pretrained embeddings, binary profile features, positional features, and AAindex features. Among these, ESM C provides global deep semantic information, while BPF and position features preserve local sequence patterns, and AAindex supplements physicochemical properties. This hybrid encoding approach improves the model's peptide sequence representation capability by integrating learned features with domain knowledge features.

In 2024, EvolutionaryScale released the ESM3 [34] series of models along with a parallel series, ESM C, which is specifically designed to capture the intrinsic biological representations of proteins. Based on the Transformer architecture, ESM C automatically learns general protein representations through unsupervised training on large-scale protein sequence data, significantly surpassing its predecessor ESM2 [35] in both predictive performance and computational efficiency. Compared to some very large models, ESM C achieves faster inference speed and higher computational efficiency with fewer parameters (e.g., the 300M and 600M versions). In this study, we used ESMC_600M as the foundational feature extraction model. The model automatically produces residue-level embeddings with dimensions $L \times 1152$, where L is the peptide sequence length. These embeddings are subsequently processed by the Sequence Multi-Scale Network.

The BPF encoding represents each amino acid residue as a 20-dimensional one-hot vector, where only the position corresponding to the specific residue type is set to 1 while others to 0. For example: Alanine (A) is encoded as $f(A) = [1, 0, 0, \ldots, 0]$ and Tyrosine (Y) is encoded as $f(Y) = [0, 0, 0, \ldots, 1]$. This encoding method transforms a peptide into a binary matrix with dimensions $L \times 20$.

Position encoding is used to introduce the position information of amino acids, generating position encoding using sine and cosine functions of different frequencies, as shown in Formulas 1 and 2.

$$PE(pos, 2i) = \sin\left(\frac{pos}{1000^{\frac{2i}{d_{model}}}}\right) \tag{1}$$

$$PE(pos, 2i + 1) = \cos\left(\frac{pos}{1000^{\frac{2i}{d_{model}}}}\right) \tag{2}$$

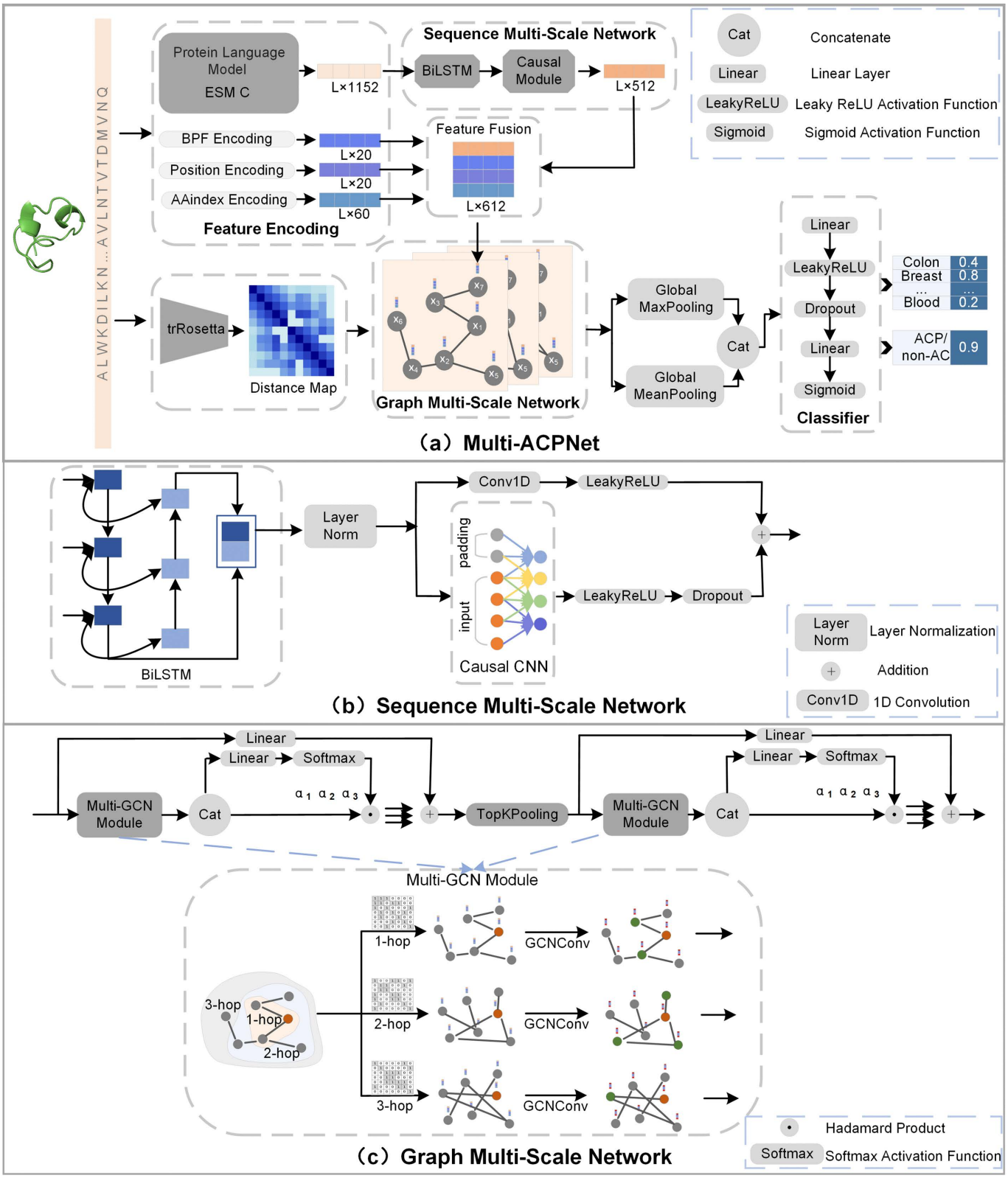

Fig 1. Overview of the proposed workflow. (A) Multi-ACPNet architecture overview. First, ESM C features are extracted and fed into the (B) Sequence Multi-Scale Network to capture multi-scale sequence features. Then, graphs are constructed using trRosetta, with domain knowledge features

and sequence-based learned features as node features. (C) The Graph Multi-Scale Network performs multi-scale structural feature learning on both the full graph and key subgraphs. The Classifier outputs the final prediction results.

https://doi.org/10.1371/journal.pcbi.1014053.g001

where, pos represents the position of amino acids in the peptide sequence, and i represents the dimension index ($(0 \leq i \leq d_{model}/2)$, where d_{model} = 20in our study corresponds to the embedding dimensionality. Based on these formulas, we obtain the position encoding with a dimension of $L \times 20$.

AAindex is a database storing the physicochemical and biochemical properties of amino acids, containing over 500 characteristics for each amino acid, such as hydrophobicity, charge, volume, etc. In AAindex version 9.2, a total of 566 physicochemical properties were collected, from which we select 553 properties without missing values. In this study, to mitigate potential issues such as redundancy, multicollinearity, and overfitting associated with high-dimensional feature usage, we employ the mRMR (Max-Relevance and Min-Redundancy) [36] algorithm to refine the original AAindex feature set. This method selects a subset of features that maximizes relevance to the target variable while minimizing inter-feature redundancy. Experimental findings in Section 3.6 indicate that the overall performance is optimal after reducing the feature dimension to 60. Beyond this point, further increasing feature dimensionality yields no substantial gains and may even lead to performance decline. This outcome confirms that mRMR-based feature selection effectively reduces noise and redundancy while retaining essential physicochemical information, thereby strengthening the model's robustness and generalization capacity. Consequently, the final AAindex feature representation has a dimensionality of $L \times 60$.

2.4. Sequence-based representation learning

ESM C is trained on massive, diverse protein sequences, acquiring universal, protein-level contextual representations. However, these representations remain relatively static and generic. Since our dual tasks, including both ACP identification and functional prediction, require learning specific, subtle functional patterns, we design a multi-scale sequence feature processing network (Sequence Multi-Scale Network; Fig 1B) to comprehensively capture multi-scale features of ACP sequences. This network dynamically transforms the general ESM C embeddings into task-specific, sequence-level representations optimized for these ACP-related tasks.

Peptide sequences and natural language both exhibit context-sensitive characteristics. BiLSTM has successfully modeled long-range dependencies between words in Natural Language Processing (NLP), and similarly can be applied to capture functional relationships between residues. In the Sequence Multi-Scale Network, the pretrained ESM C features are first fed into a BiLSTM module. In peptide sequences, residues at both the N-terminus and C-terminus may jointly determine functionality. Traditional unidirectional LSTM would lose reverse dependency information, whereas BiLSTM employs two opposing LSTMs to process the sequence. This architecture effectively models bidirectional long-range interactions between amino acid residues in the sequence. LSTM selectively remembers or forgets information in a sequence by means of unique gating mechanisms (forget, input, and output gates), thus capturing key residues in the sequence that are far apart but functionally relevant.

The functions of many ACPs rely on short, conserved local motifs, and causal convolution serves as an ideal tool for capturing such features through its local perception and historical dependency. The output of the BiLSTM is then fed into a local information processing module that adopts a dual-branch parallel architecture. The first branch uses causal convolution, LeakyReLU, and Dropout. Causal convolution maintains consistent input and output sequence lengths by zero-padding on the left side of the sequence, while ensuring that the output at each time step relies only on sequence information from the current moment and before. We use a causal convolution with a kernel size k of 3 and a zero padding of $k - 1$, guaranteeing that each output step relies only on the current and two preceding input elements in the sequence. This preserves the causality of local functions while capturing amino acid local patterns. The second branch utilizes standard 1D convolution ($k = 1$) and LeakyReLU. The two branches are combined through summation, producing an output feature of dimensions$L \times 512$, which preserves local details while integrating global context.

2.5. Structure-based geometric learning

The function of ACPs may be influenced by structural features at different scales. Therefore, we innovatively propose a framework for learning the structure of ACPs based on Graph Neural Network (GNN), which achieves multi-granularity modeling of spatial interactions in ACPs and improves prediction accuracy by integrating multi-scale neighborhoods with adaptive fusion mechanism.

2.5.1. Graph construction. We employ trRosetta [37] to predict the distance map of peptides. trRosetta predicts the probability distribution of C_β distances between residues directly from multiple sequence alignment (MSA) by end-to-end training. By comparing the experimental performance, we set a distance threshold of D_{th} = 10Å to construct the initial adjacency matrix $A \in R^{L\times L}$:

$$A_{ij} = \begin{cases} 1, & if\ d_{ij} \le D_{th} \\ 0, & otherwise \end{cases} \tag{3}$$

where d_{ij} denotes the Euclidean distance between the C_β atoms of residue i and residue j.

The node features of the graph are constructed by concatenating the outputs from the Sequence Multi-Scale Network, BPF encoding features, positional encoding features, and AAindex encoding features, generating an $L \times 612$ dimensional node feature matrix $H^{(0)}$.

To construct the multi-scale graph, multi-hop adjacency matrices are generated via matrix power operations:

$$A_{1hop} = A\ A_{2hop} = sign\left(A^2\right)\ A_{3hop} = sign\left(A^3\right) \tag{4}$$

where, $sign(\cdot)$ denotes the sign function for matrix binarization. In the k-hop adjacency matrix, a value of 1 indicates the existence of a path of length k between the two nodes.

2.5.2. Graph network architecture. We develop the Graph Multi-Scale Network (Fig 1C) to perform hierarchical graph convolutions. The network first captures multi-scale structural features across the global peptide graphs. Since the full graph may contain non-functional residues that could interfere with key feature learning, we further retain critical residue nodes to extract multi-scale features from important subgraphs. The network primarily consists of three components: the Multi-GCN Module, adaptive fusion and top-k pooling.

The Multi-GCN Module performs parallel graph convolution operations. After constructing the graphs at three scales (1-hop, 2-hop, 3-hop), graph convolutions are performed in parallel according to Formula 5 to independently process neighborhood information at each scale.

$$H_k^{(l)} = \sigma\left(\hat{D}_k^{-\frac{1}{2}}\hat{A}_k\hat{D}_k^{-\frac{1}{2}}H^{(l-1)}W_k\right) \tag{5}$$

where $H^{(l)} \in R^{L\times d'}$ and $H^{(l-1)} \in R^{L\times d}$ denote the output feature matrices of the l-th and (l-1)-th graph convolution layers respectively, $H^{(0)}$ is the initial input, σ is the LeakyReLU activation function, $\hat{A} = A + I$ is the adjacency matrix with self-loops added, $\hat{D}$ is the degree matrix, $W \in R^{d\times d'}$ is a trainable weight matrix, and $k \in \{1,\ 2,\ 3\}$ represents the graph convolution at 3 scales.

Through a single parallel graph convolution operation, we obtain three feature matrices: $H_1^{(1)} \in R^{L\times d'}, H_2^{(1)} \in R^{L\times d'}, H_3^{(1)} \in R^{L\times d'}$. To achieve adaptive fusion, we propose a dynamic weighting mechanism, calculated as follows:

$$H_{fusion} = \sum_{k=1}^{3} H_k^{(1)} \odot (\alpha_k \otimes 1^T) \tag{6}$$

where $\alpha_k \in R^{L\times 1}$is the learned attention weights, and $1^T \in R^{1\times d'}$ is a row vector with all elements equal to 1. The symbol $\otimes$ represents outer product. The weights are expanded to an $L \times d'$-dimensional matrix via the outer product, where each column of the expanded matrix is a copy of α_k. Features are scaled by the attention weights through Hadamard product $\odot$. The fused feature $H_{fusion} \in R^{L\times d'}$ is obtained by summing all features after attention-weighted scaling.

To obtain the attention weights $\alpha_1,\alpha_2,\alpha_3$, we first concatenate the features $H_k^{(1)}$ and apply a linear transformation:

$$S = \left(\left[H_1^{(1)} \parallel H_2^{(1)} \parallel H_3^{(1)}\right]\right)W + b \tag{7}$$

where $\parallel$ represents concatenation, $W \in R^{3d'\times 3}$ and $b \in R^3$ are learnable weight and bias, $S \in R^{L\times 3}$ is projected into a 3-D space corresponding to the three scales. Then, normalize using softmax to get α_k:

$$\alpha_k = \frac{\exp(S_k)}{\sum_{j=1}^{3}\exp(S_j)} \quad \forall k \in \{1,2,3\} \tag{8}$$

To ensure that critical local features are not completely covered by subsequent processing, we introduce residual connectivity:

$$H_{out} = H_{fusion} + WH^{(0)} \tag{9}$$

The above describes entire process of the Multi-GCN Module and adaptive fusion over the complete graph. To focus on critical peptide structural regions while suppressing noisy nodes, we subsequently employ top-k pooling, which scores nodes via a learnable projection vector and retains only the top-n nodes. Specifically, we preserve the top 50% highest-scoring nodes to perform multi-scale graph convolution and adaptive fusion over this subgraph.

2.5.3. Output layers. To enable downstream classification, we employ both global average pooling and global maximum pooling to form the final graph representation vector. The classifier comprises two fully connected layers for nonlinear mapping.

For the ACP recognition task, we use a single output neuron with sigmoid activation to distinguish ACPs from non-ACPs. For functional prediction, samples can simultaneously belong to multiple categories, thus, the output layer contains seven neurons with sigmoid activations, each independently computing the probability for one category.

3. Results and discussion

3.1. Evaluation metrics

In the ACP recognition task, we adopt widely used evaluation criteria including Sensitivity (SE), Specificity (SP), Accuracy (ACC), Precision, F1-score and the Matthews correlation coefficient (MCC), Area Under the ROC Curve (AUC). These metrics are defined as follows:

$$\begin{cases} SE = \frac{TP}{TP+FN} \\ SP = \frac{TN}{TN+FP} \\ ACC = \frac{TP+TN}{TP+TN+FP+FN} \quad MCC = \frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \\ \text{Precision} = \frac{TP}{TP+FP} \\ \text{F1}-\text{score} = \frac{2\times SE\times \text{Precision}}{SE+\text{Precision}} \\ AUC : \text{Area under the ROC Curve} \end{cases} \tag{10}$$

where TP (true positives) denotes the number of correctly predicted positive samples, and TN (true negatives) represents the number of correctly predicted negative samples. FP (false positives) and FN (false negatives) indicate misclassified negative and positive samples, respectively.

For the multi-label functional prediction task, we employ label-based multilabel evaluation with two averaging modes: Macro-averaging and Micro-averaging.

Macro-averaging computes metrics separately for each class and then takes the arithmetic mean:

$$\text{MacroMetric} = \frac{1}{L}\sum_{l=1}^{L} evalMetric(TP_l, FP_l, TN_l, FN_l) \tag{11}$$

where L denotes the total number of label categories. Micro-averaging computes global metrics by aggregating TP, FP, TN, and FN for all categories. The calculation process is as follows:

$$\text{MicroMetric} = evalMetric\left(\sum_{l=1}^{L} TP_l, \sum_{l=1}^{L} FP_l, \sum_{l=1}^{L} TN_l, \sum_{l=1}^{L} FN_l\right) \tag{12}$$

Additionally, we employ Hamming loss, a key multilabel classification metric that quantifies the discrepancy between predicted and ground-truth labels. It is defined as:

$$\text{Hamming Loss} = \frac{1}{N\times L}\sum_{n=1}^{N}\sum_{l=1}^{L} \mathbb{I}\left(y_{nl} \neq \hat{y}_{nl}\right) \tag{13}$$

where N denotes the number of samples, $y_{nl} \in \{0,\ 1\}$ indicates the ground-truth label of sample n for category l (1 indicates presence of the function, 0 indicates absence). $\hat{y}_{nl} \in \{0,\ 1\}$ represents the predicted label, and $\mathbb{I}(\cdot)$ is the indicator function that equals 1 for misclassified labels and 0 otherwise.

3.2. Experimental setup

We implement the Multi-ACPNet model based on the PyTorch framework, using Adam as the optimizer with an initial learning rate of 0.0001. The model computations are GPU accelerated on an NVIDIA GeForce RTX 4090, running on a Windows system with an Intel Core i9-14900K processor and 64GB RAM. The hyperparameters are optimized on the AntiCP 2.0 training set through grid search with 5-fold cross-validation (see Table C in S1 Text for details). Notably, we maintain identical hyperparameters configurations for both binary classification and multi-label prediction tasks, achieving task switching solely by modifying the output layer dimensionality. This design significantly enhances the model's extensibility while preserving its generalizability across different prediction scenarios. To prevent potential data leakage arising from sequence similarity across different datasets, we perform rigorous sequence-level redundancy removal prior to hyperparameter optimization. Specifically, to ensure that no highly similar sequences exist between the training data used for hyperparameter optimization and the independent evaluation sets, we systematically remove sequence redundancy using the CD-HIT tool. Operationally, we first merge the AntiCP 2.0 training set with the AntiCP 2.0 test set, the ACP-Mixed-80 test set, and the multi-label functional prediction dataset. CD-HIT is then executed to perform full-sequence alignment and clustering. During this process, only entries from the training set partition that exhibit >90% similarity with sequences in any evaluation set are removed, while the evaluation sets themselves remain unaltered to preserve their completeness for subsequent fair benchmarking comparisons with other methods. This procedure ensures strict sequence-level independence between the training set and all evaluation sets. Table D in S1 Text presents the sequence similarity analysis between the AntiCP 2.0 training set and each evaluation dataset. It is important to clarify that for the

multi-label functional dataset, all sequences are included in the similarity analysis because the entire dataset undergoes 10-fold cross-validation (as described in Section 3.4), meaning every sequence serves as a test sample during the evaluation process. Fig 2 displays the ROC curves of the optimal model evaluated through 5-fold cross-validation on both the AntiCP 2.0_Main and AntiCP 2.0_Alternate training datasets. The corresponding confusion matrices and key evaluation metrics are presented in Fig A in S1 Text.

To conduct the evaluation of each dataset, we train a separate model. We first load the model hyperparameters. Then, we split the training set into training and validation subsets at a 5:1 ratio. Model selection is performed based on validation performance, and the model achieving the highest Accuracy on the validation set is subsequently used for final evaluation on the independent test set. To ensure a fair comparison with existing benchmark methods (which are trained and evaluated on complete public datasets), we likewise utilize the complete original datasets. Concurrently, we systematically analyze the internal sequence similarity within each dataset: the sequence similarity between the training and test sets of ACP-Mixed-80 is below 80%, and the internal similarity of the multi-label functional prediction dataset is below 90%. These thresholds align with the dataset preprocessing standards described in Section 2.1. For AntiCP 2.0_Main and AntiCP 2.0_Alternate, a certain degree of sequence similarity exists between their training and independent test sets (details in Table D in S1 Text), but no identical sequences are present. This analysis ensures effective separation between training and test sets at the sequence-independence level, thereby maintaining evaluation fairness while effectively preventing data leakage.

3.3. Comparison of the proposed method with existing methods for classifying ACPs and non-ACPs

In the task of ACP identification, we compare Multi-ACPNet with existing ACP prediction methods, and the results of all the compared methods are obtained from their original publications. Table 1 presents the performance comparison of the proposed Multi-ACPNet with the existing advanced methods on the ACP-Mixed-80 dataset. Our method achieves superior performance across multiple key metrics: it achieves the highest Accuracy, MCC, Precision and AUC scores.

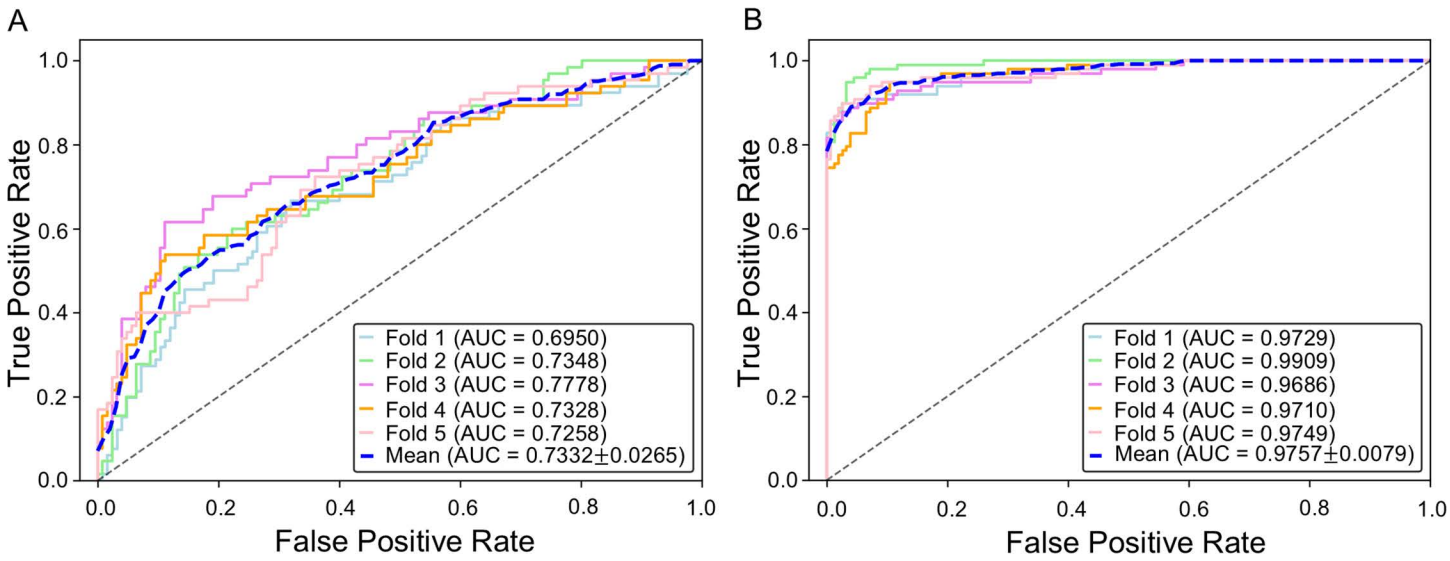

Fig 2. ROC curves of the optimal model through 5-fold cross-validation on AntiCP 2.0 training datasets. (A) Cross-validation result on the AntiCP 2.0_Main training dataset. (B) Cross-validation result on the AntiCP 2.0_Alternate training dataset.

https://doi.org/10.1371/journal.pcbi.1014053.g002

Table 1. Performance comparison between the proposed method and existing methods on the ACP-Mixed-80 dataset[a].

Methods	Year	ACC	Precision	SE	SP	MCC	AUC
AntiCP-ACC [7]	2013	0.7635	/	0.8851	0.6419	0.5433	0.1213
AntiCP-DPC [7]	2013	0.7432	/	**0.9190**	0.5676	0.5196	0.1087
ACPred-Fuse [26]	2020	0.7365	/	0.5203	**0.9527**	0.5246	0.1430
ACPred-LAF [21]	2021	0.8115	/	0.7213	0.9016	0.6333	0.8267
PreTP-EL [38]	2021	0.5820	0.5710	0.6560	0.5080	0.1660	0.3820
AI4ACP [17]	2022	0.7300	0.6750	0.8850	0.5740	0.4830	0.8140
PreTP-Stack [39]	2023	0.4920	0.4950	0.8520	0.1310	-0.0240	0.6750
ACP-MLC [30]	2023	0.7870	0.7780	0.8030	0.7700	0.5740	0.8880
CAPTURE [6]	2024	0.8400	/	/	/	0.6900	/
MA-PEP [40]	2024	0.8443	0.8387	0.8525	0.8361	0.6886	0.8987
ACP-PDAFF [41]	2024	0.8600	/	0.8200	0.9000	0.7100	/
Multi-ACPNet (ours)	2025	**0.8770**	**0.8594**	0.9016	0.8525	**0.7550**	**0.9105**

[a]All the results reported in this table except for the ones of Multi-ACPNet are obtained from publications./ means that there is no value in the corresponding item, bold indicates the highest value.

https://doi.org/10.1371/journal.pcbi.1014053.t001

Notably, while maintaining high Sensitivity (0.9016), our method also attains competitive Specificity (0.8525), achieving an excellent balance between these metrics. In contrast, existing methods commonly show performance imbalance issues.

For the AntiCP 2.0_Main and AntiCP 2.0_Alternate datasets, we first train models on the complete training sets and evaluate them on the independent test sets, consistent with other comparative methods. As shown in Table 2, our Multi-ACPNet achieves the highest ACC (0.8140 and 0.9536) and MCC (0.6283 and 0.9103) values on both datasets. And, on the AntiCP 2.0_Alternate dataset, our predictor attains a remarkable SP of 0.9948, outperforming all existing methods. To preclude potential performance overestimation caused by redundant sequences, we remove training data with >90% sequence similarity to the test set and evaluate model performance under this condition (this version is denoted as Multi-ACPNet-CD). Experimental results demonstrate that even after removing redundant data, our model still exhibits superior performance compared to existing methods that do not perform redundancy elimination. On the AntiCP 2.0_Main dataset, most existing methods exhibit either high SE with low SP, or conversely low SE with high SP. For instance, AntiCP achieves perfect SE (1) but extremely low SP (0.0116), reflecting complete positive sample detection at the cost of an excessively high false positive rate. Conversely, ACPred-LAF shows the highest SP (0.8895) but relatively low SE (0.6337), suggesting accurate negative sample exclusion but compromised detection rate for true positives. However, our approach maintains a good balance (SE = 0.8214, SP = 0.8021) while achieving competitive rankings, substantially reducing misclassification risks in practical applications.

Multi-ACPNet achieves a false positive rate (FPR) of 19.79% on AntiCP 2.0_Main (where negative samples are AMPs) and a Specificity of 0.9948 (FPR as low as 0.52%) on AntiCP 2.0_Alternate (where negative samples are random peptides). This indicates that compared to structurally random peptides, the model finds it more difficult to distinguish AMPs from ACPs. However, the two models are independently trained on different data distributions, which could itself contribute to performance differences. Therefore, we conduct an additional experiment: we merge the training sets of both datasets to train a unified model and evaluate it separately on their respective independent test sets. This approach ensures that the model is exposed to both negative sample types during training, providing a more controlled test of the hypothesis that distinguishing ACPs from AMPs is inherently more difficult. The results (Fig B in S1 Text) show that the model still attains a high FPR of 19.77% on AntiCP 2.0_Main, while achieving only 1.55% on AntiCP 2.0_Alternate. These experiments suggest that ACPs and AMPs may possess similar structural characteristics. Consequently, employing AMPs as negative

Table 2. Performance comparison between the proposed method and existing methods on the AntiCP 2.0 independent test dataset[a].

		AntiCP 2.0_Main				AntiCP 2.0_ Alternate			
Methods	**Year**	**ACC**	**SE**	**SP**	**MCC**	**ACC**	**SE**	**SP**	**MCC**
AntiCP 2.0 [8]	2021	0.7543	0.7764	0.7341	0.5100	0.9201	0.9227	0.9175	0.8400
AntiCP [7]	2017	0.5058	**1.0000**	0.0116	0.0700	0.8995	0.8969	0.9020	0.8000
ACPred [42]	2018	0.5347	0.8555	0.2139	0.0900	0.8531	0.8711	0.8351	0.7100
ACPred-FL [25]	2018	0.4480	0.6705	0.2254	0.1200	0.4380	0.6021	0.2558	0.1500
ACPred-LAF [21]	2021	0.7616	0.6337	**0.8895**	0.5413	0.8918	0.8660	0.9175	0.7845
ACPred-Fuse [26]	2020	0.6890	0.6920	0.6860	0.3800	0.7890	0.6440	0.9330	0.6000
iACP-DRLF [13]	2021	0.7750	0.8070	0.7430	0.5500	0.9300	0.8960	0.9640	0.8600
ME-ACP [22]	2022	0.7920	0.7490	0.8350	0.5860	0.9330	0.9170	0.9480	0.8660
GRCI-Net [43]	2021	0.7460	0.7540	0.8350	0.5860	0.8760	0.8700	0.8810	0.7510
AI4ACP [[17]	2022	0.7180	0.8020	0.6330	0.4420	0.8940	0.8710	0.9180	0.7900
ACPred-BMF [20]	2022	0.8081	0.8837	0.7326	0.6200	0.9355	0.9227	0.9485	0.8700
ACP-check [44]	2022	0.7800	0.8000	0.7700	0.5600	0.9300	**0.9300**	0.9300	0.8600
ACP-OPE [45]	2023	0.7895	0.8153	0.7676	/	/	/	/	/
iACP-RF [46]	2023	0.7590	0.7560	0.7620	0.5200	0.9310	0.8920	0.9690	0.8600
CAPTURE [6]	2024	0.7670	/	/	0.5400	0.9410	/	/	0.8800
Contrastive [47]	2024	0.8081	0.8023	0.8140	0.6163	0.9381	0.8969	0.9794	0.8793
MA-PEP [40]	2024	0.8081	0.8779	0.7384	0.6224	0.9356	0.9227	0.9485	0.8714
ACP-PDAFF [41]	2024	0.8000	0.8300	0.7600	0.6000	0.9400	0.9200	0.9600	0.8800
Multi-ACPNet-CD(ours)	2025	0.8110	0.8314	0.7907	0.6226	0.9510	0.9124	0.9897	0.9048
Multi-ACPNet (ours)	2025	**0.8140**	0.8214	0.8021	**0.6283**	**0.9536**	0.9124	**0.9948**	**0.9103**

[a]All the results reported in this table except for the ones of Multi-ACPNet and Multi-ACPNet-CD are obtained from publications./ means that there is no value in the corresponding item, bold indicates the highest value.

https://doi.org/10.1371/journal.pcbi.1014053.t002

samples substantially elevates the classification challenge, which likely accounts for the notably inferior performance of Multi-ACPNet on AntiCP 2.0_Main compared to AntiCP 2.0_Alternate.

A common issue in the ACP prediction literature is the frequent neglect of model scale signatures, such as the number of parameters, computational cost, and inference time. However, this evaluation is standard practice in the broader machine learning literature to ensure fair and insightful comparisons [48]. To comprehensively evaluate model efficiency, we compare the model scale and inference performance of all deep learning methods listed in Tables 1 and 2. Under unified hardware conditions, we systematically measure the number of parameters (Params), floating-point operations per sample (FLOPs), and inference time per sample for each model. As shown in Table E in S1 Text, although Multi-ACPNet exhibits a higher number of parameters and greater computational cost per sample than all compared deep learning models, its inference time remains competitive without significant delay. In contrast, the inference time for a single sample in models such as AI4ACP, GRCI-Net, and ACPred-BMF exceeds one second, indicating substantially lower efficiency. Furthermore, we evaluate the batch inference efficiency of the model by processing the entire test set as a single batch. The total processing time for the ACP-Mixed-80 dataset (122 samples) is only 35.68 ms, while the AntiCP 2.0_Main (344 samples) and AntiCP 2.0_Alternate (388 samples) datasets require only 110.50 ms and 168.80 ms, respectively. These results demonstrate that Multi-ACPNet possesses strong real-time processing capability. To rule out the possibility that performance improvements stem merely from increased model scale, we conduct in-depth ablation experiments in Section 3.5. The results indicate that the performance advantage primarily arises from the deep integration of sequence and structural information, rather than from simply increasing the number of parameters.

The comparative results demonstrate that our method maintains reasonable inference overhead and achieves an improved balance between Sensitivity and Specificity across all three datasets while attaining optimal performance in other metrics. This is primarily attributed to the introduction of graph convolution in the proposed model, which captures peptide structural features, overcoming the limitation of previous methods that relied solely on sequence information. Our network employs BiLSTM and causal convolution to extract multi-scale sequential patterns (preventing missed detections) while utilizing distance maps to identify spatial interactions (reducing false positives). The synergy between sequence and structural features significantly enhances the recognition capability for complex peptides.

3.4. Performance of functional activity prediction of ACP against various cancers

In the functional activity prediction task, we compare our model with two benchmark methods: the current state-of-the-art DUO-ACP [49] and ACP-MLC. For a fair comparison, we use the same strategy of 10-fold cross-validation as employed in the papers.

Table 3 presents the comparison of three models on seven evaluation metrics, including Macro-averaging scores and Hamming loss. The results reveal that our Multi-ACPNet achieves statistically superior performance on all metrics. Specifically, it achieves an ACC of 0.8697 and MCC of 0.7309, significantly outperforming the other two models. Notably, it attains a Hamming loss of only 0.1303, indicating minimal discrepancy between predicted and ground truth labels. This marked improvement in predictive performance is achieved with only a marginal increase in inference time, underscoring the efficiency of our architecture. Furthermore, in order to deeply analyze the model's specific performance for each cancer category, we comprehensively evaluate the predictive performance across seven tumor cell categories (Table F in S1 Text). Multi-ACPNet achieves excellent prediction ability on all cancer categories. However, it shows relatively lower Sensitivity in the Blood category, primarily due to the limited samples that hindered effective feature learning. To demonstrate our model's superior performance, we similarly evaluate all three competing models across cancer types. It should be noted that only four metrics (ACC, AUC, F1-score, and MCC) were reported for DUO-ACP in the original publication. Here we also show the performance of the DUO-ACP model on only four metrics. As illustrated in Fig 3, ACP-MLC shows competitive performance in Colon cancer but shows limited performance in other tissue types. DUO-ACP generally achieves higher scores across its four reported metrics, yet its overall performance remains inferior to our Multi-ACPNet, which attains optimal results for most evaluation metrics across all cancer types.

For Micro-averaging, Multi-ACPNet achieves an ACC of 0.8697, Precision of 0.8763, F1-score of 0.8472, and SE of 0.8371, outperforming DUO-ACP by 1.52% in F1-score and 1.31% in SE. These results demonstrate that our proposed model can more reliably handle the complexity of multi-label tasks and exhibits superior performance in ACP activity prediction. This advantage is mainly due to our innovative design of a multi-scale sequence-structure cooperative learning framework, which simultaneously captures short-chain functional motifs and long-range sequence patterns at the sequence level, while resolving multi-scale spatial structures of peptide chains at the structural level. This integrated approach enables the model to more effectively capture discriminative features across different cancer types and specifically recognize the anticancer activity of peptides against various cancer cell lines, thereby maintaining optimal performance under a comprehensive evaluation system.

Table 3. Performance comparison between the proposed method and existing methods for ACP functional activity prediction[a].

Method	Params(M)/ FLOPs(G)/ Inf-Time(ms)	AUC	ACC	SE	SP	F1-score	MCC	Hamming loss
ACP-MLC [30]	/	0.8680	0.7730	0.6080	0.8540	0.6760	0.5090	0.1570
DUO-ACP [49]	3.8512/0.0358/2.9390	0.8860	0.8350	0.8120	0.8190	0.8190	0.6470	0.1650
Multi-ACPNet (ours)	10.6752/0.5344/8.8156	**0.9033**	**0.8697**	**0.8371**	**0.8816**	**0.8472**	**0.7309**	**0.1303**

[a]All the results reported in this table except for the ones of Multi-ACPNet are obtained from publications. Bold indicates the highest value.

https://doi.org/10.1371/journal.pcbi.1014053.t003

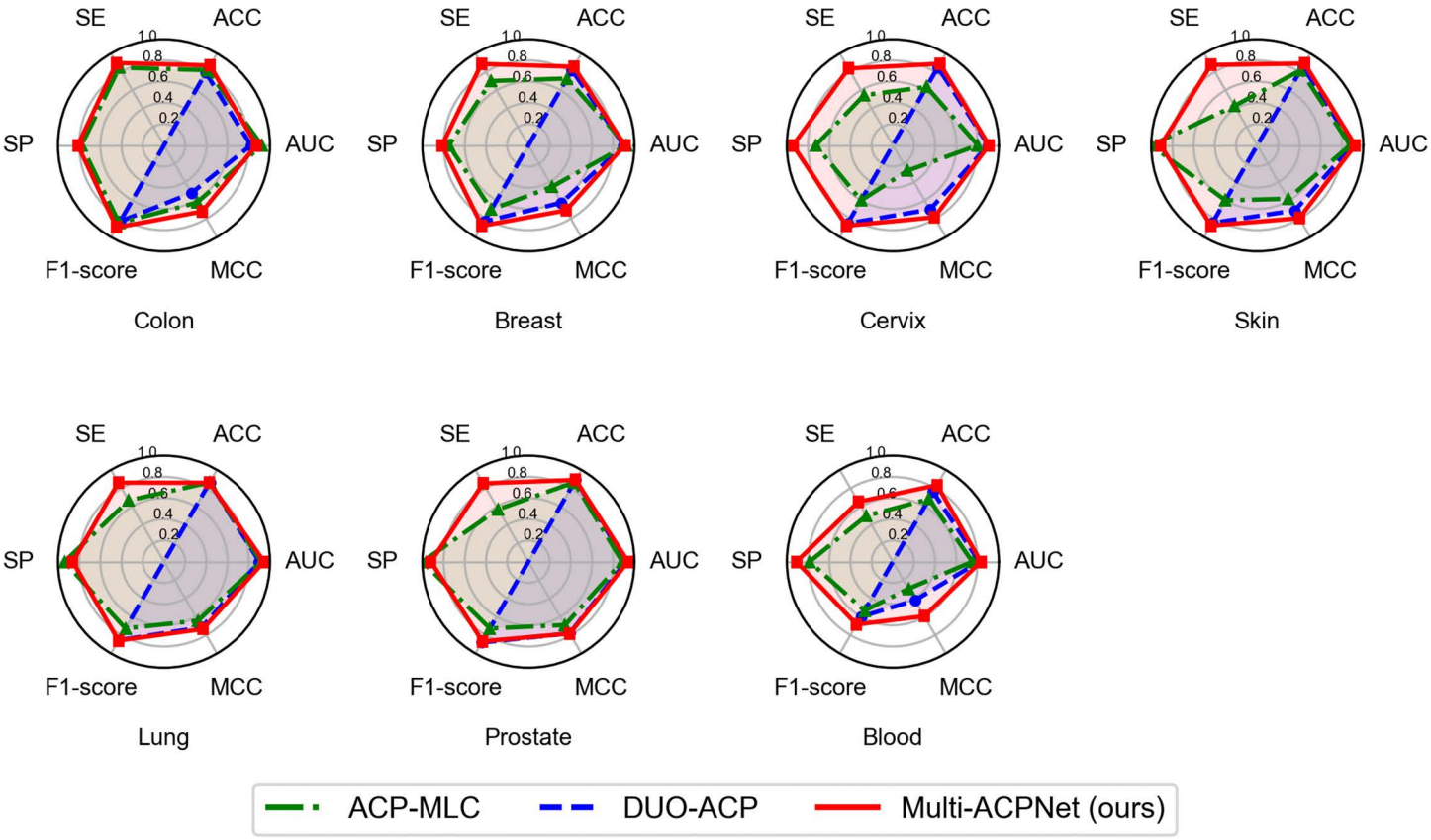

Fig 3. Performance comparison between the proposed method and existing methods across seven cancer types.

https://doi.org/10.1371/journal.pcbi.1014053.g003

3.5. Effectiveness analysis of sequence and structure learning

To analyze the synergistic mechanism between sequence and structural information, validate the necessity of sequence learning, and demonstrate the effectiveness of structural information, we design two ablated models while maintaining identical parameter counts: (1) ACPNet_no_Graph, which, following a similar approach to Ortega et al. [50], retains only sequence learning by setting all elements in the adjacency matrix to 1 and all distance values to 1, thereby disabling graph structural learning, and (2) ACPNet_no_Seq, which disables sequence information by setting all input ESM C features to 1. Comprehensive evaluations on the AntiCP 2.0_Main dataset (Fig 4) reveal that ACPNet_no_Graph outperforms the structure-only baseline (ACPNet_no_Seq) across all evaluation metrics. Specifically, it achieves a significantly higher MCC (0.5988). Ultimately, Multi-ACPNet attains optimal performance by integrating both sequence and structural features.

To better demonstrate the effectiveness of sequence-structure synergy and enhance model interpretability, we employ t-distributed stochastic neighbor embedding (t-SNE) [51] for feature visualization. Specifically, we extract features from the penultimate fully connected layer of models and reduce them into a 2D space. Fig 5 presents the visualization results on the AntiCP 2.0_Main independent test set. ACPNet_no_Seq exhibits limited clustering capability, showing satisfactory aggregation only for partial negative samples. ACPNet_no_Graph produces fewer clustering errors. Multi-ACPNet, through sequence-structure fusion, generates two distinct clusters for ACPs and non-ACPs, confirming its superior capability in learning both common and discriminative features. Notably, there remains a substantial number of challenging samples that are difficult to classify correctly, improving the model's ability to discriminate these difficult samples will

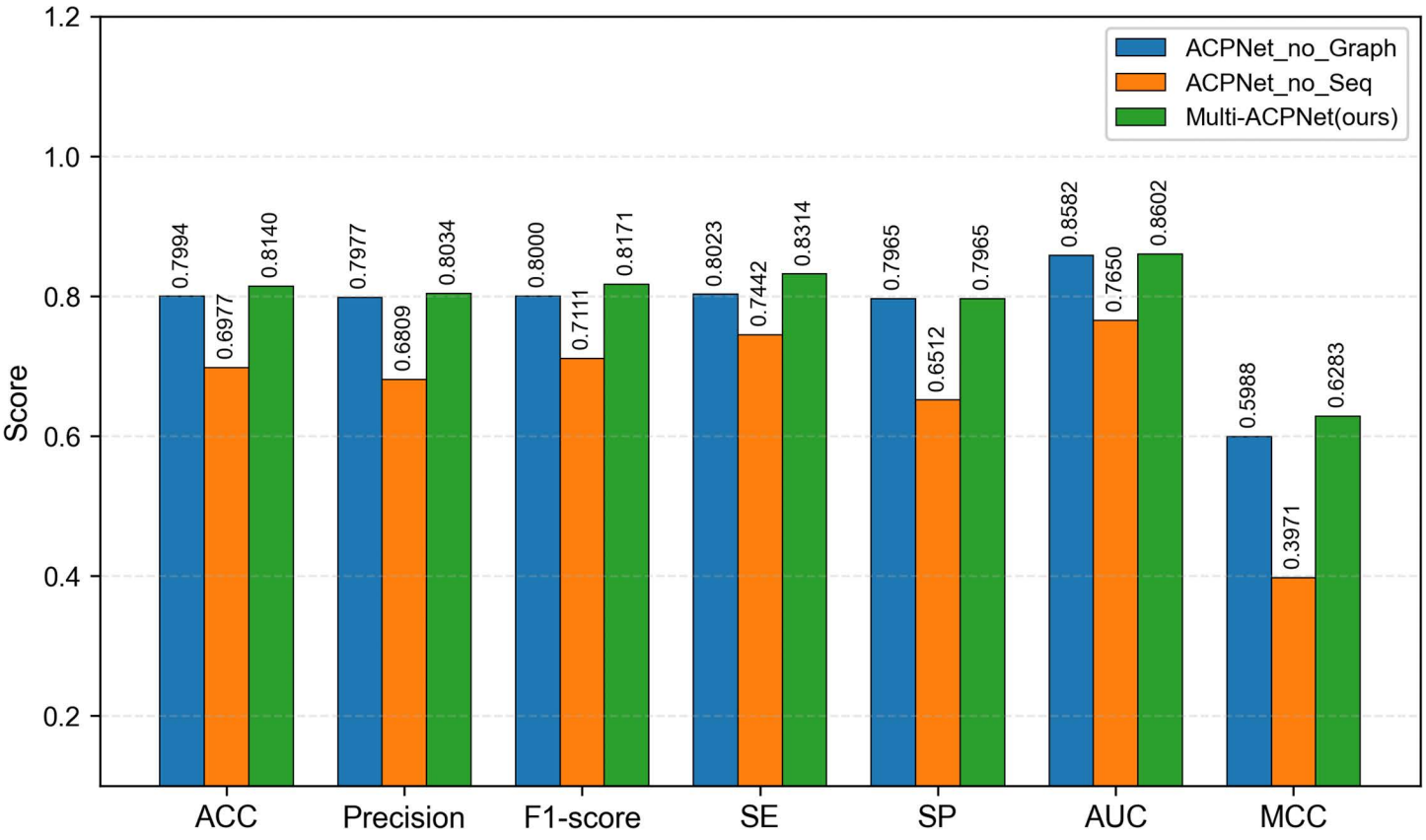

Fig 4. Performance comparison of sequence-only and structure-only models on the AntiCP 2.0_Main test set.

https://doi.org/10.1371/journal.pcbi.1014053.g004

be the focus of our future research. Furthermore, Figs C to F in S1 Text present the visualization results on the AntiCP 2.0_Alternate and ACP-Mixed-80 datasets. Consistent with the observations on the AntiCP 2.0_Main dataset, ACPNet_no_Graph achieves superior classification performance compared to ACPNet_no_Seq in the ACP identification task. However, as shown in Fig G in S1 Text, a different pattern emerges on the ACPs functional prediction dataset (using Macro-averaging metrics). For this multi-label functional classification task, ACPNet_no_Seq outperforms ACPNet_no_Graph on most metrics, suggesting that structural information may play a more dominant role in distinguishing specific cancer-type activities. Nonetheless, the full Multi-ACPNet model achieves the most optimal classification results across all test sets, further validating the effectiveness and adaptability of our sequence-structure fusion approach for both binary ACP identification and multi-label functional prediction tasks.

The above phenomenon indicates that in our proposed model, the Sequence Multi-Scale Network dominates the identification process of ACPs. On the other hand, the Graph Multi-Scale Network further captures residue interactions across different spatial scales, elucidating critical tertiary structural features of peptides, thereby playing a pivotal role in enhancing model performance. The sequence and structural modalities exhibit complementary advantages through synergistic cooperation, collectively establishing a comprehensive mechanism for ACP recognition. These findings demonstrate the macro-effectiveness of the synergistic mechanism between sequence and structural modalities. To further validate the contributions of key internal modules, we conduct ablation studies targeting critical components within the sequence and graph networks. The specific configurations of the ablated models are as follows:

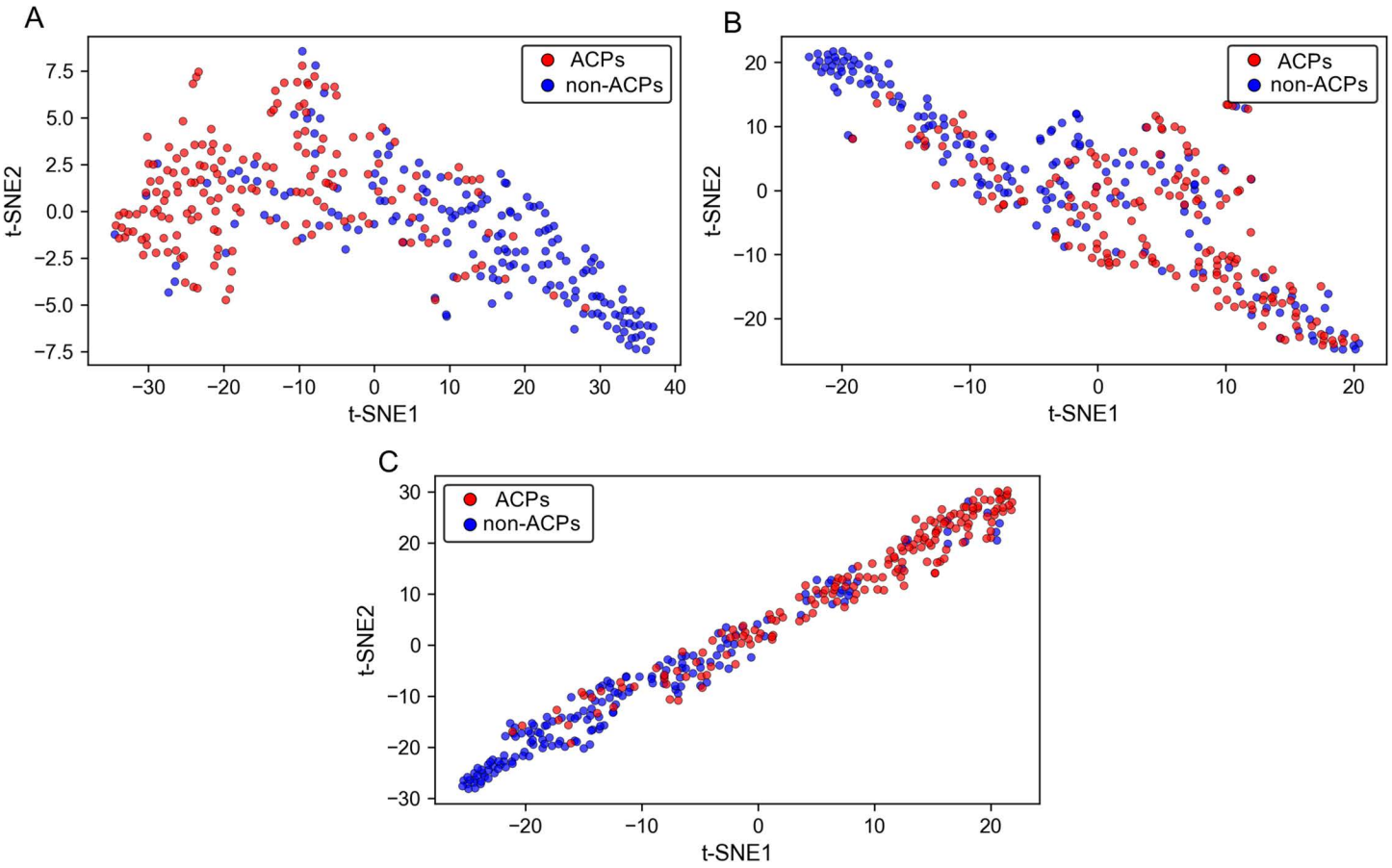

Fig 5. t-SNE visualization of feature representations from the penultimate layer on AntiCP 2.0_Main test set. (A) ACPNet_no_Graph feature distribution. (B) ACPNet_no_Seq feature distribution. (C) Multi-ACPNet feature distribution.

https://doi.org/10.1371/journal.pcbi.1014053.g005

(1) No-BiLSTM: The BiLSTM layer is removed from the Sequence Multi-Scale Network, retaining only the causal convolutional module for sequence feature extraction.

(2) No-CausalCNN: The causal convolutional module is removed, with sequence features extracted solely via BiLSTM.

(3) No-TopKPooling: The top-k pooling layer and the subsequent subgraph convolution module are removed from the Graph Multi-Scale Network. Global pooling is applied directly to the full-graph features instead, thereby eliminating the key-residue selection and hierarchical subgraph learning mechanism.

(4) No-Seq: The entire sequence branch (including both BiLSTM and causal convolution) is removed, and the model relies solely on the multi-scale GNN for prediction.

(5) No-Graph: The graph structure learning branch is removed. The output of the Sequence Multi-Scale Network is fed directly into the classifier, relying solely on sequence information.

The experimental results are presented in Tables G to J in S1 Text. The dominant role of sequence information is clearly evidenced by the dramatic performance drop observed when the entire sequence branch is removed (No-Seq). For instance, on the AntiCP 2.0_Main and ACP functional prediction datasets, the MCC scores of No-Seq decrease by

0.3024 and 0.2327, respectively, compared to the full model. Within the sequence branch, both long-range dependencies and local patterns are essential. On the functional prediction dataset, removing BiLSTM (No-BiLSTM) causes a more severe performance degradation than removing causal convolution (No-CausalCNN), underscoring the greater importance of global context modeling for this task. Conversely, on the three ACP identification datasets, removing causal convolution results in a more significant performance degradation, highlighting the critical role of local motif extraction in distinguishing ACPs from non-ACPs. In the graph branch, removing top-k pooling (No-TopKPooling) consistently reduces performance across all datasets, validating the importance of key-residue selection and hierarchical subgraph learning.

Critically, performance is not merely a function of model scale. For the first ablation study, where we create ACPNet_no_Seq and ACPNet_no_Graph by holding the number of parameters constant and manipulating the input features to disable sequence or structural learning, respectively, the model performance changes substantially. This confirms that the observed effectiveness depends on the integration of both information modalities, not on scale alone. In the second ablation study involving the removal of key components, no consistent pattern emerges where larger models yield better performance. Comparisons with external algorithms further support this conclusion. For instance, on the ACP-Mixed-80 dataset, the lightweight No-Seq variant (0.44M parameters) outperforms much larger models such as ACPred-LAF (2.46M). Similarly, the No-BiLSTM model (4.11M) achieves better results than the larger MA-PEP (6.41M) and ACP-PDAFF (7.88M). On the AntiCP 2.0_Alternate dataset, No-Seq surpasses ACPred-LAF by 0.0475 in MCC despite having only 17.76% of its parameters. These comparisons clearly indicate that the superiority of Multi-ACPNet stems not from increased model size, but from the effective integration of multi-scale sequence and structural feature extraction.

3.6. Analysis of feature encoding methods

In this section, we conduct an ablation study to evaluate three feature combination schemes (progressively incorporating BPF, position, and AAindex features) across all datasets. Tables K and L in S1 Text present the experimental results. The results demonstrate that starting from BPF, the progressive integration of position and AAindex features not only introduces precise residue position information but also provides a biochemical foundation for structure-function relationships. This integration results in a significant and stable improvement in the model's predictive performance across both tasks, validating the efficacy of the multi-feature fusion strategy.

To refine the high-dimensional AAindex representation, we apply the mRMR feature selection algorithm. As shown in Tables M and N in S1 Text which present the performance across datasets at different feature dimensionalities, the model achieves its highest ACC and MCC values on the AntiCP 2.0_Main, ACP-Mixed-80, and ACP function prediction datasets when the AAindex features are reduced to 60 dimensions. Furthermore, the highest SP value is also attained on the ACP-Mixed-80 dataset, while performance on the AntiCP 2.0_Alternate dataset remains sufficiently high. Beyond 60 dimensions, further increase in feature number yields no substantial gains and may even lead to performance degradation. Therefore, by selecting 60-dimensional features, we effectively reduce computational cost while retaining the physicochemical properties most relevant to peptide prediction, thereby avoiding the redundancy and interference associated with high-dimensional representations.

In recent years, a wide variety of pre-trained language models (PLMs) have rapidly emerged. In order to demonstrate the superiority of ESMC_600M, we conduct a systematic comparison with several mainstream PLMs, including ESM2, ESM3, ProtT5 [52], and ProteinBERT [53] (detailed in Tables O and P in S1 Text). The experimental results demonstrate that the ESMC_600M model achieves the best performance across multiple key metrics on three ACP classification datasets and attains leading results on all evaluation metrics for the ACP function prediction dataset. Considering its balanced computational efficiency and comprehensive performance superiority, we ultimately select ESMC_600M as our foundational feature extraction pre-trained model.

3.7. Parameter stability analysis

To evaluate the robustness of the Multi-ACPNet model and its sensitivity to key hyperparameters, we analyze the impact of four core hyperparameters: learning rate, distance threshold (D_{th}), dropout rate, and TopKPooling ratio. Detailed performance metrics for each parameter across different datasets are provided in Tables Q to X in S1 Text, while Fig H in S1 Text visually illustrates the trend of AUC as the parameters vary. To further quantify the model's sensitivity to these parameters, we calculate the Coefficient of Variation (CV) of the AUC within the tested ranges for each parameter. The CV is defined as the ratio of the standard deviation to the mean, where a lower value indicates that model performance is less sensitive to fluctuations in that parameter and thus reflects stronger robustness.

The comprehensive analysis shows that Multi-ACPNet maintains relatively stable performance across a wide range of most hyperparameters, though its sensitivity varies significantly depending on the parameter and the dataset. The AntiCP 2.0_Alternate dataset consistently exhibits the highest AUC and the smallest CV values across all four parameter variations, indicating that the prediction task for this dataset is relatively simpler, and the model's performance is least sensitive to parameter adjustments. In contrast, the ACP-Mixed-80 dataset shows the largest CV values when D_{th}, dropout rate, and TopKPooling ratio are varied, suggesting that the prediction performance for this dataset is more sensitive to adjustments in structure construction and regularization strategies. We also observe that the ACPs functional prediction dataset exhibits substantial fluctuations across different learning rates, with a CV as high as 5.35%. Notably, when the learning rate is set to 1e-5, performance deteriorates drastically, with the MCC dropping to as low as 0.4181. This highlights the critical importance of selecting an appropriate learning rate for ensuring stable and reliable model performance.

3.8. Interpretability analysis

Beyond the t-SNE-based feature visualization in Section 3.5, we further employ agglomerative clustering [54] on the learned feature representations to investigate the decision-making mechanism of Multi-ACPNet and enhance the model's interpretability. We extract 256-dimensional feature vectors from the penultimate layer of the model and select 20 positive ACPs and 20 non-ACPs from the ACP-Mixed-80 independent test set, resulting in a total of 40 samples for cluster analysis. As shown in Fig 6, the clustering results of the 40 samples in the 256-dimensional feature space demonstrate a clear separation between the two classes. The ACPs (bottom) and non-ACPs (top) form two distinct clusters with a noticeable boundary between them, indicating that the features learned by the model effectively differentiate the two types of peptides. In particular, the ACPs exhibit a more consistent distribution in the feature space, reflecting shared physicochemical properties that are likely critical determinants of their anticancer function. In contrast, the non-ACPs show greater diversity in their feature representation, which aligns with the inherent heterogeneity of this category—encompassing

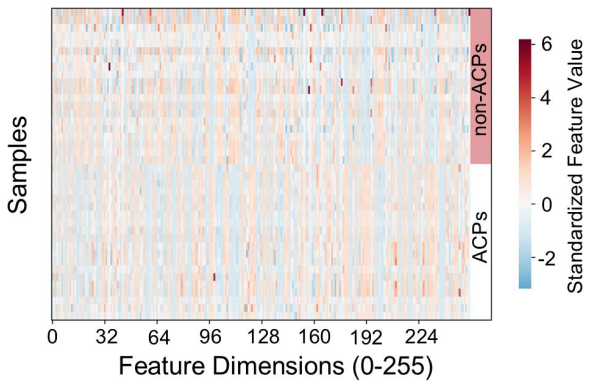

Fig 6. Agglomerative clustering of feature vectors from the penultimate network layer on ACP-Mixed-80 test set.

https://doi.org/10.1371/journal.pcbi.1014053.g006

various functional peptides such as antimicrobial peptides, signaling peptides, and random peptides without anticancer activity. Although the high-dimensional features learned by the model are abstract in nature, the distinct clustering patterns strongly suggest that they capture systematic biological differences. These salient feature dimensions not only enhance the credibility of the classification decisions but also provide important clues for subsequent biological interpretation: they may correspond to functional motifs or structural domains that have not been fully characterized, offering guidance for the rational design and functional discovery of novel ACPs.

Based on the residue importance scores derived from TopKPooling, we select 50 correctly predicted ACP sequences from the ACP-Mixed-80 test set to visualize positional residue importance. In the corresponding heatmap (Fig IA in S1 Text), a redder color indicates a greater contribution of the residue at that position to the correct prediction of ACPs. The results clearly demonstrate that residues in the *N*-terminal region, especially the first amino acid position, contribute most significantly. This finding further confirms the critical influence of the *N*-terminal sequence on the anticancer activity of peptides, which is consistent with previous research [15].

Fig IB in S1 Text further presents the sequence logos of the first ten amino acid positions at the *N*-terminus for the positive samples in the ACP-Mixed-80 test set. Statistical analysis reveals that Glycine (G) occurs most frequently at the first position, while Phenylalanine (F), Leucine (L), and Lysine (K) are significantly enriched within the first ten positions. The crucial functional roles of these amino acids in ACPs have been well-established in prior studies [55,56]. For instance, cationic residues (e.g., K) can penetrate and disrupt cancer cell membranes, thereby inducing cytotoxicity. Concurrently, hydrophobic residues (e.g., F) enhance anticancer efficacy by destabilizing membrane integrity. Importantly, our model successfully captures and prioritizes these functionally critical residues, demonstrating its biological interpretability and alignment with known mechanisms of action.

3.9. Extended application: toxicity prediction

From a translational perspective, the therapeutic potential of ACPs depends not only on their anticancer efficacy but equally on their safety profiles—peptides cytotoxic to normal cells have limited clinical applicability. To comprehensively evaluate peptide candidates, we apply Multi-ACPNet to toxicity prediction, extending its application scope to address this critical safety dimension in peptide therapeutic development.

In the absence of publicly available datasets specifically for ACP toxicity, we construct two dedicated datasets to comprehensively evaluate model performance: a general peptide toxicity dataset S1 for benchmark validation, and an ACP toxicity test set S2 for targeted evaluation. First, to establish the initial data pool, we retrieve 1,783 reviewed toxic peptide sequences from the UniProt [57] database using the search query “KW-0800 AND (length: [10 TO 50])”, and obtain 10,493 non-toxic peptide sequences using the query “NOT KW-0800 AND NOT KW-0020 AND (length: [10 TO 50])”. To specifically assess the model’s ability to predict the toxicity of ACPs, we further screen sequences that overlap with the above-mentioned data pool from the four ACP datasets employed in this study: ACP-Mixed-80, AntiCP 2.0_Main, AntiCP 2.0_Alternate, and the ACP functional activity prediction dataset. From this overlap, we construct an independent test set S2, which contains 41 non-toxic ACPs and, due to limited data availability, only 14 toxic ACPs. After excluding the sequences included in S2 from data pool, we reduce redundancy in the remaining peptides using CD-HIT with a 90% threshold, producing 1,202 unique toxic peptide sequences and 5,811 non-toxic sequences. To ensure class balance, we randomly select an equal number of non-toxic peptides. We then randomly select approximately 15% of peptides from both the toxic and non-toxic subsets to form the test set, yielding the general peptide toxicity dataset S1 with 2,050 training and 354 test samples.

We first evaluate the model’s performance on the general peptide toxicity dataset S1 and compare it with the recently proposed ToxGIN [58] algorithm, a Graph Isomorphism Network (GIN)-based peptide toxicity predictor. We uniformly employ cross-validation on the training set to determine the model weights and perform the final evaluation on the test set. As shown in Table Y in S1 Text, our model outperforms ToxGIN across all five evaluation metrics with significantly fewer parameters, achieving a Specificity of 0.9096 and a MCC of 0.7748.

We then perform independent external validation of the two trained general models on S2, specifically to examine their generalization capability and practical potential in real-world ACPs toxicity prediction tasks. As shown in Fig 7, ToxGIN demonstrates a slightly higher recognition rate for non-toxic ACPs. However, it exhibits a substantially higher error rate on toxic samples, misclassifying six out of the 14 toxic ACPs as non-toxic. In contrast, our model shows a more balanced and robust performance: it correctly identifies the majority of toxic peptides, while maintaining a comparable recognition rate for non-toxic samples. These results demonstrate that, although our model is not specifically designed for toxicity prediction, it still yields competitive performance in predicting the toxicity of ACPs, highlighting its potential utility in screening for non-toxic ACP candidates.

However, the ACP toxicity dataset S2 is limited in size, which to some extent affects the statistical power of model performance evaluation on this subset. Future research will focus on collecting more experimentally validated toxic and non-toxic ACP data to build a more representative benchmark dataset.

4. Limitations

The proposed method possesses clear practical value and significance. From an application perspective, our proposed framework demonstrates robust performance across three critical tasks in peptide drug discovery: anticancer peptide (ACP) identification, functional activity prediction, and toxicity assessment. This capability allows it to serve as an efficient computational screening tool that can accelerate the discovery of novel anticancer peptides by evaluating their efficacy, mechanism, and safety. Methodologically, this work systematically validates the complementarity and synergistic effects between sequence and structural information in peptide prediction, providing a transferable framework for similar tasks. A key future direction is to develop an integrated multi-task model capable of simultaneously identifying ACPs, assessing their toxicity, and predicting their functional activity. Leveraging the powerful multi-scale feature extraction capabilities of Multi-ACPNet and by incorporating a multi-task learning architecture, it is expected to optimize three closely related prediction objectives in parallel based on shared sequence-structure representations. This approach holds promise for providing a powerful computational tool to enable efficient and rational design as well as virtual screening of peptide-based drugs.

Despite these promising aspects and future directions, it is important to acknowledge several current limitations of the study. First, the model's performance is constrained by the limited availability of high-quality tertiary structural features of peptides. Although we employ advanced tools such as trRosetta for structure prediction, discrepancies inevitably exist between predicted and real experimental structures, which may introduce potential biases during model training

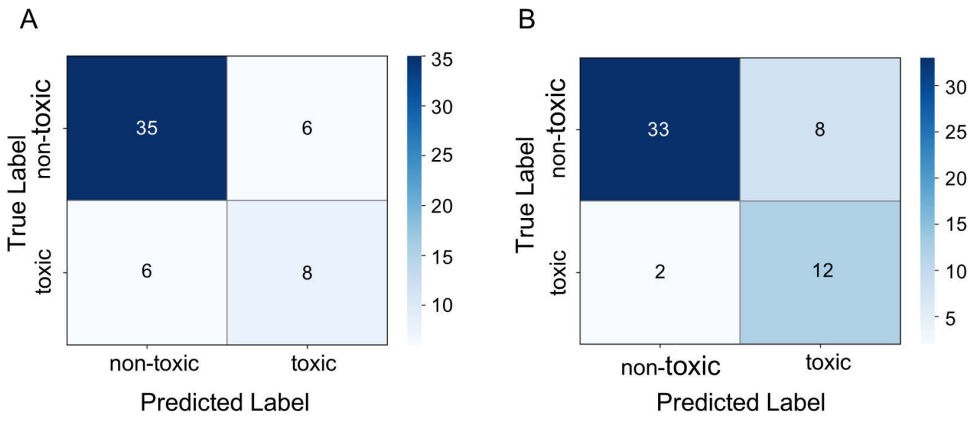

Fig 7. Comparative performance on the ACP toxicity test set ((s2). (A) Confusion matrix for ToxGIN. (B) Confusion matrix for Multi-ACPNet.

https://doi.org/10.1371/journal.pcbi.1014053.g007

and generalization. Future work could explore more robust structure representation learning methods. Furthermore, the current model framework is primarily validated on short peptide sequences. For ultra-long peptides or small proteins, their more complex long-range interactions and folding patterns may require specialized architectural adjustments. Extending the model to handle a broader range of peptide and protein lengths will be a key direction for improving its general applicability. Additionally, the current model architecture is relatively complex. Future work will focus on further optimizing the model architecture to reduce its complexity and parameter size while maintaining predictive performance, thereby shortening inference time and enhancing deployment efficiency. Finally, the imbalanced data distribution across cancer types potentially affects prediction performance for minority classes. As the current model is trained primarily on public datasets and only predicts seven anticancer activity types, future work will focus on enhancing the model's generalizability and expanding its predictive coverage through integration of more diverse datasets.

5. Conclusion

This study proposes Multi-ACPNet, an innovative dual-function framework capable of simultaneous ACP identification and activity type classification, addressing the limitation of existing methods that rely solely on sequence information. The model achieves multi-level information extraction from sequence features to spatial structural dependencies through deep integration of domain prior knowledge with ESM-2 embeddings, and an original hybrid architecture combining BiLSTM-causal convolution with multi-scale GCN. Experimental results demonstrate that this novel approach significantly improves prediction performance for both tasks, representing a critical advancement for biomedical applications.

Supporting information

S1 Text. Supplementary Figures and Tables.
(DOCX)

Author contributions

Funding acquisition: Lu Meng.

Investigation: Lu Meng.

Methodology: Lu Meng, Lijun Zhou.

Project administration: Lu Meng.

Resources: Lu Meng.

Software: Lijun Zhou.

Validation: Lijun Zhou.

Visualization: Lijun Zhou.

Writing – original draft: Lijun Zhou.

Writing – review & editing: Lu Meng.

References

1. Otvos L Jr. Peptide-based drug design: here and now. Methods Mol Biol. 2008;494:1–8. https://doi.org/10.1007/978-1-59745-419-3_1 PMID: 18726565
2. Vlieghe P, Lisowski V, Martinez J, Khrestchatisky M. Synthetic therapeutic peptides: science and market. Drug Discov Today. 2010;15(1–2):40–56. https://doi.org/10.1016/j.drudis.2009.10.009 PMID: 19879957
3. Sun X, Liu Y, Ma T, Zhu N, Lao X, Zheng H. DCTPep, the data of cancer therapy peptides. Sci Data. 2024;11(1):541. https://doi.org/10.1038/s41597-024-03388-9 PMID: 38796630

4. Basith S, Manavalan B, Hwan Shin T, Lee G. Machine intelligence in peptide therapeutics: A next-generation tool for rapid disease screening. Med Res Rev. 2020;40(4):1276–314. https://doi.org/10.1002/med.21658 PMID: 31922268
5. Liang X, Li F, Chen J, Li J, Wu H, Li S, et al. Large-scale comparative review and assessment of computational methods for anti-cancer peptide identification. Brief Bioinform. 2021;22(4):bbaa312. https://doi.org/10.1093/bib/bbaa312 PMID: 33316035
6. Ghafoor H, Asim MN, Ibrahim MA, Ahmed S, Dengel A. CAPTURE: Comprehensive anti-cancer peptide predictor with a unique amino acid sequence encoder. Comput Biol Med. 2024;176:108538. https://doi.org/10.1016/j.compbiomed.2024.108538 PMID: 38759585
7. Tyagi A, Kapoor P, Kumar R, Chaudhary K, Gautam A, Raghava GPS. In silico models for designing and discovering novel anticancer peptides. Sci Rep. 2013;3:2984. https://doi.org/10.1038/srep02984 PMID: 24136089
8. Agrawal P, Bhagat D, Mahalwal M, Sharma N, Raghava GPS. AntiCP 2.0: An updated model for predicting anticancer peptides. Cold Spring Harbor Laboratory. 2020;3.
9. Karim T, Shaon MSH, Sultan MF, Hasan MZ, Kafy A-A. ANNprob-ACPs: A novel anticancer peptide identifier based on probabilistic feature fusion approach. Comput Biol Med. 2024;169:107915. https://doi.org/10.1016/j.compbiomed.2023.107915 PMID: 38171261
10. Yi H-C, You Z-H, Zhou X, Cheng L, Li X, Jiang T-H, et al. ACP-DL: A Deep Learning Long Short-Term Memory Model to Predict Anticancer Peptides Using High-Efficiency Feature Representation. Mol Ther Nucleic Acids. 2019;17:1–9. https://doi.org/10.1016/j.omtn.2019.04.025 PMID: 31173946
11. Liu J, Li M, Chen X. AntiMF: A deep learning framework for predicting anticancer peptides based on multi-view feature extraction. Methods. 2022;207:38–43. https://doi.org/10.1016/j.ymeth.2022.07.017 PMID: 36100141
12. Zhang S, Zhao Y, Liang Y. AACFlow: an end-to-end model based on attention augmented convolutional neural network and flow-attention mechanism for identification of anticancer peptides. Bioinformatics. 2024;40(3):btae142. https://doi.org/10.1093/bioinformatics/btae142 PMID: 38452348
13. Lv Z, Cui F, Zou Q, Zhang L, Xu L. Anticancer peptides prediction with deep representation learning features. Brief Bioinform. 2021;22(5):bbab008. https://doi.org/10.1093/bib/bbab008 PMID: 33529337
14. Yuan Q, Chen K, Yu Y, Le NQK, Chua MCH. Prediction of anticancer peptides based on an ensemble model of deep learning and machine learning using ordinal positional encoding. Brief Bioinform. 2023;24(1):bbac630. https://doi.org/10.1093/bib/bbac630 PMID: 36642410
15. Yao L, Xie P, Guan J, Chung C-R, Zhang W, Deng J, et al. ACP-CapsPred: an explainable computational framework for identification and functional prediction of anticancer peptides based on capsule network. Brief Bioinform. 2024;25(5):bbae460. https://doi.org/10.1093/bib/bbae460 PMID: 39293807
16. Yao L, Li W, Zhang Y, Deng J, Pang Y, Huang Y, et al. Accelerating the Discovery of Anticancer Peptides through Deep Forest Architecture with Deep Graphical Representation. Int J Mol Sci. 2023;24(5):4328. https://doi.org/10.3390/ijms24054328 PMID: 36901759
17. Sun Y-Y, Lin T-T, Cheng W-C, Lu I-H, Lin C-Y, Chen S-H. Peptide-Based Drug Predictions for Cancer Therapy Using Deep Learning. Pharmaceuticals (Basel). 2022;15(4):422. https://doi.org/10.3390/ph15040422 PMID: 35455418
18. Zhou C, Peng D, Liao B, Jia R, Wu F. ACP_MS: prediction of anticancer peptides based on feature extraction. Brief Bioinform. 2022;23(6):bbac462. https://doi.org/10.1093/bib/bbac462 PMID: 36326080
19. Bian J, Liu X, Dong G, Hou C, Huang S, Zhang D. ACP-ML: A sequence-based method for anticancer peptide prediction. Comput Biol Med. 2024;170:108063. https://doi.org/10.1016/j.compbiomed.2024.108063 PMID: 38301519
20. Han B, Zhao N, Zeng C, Mu Z, Gong X. ACPred-BMF: bidirectional LSTM with multiple feature representations for explainable anticancer peptide prediction. Sci Rep. 2022;12(1):21915. https://doi.org/10.1038/s41598-022-24404-1 PMID: 36535969
21. He W, Wang Y, Cui L, Su R, Wei L. Learning embedding features based on multisense-scaled attention architecture to improve the predictive performance of anticancer peptides. Bioinformatics. 2021;37(24):4684–93. https://doi.org/10.1093/bioinformatics/btab560 PMID: 34323948
22. Feng G, Yao H, Li C, Liu R, Huang R, Fan X, et al. ME-ACP: Multi-view neural networks with ensemble model for identification of anticancer peptides. Comput Biol Med. 2022;145:105459. https://doi.org/10.1016/j.compbiomed.2022.105459 PMID: 35358753
23. Geng A, Luo Z, Li A, Zhang Z, Zou Q, Wei L, et al. ACP-CLB: An Anticancer Peptide Prediction Model Based on Multichannel Discriminative Processing and Integration of Large Pretrained Protein Language Models. J Chem Inform Model. 2025;65(5):2336–49.
24. Khan S. Deep-Representation-Learning-Based Classification Strategy for Anticancer Peptides. Mathematics. 2024;12(9):1330. https://doi.org/10.3390/math12091330
25. Wei L, Zhou C, Chen H, Song J, Su R. ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. Bioinformatics. 2018;34(23):4007–16. https://doi.org/10.1093/bioinformatics/bty451 PMID: 29868903
26. Rao B, Zhou C, Zhang G, Su R, Wei L. ACPred-Fuse: fusing multi-view information improves the prediction of anticancer peptides. Brief Bioinform. 2020;21(5):1846–55. https://doi.org/10.1093/bib/bbz088 PMID: 31729528
27. Akbar S, Hayat M, Tahir M, Chong KT. cACP-2LFS: Classification of Anticancer Peptides using Sequential Discriminative model of KSAAP and Two-Level Feature Selection Approach. IEEE Access. 2020;PP:1.
28. Manavalan B, Basith S, Shin TH, Choi S, Kim MO, Lee G. MLACP: machine-learning-based prediction of anticancer peptides. Oncotarget. 2017;8(44):77121–36. https://doi.org/10.18632/oncotarget.20365 PMID: 29100375
29. Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics. 2010.
30. Deng H, Ding M, Wang Y, Li W, Liu G, Tang Y. ACP-MLC: A two-level prediction engine for identification of anticancer peptides and multi-label classification of their functional types. Comput Biol Med. 2023;158:106844. https://doi.org/10.1016/j.compbiomed.2023.106844 PMID: 37058760

31. Tyagi A, Tuknait A, Anand P, Gupta S, Sharma M, Mathur D, et al. CancerPPD: a database of anticancer peptides and proteins. Nucleic Acids Res. 2015;43(Database issue):D837–43. https://doi.org/10.1093/nar/gku892 PMID: 25270878
32. Team E. ESM Cambrian: Revealing the mysteries of proteins with unsupervised learning: EvolutionaryScale Website; 2024. Available from: https://evolutionaryscale.ai/blog/esm-cambrian
33. Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M. AAindex: amino acid index database, progress report 2008. Nucl Acids Res. 2007;36(Database):D202–5.
34. Hayes T, Rao R, Akin H, Sofroniew NJ, Oktay D, Lin Z, et al. Simulating 500 million years of evolution with a language model. Science. 2025;387(6736):850–8. https://doi.org/10.1126/science.ads0018 PMID: 39818825
35. 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(6637):1123–30. https://doi.org/10.1126/science.ade2574 PMID: 36927031
36. Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell. 2005;27(8):1226–38. https://doi.org/10.1109/TPAMI.2005.159 PMID: 16119262
37. Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, Baker D. Improved protein structure prediction using predicted interresidue orientations. Proc Natl Acad Sci U S A. 2020;117(3):1496–503. https://doi.org/10.1073/pnas.1914677117 PMID: 31896580
38. Guo Y, Yan K, Lv H, Liu B. PreTP-EL: prediction of therapeutic peptides based on ensemble learning. Brief Bioinform. 2021;22(6):bbab358. https://doi.org/10.1093/bib/bbab358 PMID: 34459488
39. Yan K, Lv H, Wen J, Guo Y, Xu Y, Liu B. PreTP-Stack: Prediction of Therapeutic Peptide Based on the Stacked Ensemble Learning. IEEE/ACM Trans Comput Biol Bioinf. 2023;20(2):1337–44. https://doi.org/10.1109/tcbb.2022.3183018
40. Liang X, Zhao H, Wang J. MA-PEP: A novel anticancer peptide prediction framework with multimodal feature fusion based on attention mechanism. Protein Sci. 2024;33(4):e4966. https://doi.org/10.1002/pro.4966 PMID: 38532681
41. Wang X, Wang S. ACP-PDAFF: Pretrained model and dual-channel attentional feature fusion for anticancer peptides prediction. Comput Biol Chem. 2024;112:108141. https://doi.org/10.1016/j.compbiolchem.2024.108141 PMID: 38996756
42. Schaduangrat N, Nantasenamat C, Prachayasittikul V, Shoombuatong W. ACPred: A Computational Tool for the Prediction and Analysis of Anticancer Peptides. Molecules. 2019;24(10):1973. https://doi.org/10.3390/molecules24101973 PMID: 31121946
43. You H, Yu L, Tian S, Ma X, Xing Y, Song J, et al. Anti-cancer Peptide Recognition Based on Grouped Sequence and Spatial Dimension Integrated Networks. Interdiscip Sci. 2022;14(1):196–208. https://doi.org/10.1007/s12539-021-00481-0 PMID: 34637113
44. Zhu L, Ye C, Hu X, Yang S, Zhu C. ACP-check: An anticancer peptide prediction model based on bidirectional long short-term memory and multi-features fusion strategy. Comput Biol Med. 2022;148:105868. https://doi.org/10.1016/j.compbiomed.2022.105868 PMID: 35868046
45. Liang X, Li F, Chen J, Li J, Wu H, Li S, et al. Large-scale comparative review and assessment of computational methods for anti-cancer peptide identification. Brief Bioinform. 2021;22(4):bbaa312. https://doi.org/10.1093/bib/bbaa312 PMID: 33316035
46. Azim SM, Sabab NHN, Noshadi I, Alinejad-Rokny H, Sharma A, Shatabda S, et al. Accurately predicting anticancer peptide using an ensemble of heterogeneously trained classifiers. Inform Med Unlock. 2023;42:101348. https://doi.org/10.1016/j.imu.2023.101348
47. Lee B, Shin D. Contrastive learning for enhancing feature extraction in anticancer peptides. Brief Bioinform. 2024;25(3):bbae220. https://doi.org/10.1093/bib/bbae220 PMID: 38725157
48. Siméoni O, Vo HV, Seitzer M, Baldassarre F, Oquab M, Jose C, et al. Dinov3. arXiv preprint arXiv:250810104. 2025.
49. Wang S, Ma B. Anti-Cancer Peptides Identification and Activity Type Classification With Protein Sequence Pre-Training. IEEE J Biomed Health Inform. 2025;29(3):1692–701. https://doi.org/10.1109/JBHI.2024.3358632 PMID: 40048353
50. Ortega-Ochoa R, Aspuru-Guzik A, Vegge T, Buonassisi T. A tomographic interpretation of structure-property relations for materials discovery. arXiv preprint arXiv:250118163. 2025. https://arxiv.org/abs/250118163
51. Laurens VDM, Hinton G. Visualizing Data using t-SNE. J Mach Learn Res. 2008;9(2605):2579–605.
52. 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. 2022;44(10):7112–27. https://doi.org/10.1109/TPAMI.2021.3095381 PMID: 34232869
53. Brandes N, Ofer D, Peleg Y, Rappoport N, Linial M. ProteinBERT: a universal deep-learning model of protein sequence and function. Bioinformatics. 2022;38(8):2102–10. https://doi.org/10.1093/bioinformatics/btac020 PMID: 35020807
54. Rokach L, Maimon O. Clustering Methods. Springer US; 2005.
55. Huang Y, Huang J, Chen Y. Alpha-helical cationic antimicrobial peptides: relationships of structure and function. Protein Cell. 2010;1(2):143–52. https://doi.org/10.1007/s13238-010-0004-3 PMID: 21203984
56. Luan CH, Parker TM, Gowda DC, Urry DW. Hydrophobicity of amino acid residues: differential scanning calorimetry and synthesis of the aromatic analogues of the polypentapeptide of elastin. Biopolymers. 1992;32(9):1251–61. https://doi.org/10.1002/bip.360320914 PMID: 1420992
57. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res. 2023;51(D1):D523–31. https://doi.org/10.1093/nar/gkac1052 PMID: 36408920
58. Yu Q, Zhang Z, Liu G, Li W, Tang Y. ToxGIN: an In silico prediction model for peptide toxicity via graph isomorphism networks integrating peptide sequence and structure information. Brief Bioinform. 2024;25(6):bbae583. https://doi.org/10.1093/bib/bbae583 PMID: 39530430