PLOS Computational Biology

METHODS

PepLM-GNN: A graph neural network framework leveraging pre-trained language models for peptide-protein binding prediction

Ke Yan[1,2], Meijing Li[1], Shutao Chen[1], Tianyi Liu[1], Jing Hao[1,2], Bin Liu[1,2,3]*, Zhen Li[3]*

1 School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China, **2** Zhongguancun Academy, Beijing, China, **3** SMBU-MSU-BIT Joint Laboratory on Bioinformatics and Engineering Biology, Shenzhen MSU-BIT University, Shenzhen, Guangdong, China

* bliu@bliulab.net (BL); zli@bliulab.net (ZL)

Citation: Yan K, Li M, Chen S, Liu T, Hao J, Liu B, et al. (2026) PepLM-GNN: A graph neural network framework leveraging pre-trained language models for peptide-protein binding prediction. PLoS Comput Biol 22(3): e1014084. https://doi.org/10.1371/journal.pcbi.1014084

Editor: Mohammad Sadegh Taghizadeh, Shiraz University, IRAN, ISLAMIC REPUBLIC OF

Received: November 17, 2025

Accepted: March 2, 2026

Published: March 24, 2026

Data availability statement: Availability and Implementation: The data and online services of PepLM-GNN are publicly accessible via http://bliulab.net/PepLM-GNN. The source code of PepLM-GNN and related data are publicly available on GitHub (https://github.com/cokeyk/PepLM-GNN).

Abstract

Motivation

The precise prediction of peptide-protein interaction (PepPI) is a core support for promoting breakthroughs in peptide drug research, as well as understanding the regulatory mechanisms of biomolecules. Researchers have developed several computational methods to predict PepPI. However, existing computational methods also have significant limitations. At the level of data feature characterisation, the problem of PepPI does not conform to the Euclidean axioms, making it difficult for conventional prediction methods to effectively measure the underlying correlations between peptides and proteins. At the level of model generalisation performance, existing approaches are often hampered by insufficient generalisation ability, as manifested by their markedly degraded performance in cold start scenarios involving novel peptides, novel proteins, and novel binding pairs.

Results

In this study, we propose a computing framework, PepLM-GNN, that integrates a pre-trained language ProtT5 model with a hybrid graph network for accurate identification of PepPI. This model constructs a graph by using ProtT5-extracted semantic context features of peptides and proteins to form heterogeneous nodes, with edges connecting interacting peptide-protein pairs. The hybrid graph network Graph Convolutional Networks (GCN) provides the comprehensive information of the peptide and protein sequences, while employing the Graph Isomorphism Network (GIN) to capture the global interactions between them. Specifically, the GCN aggregates both the semantic context information of node sequences and local neighbourhood information, effectively representing non-Euclidean data. To capture the global associations, we adopt a GIN strategy to optimize the cross-node feature interaction and transfer process, thereby enhancing the generalisation performance of addressing the cold start

Funding: This work was supported by the Beijing Natural Science Foundation (No. L248013) to B.L., the National Natural Science Foundation of China (No. 62473049 to K.Y., U22A2039 to B.L.), the Zhongguancun Academy (Project No. 20240101) to B.L., and Beijing Institute of Technology Research Fund Program for Young Scholars to K.Y.

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

scenario. Compared with the existing advanced methods, PepLM-GNN demonstrated highly accurate performance and robustness in predicting the PepPI. We further demonstrated the capabilities of PepLM-GNN in virtual peptide drug screening, which is expected to facilitate the discovery of peptide drugs and the elucidation of protein functions.

Author summary

We propose a computational framework, PepLM-GNN, that integrates the ProtT5 pre-trained language model with a hybrid graph network. Specifically, the semantic features of peptides and proteins are extracted using ProtT5 to construct a graph. Within the hybrid graph network, GCN model aggregates semantic and local neighborhood information from node sequences, enabling an adequate representation of non-Euclidean data. Meanwhile, GIN model is utilized to optimize the process of cross-node feature interaction and transmission, thereby enhancing the generalization performance in addressing cold-start scenarios. Experimental results demonstrate that PepLM-GNN outperforms existing state-of-the-art methods in both accuracy and robustness for PepPI prediction. Moreover, PepLM-GNN can be applied to virtual peptide drug screening, thereby accelerating the development of peptide drugs. Furthermore, we have established a public online service platform (http://bliulab.net/PepLM-GNN) to facilitate the practical application.

1. Introduction

The peptide-protein interaction (PepPI) is widely present in living organisms and has been a core issue in the field of life sciences. It plays a crucial role in numerous key physiological processes, including cell signal transduction, metabolic regulation, and so on [1–3]. For instance, in the cellular signaling pathway, peptide hormones (such as insulin) can specifically bind to protein receptors on the cell surface, thereby triggering a series of cascade reactions and achieving precise regulation of biological blood glucose levels [4]. The peptide-protein complexes on the surface of antigen-presenting cells in the immune response can activate T cells to resist the invasion of pathogens [5–7].

Clarifying the PepPI is of great significance in many fields. In the process of drug development, it can help researchers clarify their goals and promote the emergence of new drugs [8–14]. Take cancer treatment as an example. Scientists can develop new anti-cancer drugs based on the abnormal PepPI within tumour cells [15]. These drugs have fewer side effects and more effective therapeutic benefits. In the field of disease mechanism research, elucidating the molecular mechanisms of the PepPI can reveal the root causes of diseases and open up new paths for their early diagnosis and effective treatment. For instance, in the study of neurodegenerative diseases, abnormal PepPI may provide new insights for combating these diseases [16].

In traditional research, PepPI have been determined through experimental methods such as yeast two-hybrid, co-immunoprecipitation, and surface plasmon resonance. However, these experimental methods differ in terms of cost, time, and throughput [2]. With the rapid update of deep learning technology, its powerful data-driven feature learning capabilities have provided certain solutions for predicting PepPI [17–19]. Compared to traditional machine learning methods that rely on manual feature engineering, deep learning models can automatically extract complex, nonlinear patterns from massive biological sequences and structural data, thereby significantly reducing the subjectivity and complexity of manual feature screening [14,20–28]. This advantage has been fully confirmed in fields such as natural language processing and image recognition [29,30]. In the PepPI study, deep learning has demonstrated great potential in uncovering hidden molecular associations, providing a novel strategy for analysing complex biological interaction mechanisms [31–34].

Among the existing deep learning-driven prediction methods for biomolecular interactions, the representative methods for PepPI prediction include CAMP [35] and IIDL-PepPI [36]. CAMP utilized the multi-channel sequence information and the conventional neural networks framework to predict the interactive and binding residues [35]. IIDL-PepPI constructed a bidirectional attention module to represent the contextual information of peptides and proteins, achieving pragmatic analysis of protein language. Meanwhile, it adopts a progressive transfer learning framework to address the peptide-protein interaction problem [36]. In addition, representative methods in the fields of protein-protein interaction (PPI) and drug-target interaction prediction also serve as references for related research. The HIGH-PPI constructed a two-view graph network to obtain the cross-protein global interaction and the local residue association within protein-protein interactions, respectively [37]. DrugBAN utilized a domain-adapted deep bilinear attention network framework to predict the local interactions of drug-target pairs [38].

However, these computational methods of PepPI prediction still have several challenges. Firstly, PepPI data belong to non-Euclidean data, where the peptide-protein interaction presents complex network topological characteristics. For instance, the same protein interacts with multiple peptides, forming distinct binding sites for each peptide. Relying solely on hand-crafted features is insufficient to capture these complex bindings explicitly. Secondly, the existing methods lie in their limited generalisation capability, especially for the cold start scenario of “novel peptides/novel proteins/novel binding pairs”.

In this study, we propose an accurate and interpretable framework, PepLM-GNN, that integrates a pre-trained language ProtT5 model and hybrid graph neural networks. The core innovation lies in optimizing PepPI modelling through the collaborative design of GCN and GIN. The nodes in the graph network are the semantic context features of peptides and proteins, while the edges are their interaction relationship. GCN extracts the local-level information via the feature associations between nodes and their direct neighbors. GIN is used to capture the global-level topological structure through the equivalence class discrimination mechanism. Specifically, the contributions of this article include:

(1) Our model employs GCN’s local neighbourhood aggregation to extract feature associations between nodes and their direct neighbours effectively. This process provides detailed support for micro-interaction in the model and meeting the non-Euclidean data modelling requirements of PepPI.

(2) To mitigate the limitations of GCN (e.g., gradient dispersion and over-smoothing that lead to the feature discriminability), we introduce GIN to enhance feature discriminability and capture the global topological structure. The hybrid framework enhances the generalisation ability and adaptability to the cold start scenario, including novel peptides, novel proteins, and novel binding pairs.

(3) An interpretable interaction subgraph is constructed by extracting important associations from the PepLM-GNN. The biological significance of the interpretable components derived from the subgraph is evaluated through a functional enrichment analysis experiment. Moreover, PepLM-GNN has been extended for applications in the alanine scanning of peptides, demonstrating a probability-driven approach to peptide drugs.

(4) To facilitate the use of researchers, we have built an online web server based on the proposed model, which can be accessed at http://bliulab.net/PepLM-GNN.

2. Results and discussion

2.1 Comparison with the other baseline methods on the benchmark dataset

In this section, we compared the model's performance with the other baseline methods on the benchmark dataset. The compared baseline methods include traditional machine learning models (LR [39], RF [40], SVM [41,42]) and classic deep learning models (CAMP [35], DrugBAN [38], IIDL-PepPI [36], HIGH-PPI [37]). The performance of the models is evaluated using four metrics: AUC, AUPR, F1, and ACC.

As shown in **Fig 1** and S1 Table, the PepLM-GNN demonstrates advantages in the binary PepPI prediction task. Its AUC of 0.8434 is superior to baseline methods, demonstrating the proposed method's excellent ability to distinguish PepPI. Compared with traditional machine learning methods that relied solely on hand-crafted features, the proposed method utilized the pre-trained ProtT5 [43] language model as a sequence encoder. This approach provides a more comprehensive and semantically rich representation of biological sequences, thereby significantly improving the model's expressive power and predictive accuracy. Compared with the IIDL-PepPI [36] that utilized the pre-trained ProtBert [44] model and progress transfer learning, the proposed method utilized the GCN framework to capture the non-Euclidean data space of PepPI. By propagating and aggregating feature information across neighbouring nodes, the proposed method effectively learns from the local interaction context. Although methods such as DrugBAN [38] and HIGH-PPI [37] utilized network topological structures, they lack high-quality sequence semantic priors and have relatively low prediction performance. Therefore, PepLM-GNN achieves superior performance in identifying PepPI and maintains stable performance compared to other baseline methods.

To further verify the statistical significance of the performance superiority of PepLM-GNN, we conducted statistical t-tests on the mean ACC values of PepLM-GNN and all comparative baseline methods derived from five-fold cross-validation on the benchmark dataset, and the specific *p*-values are summarized in **Table 1**. As shown in **Table 1**, all *p*-values of the statistical t-tests between PepLM-GNN and other baseline methods are less than 0.05, which quantitatively demonstrates that the PepLM-GNN is statistically significantly higher than the comparative methods, and the performance advantage of our model on the benchmark dataset is not caused by random factors.

2.2 Comparison with other baseline methods on independent test and cold start test datasets

To verify the model's generalisation ability, we conduct comparative experiments on four independent test datasets, including LEADS-PEP, Test167, Test251, and Test1440. The compared baseline methods contain traditional machine learning methods (SVM, RF, LR) and deep learning methods (CAMP [35], HIGH-PPI [37], DrugBAN [38], IIDL-PepPI [36], DeepRank-GNN-esm [45], DeepGNHV [46]). The hyperparameters of all baseline methods were optimized to their optimal values, ensuring a fair comparison. We evaluated the performance in terms of AUC, AUPR, F1, and ACC on the four independent test datasets, and the results are illustrated in **Fig 2**.

As shown in **Fig 2**, the results indicate that PepLM-GNN outperforms other state-of-the-art methods on binary PepPI prediction. The limited generalization of baseline methods, such as the CAMP [35], makes it challenging to capture the fundational sequence semantic knowledge, which is critical for predicting unseen PepPI. In contrast, the proposed method utilizes the pre-trained ProtT5 [43] model, which was constructed based on the large-scale unsupervised protein sequence, to provide rich semantic information. Moreover, the proposed method utilizes the GIN to capture the global topological association between the peptide and protein. By learning these inherent relational associations, PepLM-GNN has a generalizable model of interaction information, thereby improving its generalization capability for accurate binding prediction on unseen PepPI.

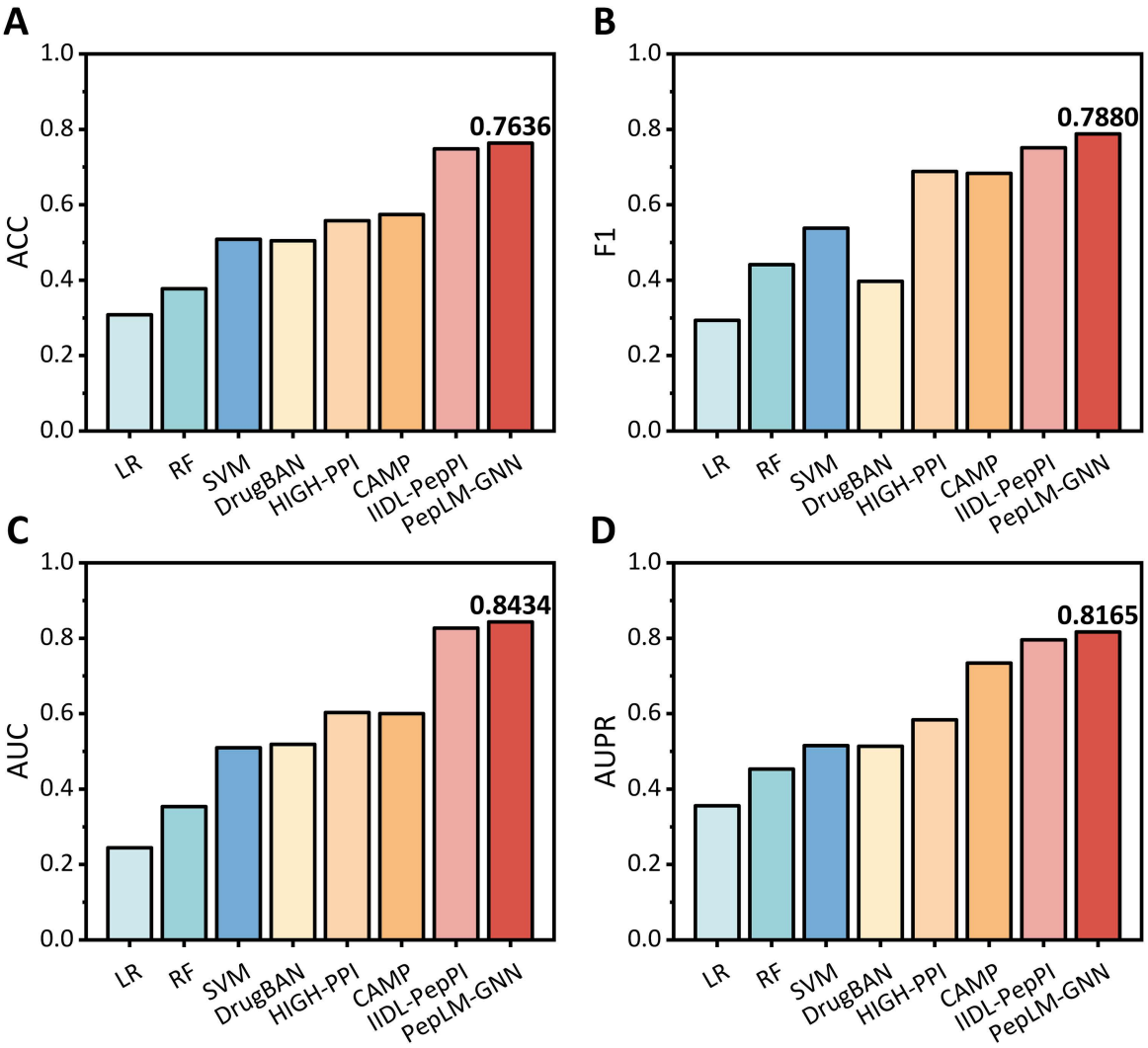

Fig 1. Performance of PepLM-GNN against other baseline methods on the benchmark dataset via five-fold cross-validation. The benchmark dataset (total 17244 samples) is derived from peptide-protein complex structures in the RCSB PDB database (before October 2022), containing 8622 positive samples (interacting peptide-protein pairs) and 8622 negative samples (non-interacting pairs) after data filtering. All reported performance metrics (e.g., ACC, AUC) represent the average values obtained from the five-fold cross-validation conducted on the benchmark dataset.

https://doi.org/10.1371/journal.pcbi.1014084.g001

Table 1. Statistical t-test *p*-values for ACC of comparative methods versus PepLM-GNN. This table reports the *p*-values of statistical t-tests between PepLM-GNN and other baseline methods, based on the mean ACC values from five-fold cross-validation on the benchmark dataset (17244 samples: 8622 positive and 8622 negative pairs). A *p*-value < 0.05 indicates that the PepLM-GNN is statistically significantly higher than that of the comparative method.

Method	LR	RF	SVM	CAMP	DrugBAN	IIDL-PepPI	HIGH-PPI
***p*-value**	8.07×10^{-8}	1.55×10^{-7}	8.18×10^{-7}	2.67×10^{-6}	7.67×10^{-7}	2.83×10^{-2}	1.93×10^{-6}

https://doi.org/10.1371/journal.pcbi.1014084.t001

To confirm the statistical significance of the performance advantage of PepLM-GNN on independent test data, we further conducted statistical t-tests on the mean ACC values. The test data was derived from the union of four independent test datasets, and the mean ACC values were calculated via five-fold cross-validation. The specific *p*-values of the t-tests

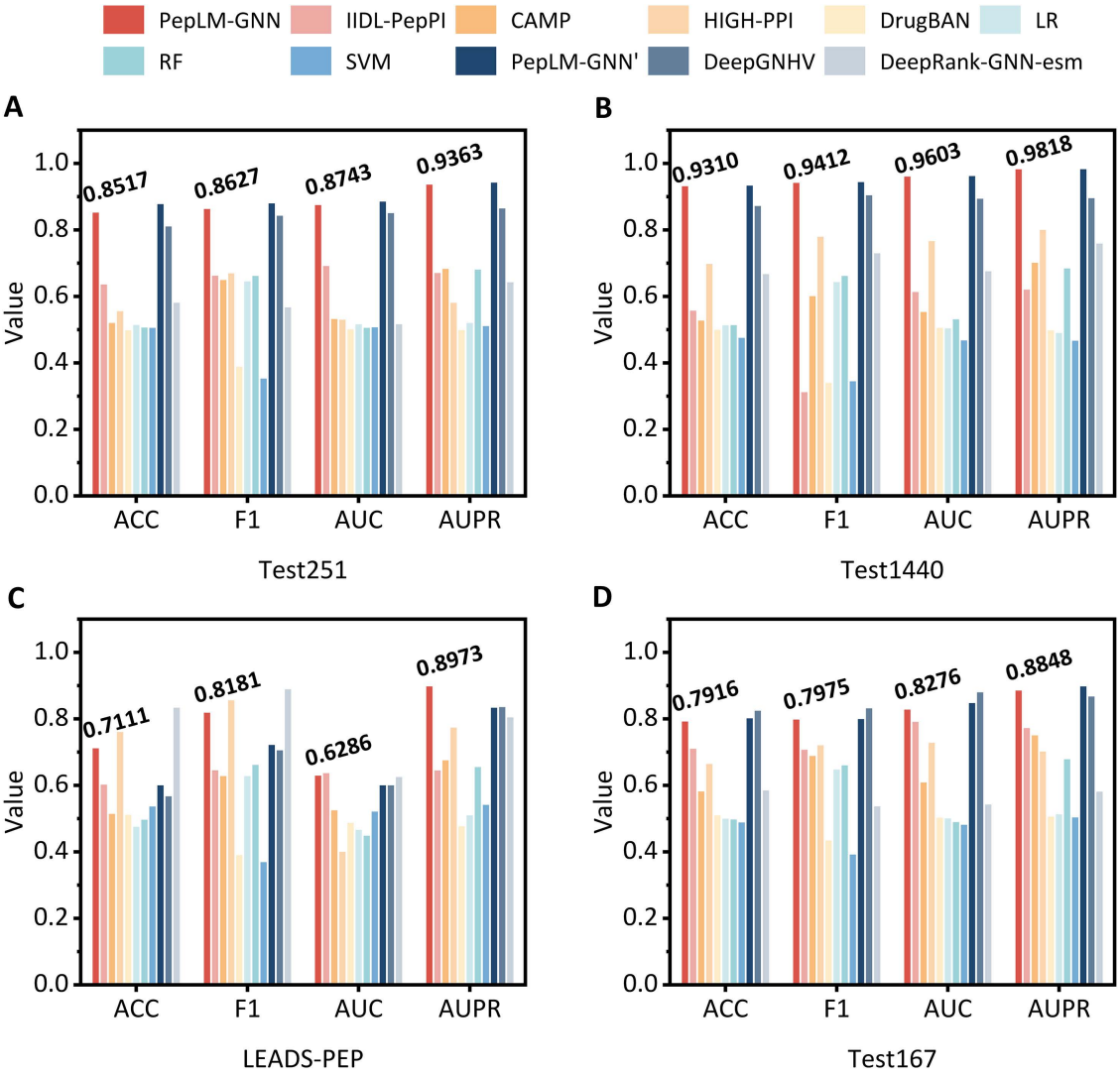

Fig 2. Performance of PepLM-GNN against other baseline methods on four independent test datasets. The four test datasets include: Test1440 (1440 positive peptide-protein pairs and 1440 negative pairs, sourced from the RCSB PDB database, January 2023-July 2024), LEADS-PEP (52 positive pairs and 52 negative pairs, a classic benchmark for evaluating peptide-protein docking performance), Test251 (249 positive pairs and 249 negative pairs), and Test167 (255 positive pairs and 255 negative pairs, derived from the RCSB PDB database, October-December 2022). PepLM-GNN', Deep-GNHV, Deep-GNN-esm are only applied to test subsets with available structural data. All performance metrics represent the mean values averaged across the five models derived from five-fold cross-validation on each independent test set.

https://doi.org/10.1371/journal.pcbi.1014084.g002

are summarized in Table 2. It can be seen from Table 2 that all *p*-values of the statistical t-tests between PepLM-GNN and the comparative methods were less than 0.05, indicating that the superior performance of PepLM-GNN on the integrated test set is statistically significant, and this advantage is stable and reliable.

Furthermore, to systematically evaluate the generalisation ability of the PepLM-GNN, we utilize the CD-HIT algorithm to construct cold-start test datasets comprising novel peptides, novel proteins, and novel protein-peptide pairs. PepLM-GNN was evaluated against other advanced deep learning methods using AUC and AUPR, and the results are shown in Fig 3.

The results showed that (1) PepLM-GNN outperformed compared methods, including IIDL-PepPI [36], CAMP [35], DrugBAN [38], and HIGH-PPI [37] in the four clustering threshold scenarios. (2) When the clustering threshold decreases

Table 2. Statistical t-test *p*-values for ACC between comparative methods and PepLM-GNN across four combined test sets. This table reports the *p*-values of statistical t-tests between PepLM-GNN and other comparative methods. The test data is the union of four independent test datasets. T-tests are based on the mean ACC values from five-fold cross-validation across five folds on the combined dataset. A *p*-value < 0.05 indicates that PepLM-GNN is statistically superior.

(a) First part of comparative methods					
Method	**LR**	**RF**	**SVM**	**CAMP**	**DrugBAN**
p-value	6.34×10^{-6}	6.29×10^{-6}	4.37×10^{-6}	1.71×10^{-5}	6.07×10^{-6}
(b) Second part of comparative methods					
Method	**IIDL-PepPI**	**HIGH-PPI**	**DeepRank-GNN-esm**	**DeepGNHV**	
p-value	1.10×10^{-3}	2.01×10^{-5}	5.72×10^{-6}	1.88×10^{-2}	

https://doi.org/10.1371/journal.pcbi.1014084.t002

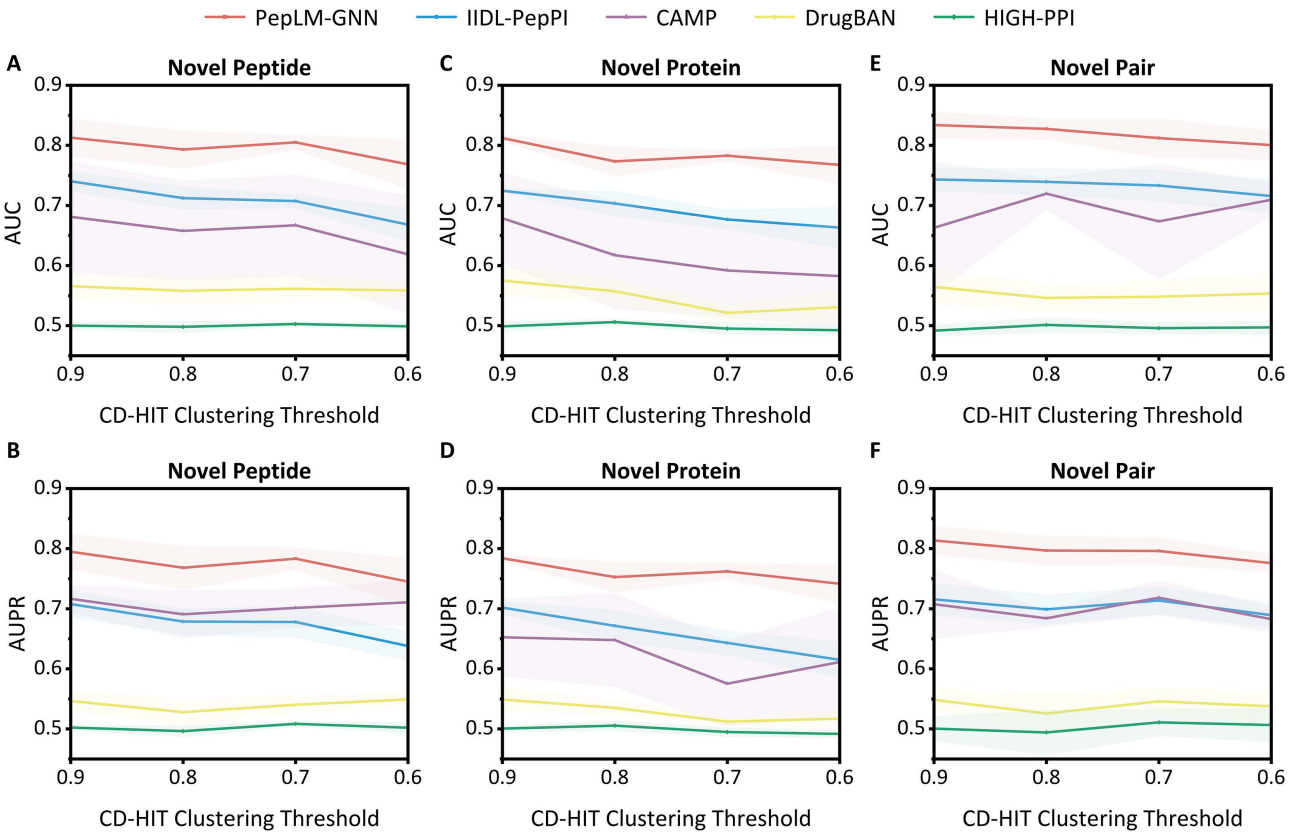

Fig 3. Comparison of PepPI predictions based on cluster-split datasets for predicting novel peptides, proteins, and peptide-protein pairs. Error bars represent the mean ± standard deviation of cross-validation experiments. The cluster-split (cold start) dataset is constructed using the CD-HIT clustering algorithm with four thresholds (0.6, 0.7, 0.8, 0.9), following the CAMP strategy: no entities from the same cluster appear in both training and test sets, resulting in three sub-datasets ("novel peptides", "novel proteins", "novel binding pairs"). All performance metrics are the mean ± standard deviation of five folds from five-fold cross-validation.

https://doi.org/10.1371/journal.pcbi.1014084.g003

and the sequence similarity between the training set and the test set weakens, the decline in the model's performance is relatively small, demonstrating a robust generalisation ability. By leveraging its equivalence class discrimination mechanism,the GIN model in the proposed method accurately capture the global topological structure of molecular interactions. When the sequence similarity decreases, cross-sample association can still be achieved through the consistency of

structural patterns, thereby enhancing the model's generalization ability. Therefore, PepLM-GNN demonstrates superior and more robust performance compared to the baseline methods on the independent test and cold start test datasets.

2.3 Comparative performance of PepLM-GNN with existing pre-trained language models

In this section, we evaluated four pre-trained language models [47] on the benchmark dataset to assess the impact of different feature encoding strategies on the PepPI prediction. The four embedding features generated by pre-trained language models are TAPE [48], ProtBert [44], ESM-2 [49], ESM-3 [50], and the proposed method based on ProtT5 [43]. TAPE [48] is mainly used for modelling the context of biological sequences, ProtBert [44] focuses on general semantic representation, and ESM-2 [49] integrates evolutionary information. As shown in **Table 3**, the PepLM-GNN model exhibits outstanding predictive performance across all four metrics. Unlike the biological language model (e.g., TAPE [48] and ProtBert [44]), the ProtT5 [43] can capture the subtle differences in key functional motifs of short peptide sequences and model the context of large-scale corpora [51]. By leveraging adaptive sequence embedding and hybrid graph networks, PepLM-GNN can effectively predict the interaction of peptide-protein pairs.

Table 4 shows that the *p*-values of the t-tests between the PepLM-GNN and all variant models are less than 0.05, indicating that the ACC of the PepLM-GNN using ProtT5 is statistically significantly higher than that of the variants with other pre-trained models substituted. This confirms that ProtT5 is more suitable for the PepPI prediction task in our model framework and can provide more effective feature representation for subsequent graph neural network processing.

2.4 Ablation study of PepLM-GNN on the benchmark dataset

To explore the contribution of three components of PepLM-GNN to the prediction performance, this section designs an ablation study on the benchmark set. We systematically developed several variant models by ablating each of the three individual modules. Model A only retains the ProtT5 [43] embedding and classification module, and the features obtained through ProtT5 [43] are directly used for classification. The GAT module was added to Model A to create Model B, which

Table 3. Performance comparison between PepLM-GNN and other pre-trained language models on the benchmark dataset in terms of ACC, F1, AUC, and AUPR. The benchmark dataset contains 17,244 samples (8,622 positive and 8,622 negative pairs). All metrics (ACC, F1, AUC, AUPR) are the mean ± standard deviation of five folds from five-fold cross-validation.

Method	ACC	F1	AUC	AUPR
ProtBert [44]	$0.6620_{\pm 0.0154}$	$0.7299_{\pm 0.0061}$	$0.7565_{\pm 0.0087}$	$0.7357_{\pm 0.0127}$
TAPE [48]	$0.6850_{\pm 0.0111}$	$0.7427_{\pm 0.0042}$	$0.7753_{\pm 0.0095}$	$0.7533_{\pm 0.0098}$
ESM-2 [49]	$0.6924_{\pm 0.0065}$	$0.7440_{\pm 0.0042}$	$0.7864_{\pm 0.0061}$	$0.7653_{\pm 0.0078}$
ESM-3 [50]	$0.7013_{\pm 0.0068}$	$0.7500_{\pm 0.0034}$	$0.7950_{\pm 0.0032}$	$0.7690_{\pm 0.0052}$
PepLM-GNN	$\mathbf{0.7636_{\pm 0.0117}}$	$\mathbf{0.7880_{\pm 0.0034}}$	$\mathbf{0.8434_{\pm 0.0060}}$	$\mathbf{0.8165_{\pm 0.0103}}$

Note: The specific versions/parameter configurations of the pre-trained language models used in this study are as follows: ProtT5: ProtT5-XL-UniRef50; ESM-2: esm2_t33_650M_UR50D; ESM-3: esmc-300m-2024-12; ProtBert: ProtBert-BFD; TAPE: ProteinBertModel (bert-base). All models adopt the default parameter settings consistent with their official pre-trained versions and original published papers.

https://doi.org/10.1371/journal.pcbi.1014084.t003

Table 4. *P*-values of statistical t-test for ACC: PepLM-GNN with different pre-trained model on five-fold cross-validation set. This table reports the *p*-values of statistical t-tests between the PepLM-GNN (using ProtT5) and other pre-trained language models. T-tests are based on the mean ACC values from five-fold cross-validation on the benchmark dataset (17244 samples: 8622 positive + 8622 negative pairs). A *p*-value < 0.05 indicates that the original PepLM-GNN has a statistically significantly higher ACC than the variant.

Method	ProtBert	TAPE	ESM-2	ESM-3
***p*-value**	3.19×10^{-5}	8.78×10^{-5}	1.29×10^{-4}	2.18×10^{-4}

https://doi.org/10.1371/journal.pcbi.1014084.t004

was then used to evaluate the impact of GAT on overall performance. The GCN module was added to Model A to create Model C, which was used to compare the effectiveness of GCN in the prediction task. The GIN module was added to Model A to create Model D, which was further used to verify the performance of GIN in the prediction task. The ablation study results for ACC and AUC are summarised in Table 5.

As shown in Table 5, model A retains only the pre-training and classification modules, performing the worst in terms of ACC and AUC. This model relies solely on sequence features for end-to-end classification and is unable to capture the interaction relationships between molecules, making it challenging to identify the PepPI. Compared with model A, models C and D exhibit higher performance, highlighting the unique value of graph networks. GCN enhances the structural connectivity of nodes by aggregating local neighbourhood features, while GIN enhances the perception of the global topological structure through iterative updates. At the same time, it can alleviate the problems of gradient dispersion and over-smoothing that are prone to occur in the deep training of GCN. These two mechanisms address the information deficiency of single-sequence features from both local and global perspectives. The research model PepLM-GNN combines the fine-grained local-structure modelling of GCN with the in-depth exploration of global topology by GIN, forming a functional synergy. Therefore, the ablation studies demonstrated that integrating the interaction network topology with sequence semantics information is essential for achieving effective prediction in the PepPI problem.

To further quantitatively verify the significant contribution of each core module to the model's performance, we conducted t-tests on the mean ACC values of the full PepLM-GNN model and each ablation variant (based on five-fold cross-validation on the benchmark dataset), and the corresponding *p*-values are reported in Table 6. It can be observed from Table 6 that all *p*-values of the statistical t-tests between the complete model and each ablation variant are less than 0.05, which statistically confirms that removing any core module will lead to a significant decrease in the ACC of the model. This fully demonstrates that each core module of PepLM-GNN (GCN, GIN, etc.) makes a statistically significant contribution to the model's PepPI prediction performance, and the model's hybrid architecture design is scientifically sound and reasonable.

Table 5. Ablation experiment performance of PepLM-GNN on the benchmark set using five-fold cross-validation. Ablation variants include: A (pretrained module + classification module), B (pretrained module + GAT + classification module), C (pretrained module + GCN + classification module), D (pretrained module + GIN + classification module), and the complete PepLM-GNN model (pretrained module + GCN + GIN + classification module). All metrics (ACC, AUC) are the mean ± standard deviation across five folds of a five-fold cross-validation.

Method	GCN module	GIN module	GAT module	ACC	AUC
A	✖	✖	✖	$0.7350_{\pm 0.0134}$	$0.8160_{\pm 0.0080}$
B	✖	✖	✓	$0.7351_{\pm 0.0105}$	$0.8223_{\pm 0.0096}$
C	✓	✖	✖	$0.7453_{\pm 0.0107}$	$0.8346_{\pm 0.0036}$
D	✖	✓	✖	$0.7475_{\pm 0.0235}$	$0.8341_{\pm 0.0149}$
PepLM-GNN	✓	✓	✖	$\mathbf{0.7636_{\pm 0.0117}}$	$\mathbf{0.8434_{\pm 0.0060}}$

https://doi.org/10.1371/journal.pcbi.1014084.t005

Table 6. *P*-values of statistical t-test for ACC: Ablation experiments of PepLM-GNN on five-fold cross-validation set. This table reports the *p*-values from t-tests comparing the complete PepLM-GNN model with each of its ablation variants. Ablation variants include: A (pretrained module + classification module), B (pretrained module + GAT + classification module), C (pretrained module + GCN + classification module), D (pretrained module + GIN + classification module). T-tests are based on the mean ACC values from five-fold cross-validation on the benchmark dataset (17244 samples: 8622 positive + 8622 negative pairs). A *p*-value < 0.05 indicates that the removed module makes a statistically significant contribution to the ACC of PepLM-GNN.

Method	A	B	C	D
***p*-value**	4.01×10^{-3}	4.06×10^{-3}	1.75×10^{-2}	2.54×10^{-2}

https://doi.org/10.1371/journal.pcbi.1014084.t006

2.5 The application of PepLM-GNN in extended tasks

To explore the interpretability and biological significance of PepPI prediction, we conducted multi-level experiments, including the key residue identification and interaction network analysis, and gene enrichment analysis.

2.5.1 Virtual screening of key residues involved in the interaction. To verify the practicality of PepLM-GNN, we combined computer simulation and experimental data to compare the model prediction results with the biometric results and verify its consistency in identifying key amino acids. We conducted alanine-scanning virtual screening of the SALL4 (1–12) peptide segment [52]. Replace each amino acid in this peptide segment with alanine in sequence to construct a series of mutants. Using the established model, predict the binding probability of each mutant to RBBp4, respectively. To quantify the impact of alanine substitution on the peptide-protein binding ability, the relative change in binding likelihood can be expressed as (binding probability of wild-type SALL4 (1–12) peptide-binding probability of alanine mutant)/binding probability of wild-type SALL4 (1–12) peptide [36].

The model predicts that the relative changes in the binding probabilities of alanine mutants corresponding to the peptide "RRK" sequence (arginine R, lysine K) are all very high. Compared with the experimental data of the binding free energy measured in the SALL4 study [52], as shown in **Fig 4A**, the trends of the two are highly consistent. Among the three residue positions with the most significant changes in binding free energy measured experimentally, two correspond to the "RRK" sequence mutants, highlighting the core role of this sequence in the interaction between SALL4 and RBBp4. This finding is consistent with the conclusion of "Arg3, Arg4, Lys5 as key interaction residues" in the SALL4 study, which verifies the consistency between model predictions and experimental measurements in identifying key amino acids. This high consistency between prediction and experiment stems from the in-depth analysis of PepPI by the model architecture. The ProtT5 [43] pre-trained model mined the semantic information of sequences such as "RRK", providing biological semantic support for identifying key residues. The graph network framework can precisely model the local residue interactions at the binding interface, perceive the global topological structure, and accurately capture the microscopic interactions between the "RRK" and RBBp4 binding sites. The synergy effect of this model architecture enables it to identify functional key regions at the sequence semantic level and analyse the specific contributions of residues at the network topology level, achieving precise identification of key binding residues.

2.5.2 Subgraph identification of FFW interacting proteins. To study the interpretability and biological significance of the model's predictions, we used the target peptide FFW as an example for subgraph analysis and functional verification. FFW is an HCC therapeutic peptide developed based on the SALL4 key sequence, and its functional core lies in the specific interaction with RBBp4. We analysed the proposed model using the GNNExplainer tool [53], extracted the protein subplots that might interact with the target peptide FFW, and presented the core proteins with the highest rankings. **Fig 4B** shows that these proteins form functional modules through close physical interactions, suggesting that they may jointly participate in the signalling pathways related to liver cancer development.

2.5.3 Functional enrichment and clinical significance validation of FFW interacting proteins. This section performs a functional enrichment analysis on the aforementioned core proteins [36]. **Fig 4C** shows that they are significantly enriched in liver cancer-related pathways (enrichment p-value 3.63×10^{-2}). This not only verified the accuracy of the model in identifying key interaction pairs (FFW-RBBp4) but also directly confirmed the mechanism by which "RRK residues regulate tumorigenesis by targeting RBBp4", which is consistent with the core role of the "RRK" sequence in SALL4. From the virtual screening of key residues in SALL4 to the validation of FFW interaction networks and pathways, the proposed model, PepLM-GNN, demonstrates stable interpretability and biological consistency in PepPI prediction.

3. Conclusion

This study proposes a PepLM-GNN method that integrates a pre-trained ProtT5 model and hybrid graph neural networks for the PepPI prediction. The model employs a GCN to extract the local neighbourhood features of molecules, enabling the capture of short-range interactions and precise modelling of non-Euclidean structures. Furthermore, GIN relies on the

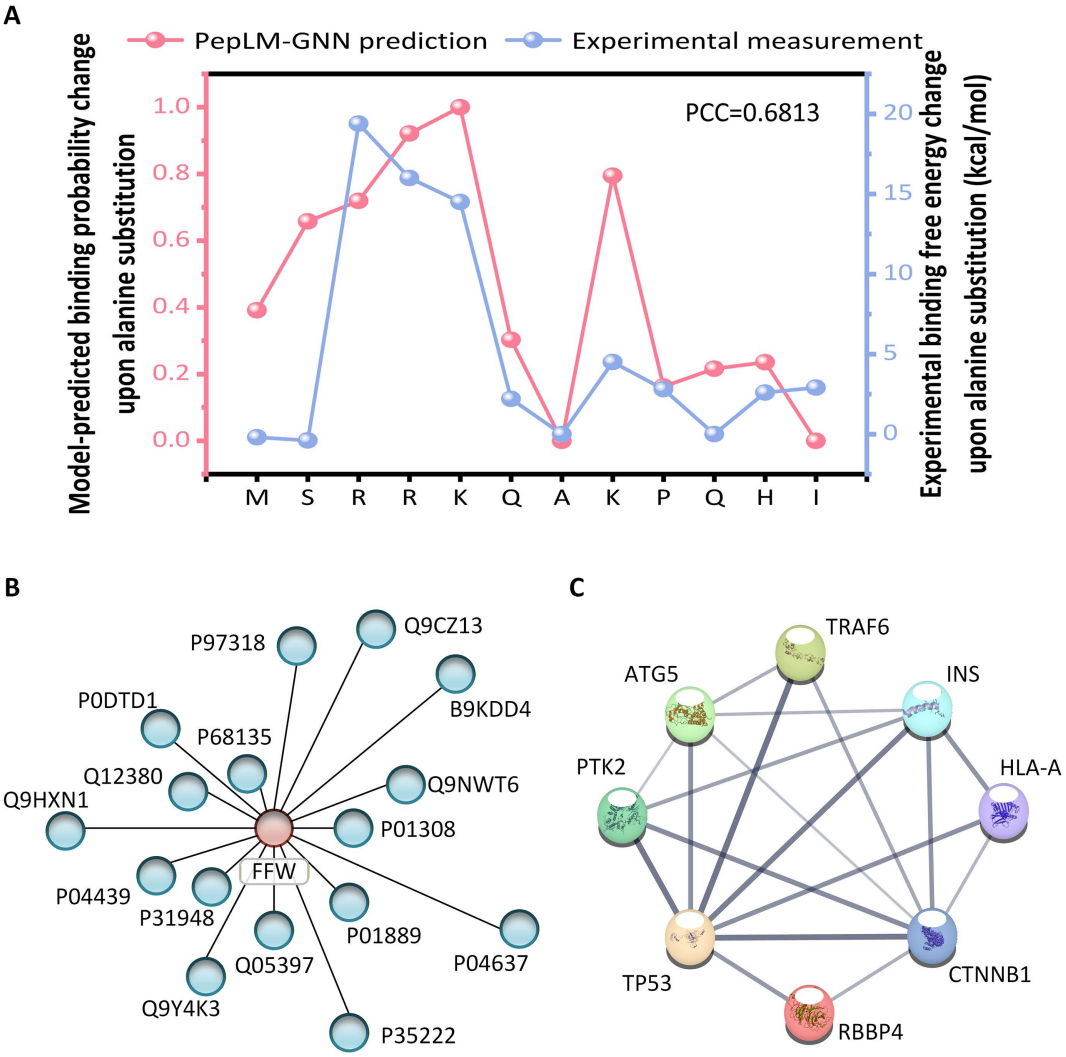

Fig 4. Application of PepLM-GNN in virtual peptide drug screening. (A) Relative changes in the predicted binding probability of RBBp4 and SALL4 peptide alanine mutants by PepLM-GNN compared with experimentally measured changes in binding free energy. **(B)** Protein subgraphs interacting with FFW peptides extracted using the GNNExplainer tool. **(C)** Functional enrichment analysis of proteins interacting with the FFW peptide.

https://doi.org/10.1371/journal.pcbi.1014084.g004

equivalence class discrimination mechanism to capture the global node interaction pattern, thereby alleviating the gradient dispersion and over-smoothing problems in deep training of GCN and improving generalisation performance. Compared to the other state-of-the-art methods, PepLM-GNN has higher prediction accuracy and interpretability. Besides, the application of PepLM-GNN in various extended tasks highlights the framework's versatility and potential to facilitate broader protein-related research. This achievement can be applied in areas such as the development of protein drugs, efficacy evaluation, and optimization of treatment strategies, thereby accelerating the drug research and development process. Additionally, a web server has been established and is accessible at http://bliulab.net/PepLM-GNN.

It should be noted that PepLM-GNN still has certain limitations. There are potential biases in dataset construction: although the existing benchmark dataset has undergone rigorous screening, it still falls short of fully covering protein-peptide pairs across different species and functional types, which may compromise the model's prediction performance

on rare or specialized samples. In addition, the graph construction step incurs a certain computational cost: the process of constructing molecular graphs based on sequence features consumes considerable computational resources, making it difficult to adapt to the rapid deployment requirements of low-configuration computing environments.

4. Materials and methods

4.1 Benchmark, independent test, and cold start test datasets

In this study, we construct three types of datasets to evaluate the performance of the PepPI prediction model, namely the benchmark dataset, four independent test datasets, and a cold start test dataset.

Benchmark dataset: The benchmark dataset used in this study is derived from the peptide-protein complex structures in the RCSB PDB database [54] before October 2022. The determination of residue pairs is based on the distance threshold method between α-carbon atoms (C_α). When the C_α atom distance between the ligand peptide and the target protein residue is less than 5 Å, it is defined as an interacting residue pair. For amino acid residues lacking C_α atoms, the nearest inter-atomic distance is used for judgment. If the structural coordinates are incomplete or the resolution is low, making accurate judgment impossible, they are uniformly regarded as non-binding residue pairs. Then we excluded peptide-protein pairs where the proportion of unknown or non-standard amino acids exceeded 20%. After the above processing, 8622 pairs of peptide-protein interaction samples were obtained, which served as positive samples for the experiment. To construct a balanced dataset, we randomly selected an equal number of peptide-protein pairs with no interaction relationship as negative samples.

Independent test dataset: In this study, an independent test dataset, Test1440, was independently constructed. The data were sourced from peptide-protein interaction complexes in the RCSB PDB database from January 2023 to July 2024, containing a total of 1440 effective complex samples. Negative samples are generated by randomly pairing peptides with proteins to construct non-interacting pairs. In addition, we have introduced three independent test sets that have been published and widely verified in various fields. The LEADS-PEP test set [55], as a classic benchmark dataset for evaluating peptide-protein docking performance, contains a total of 52 pairs of positive and negative samples. The Test251 test set [56,57] is a commonly used independent validation dataset in the field, containing 249 pairs of positive and negative samples, and has undergone strict data screening to ensure no redundancy with the training data. The Test167 test set [36], with data derived from the peptide-protein complex in the RCSB PDB database from October to December 2022, was processed through steps such as extracting residue level labels with the PBD-BRE tool [58], filtering samples with a proportion of non-standard amino acids exceeding 20%, and excluding samples with a sequence similarity of over 80% to the training set/validation set. Ultimately, 255 pairs of positive and negative samples were retained.

Cold-start test dataset: According to the CAMP strategy [33], we independently cluster proteins and peptides using the CD-HIT algorithm (i.e., the clustering processes for the two sequence types are carried out separately, only for a single sequence type). The core data partitioning rule requires that entities from the same cluster cannot simultaneously appear in the training set and the test set, thereby ensuring the “novelty” of the cold start scenario. Based on this, we divide the benchmark data into three cold-start subsets. “New peptides” (only peptide clustering, with no overlapping peptide clusters between the training set and the test set), “new proteins” (only protein clustering, with no overlapping protein clusters between the training set and the test set), and “new binding pairs” (relying on the independent clustering results of proteins and peptides). The cold-start test dataset uses four consistent CD-HIT clustering thresholds (0.6, 0.7, 0.8, and 0.9) to construct protein and peptide sequences. It simultaneously performs independent clustering of the two sequence types under the same thresholds.

4.2 The architecture of the PepLM-GNN

To address the precise modelling and interpretable requirements of PepPI, this study proposes the PepLM-GNN method, a hybrid graph neural network (GNN) that achieves efficient prediction through feature learning and network topology. In

the graph construct stage, a global PepPI network is established. In this network, nodes represent individual peptide or protein entities, while edges denote interaction relationships. This structure forms complex associations (e.g., a single peptide connecting to multiple target proteins, and vice versa). The hybrid graph network model adopts the collaborative architecture of GCN and GIN to achieve the local-level interaction and global-level topology information capture, respectively. GCN relies on the local neighbourhood aggregation mechanism to obtain micro-interaction details from the local topology of nodes, adapting to the non-Euclidean data characteristics of PepPI. GIN encodes the global topological dependencies of peptide-protein molecules through an equivalence class discrimination mechanism and iterative feature aggregation, while alleviating the problems of gradient dispersion and over-smoothing in the deep training of GCN. This optimizes the differentiated representation of node features, ensuring that the model can distinguish the interaction patterns of different PepPI. The model architecture is shown in Fig 5. Our method consists of four modules: ProtT5, graph convolution, graph isomorphism, and classification.

4.2.1 ProtT5-based feature extraction. This study constructed a sequence feature analysis architecture based on a pre-trained ProtT5 model to extract the sequence semantic context features of peptides and proteins [59]. The pre-trained ProtT5 model employs a text generation architecture and a cross-modal coding mechanism to capture the global semantic context of sequences, including their inherent long-term dependencies in sequences. When a peptide or protein

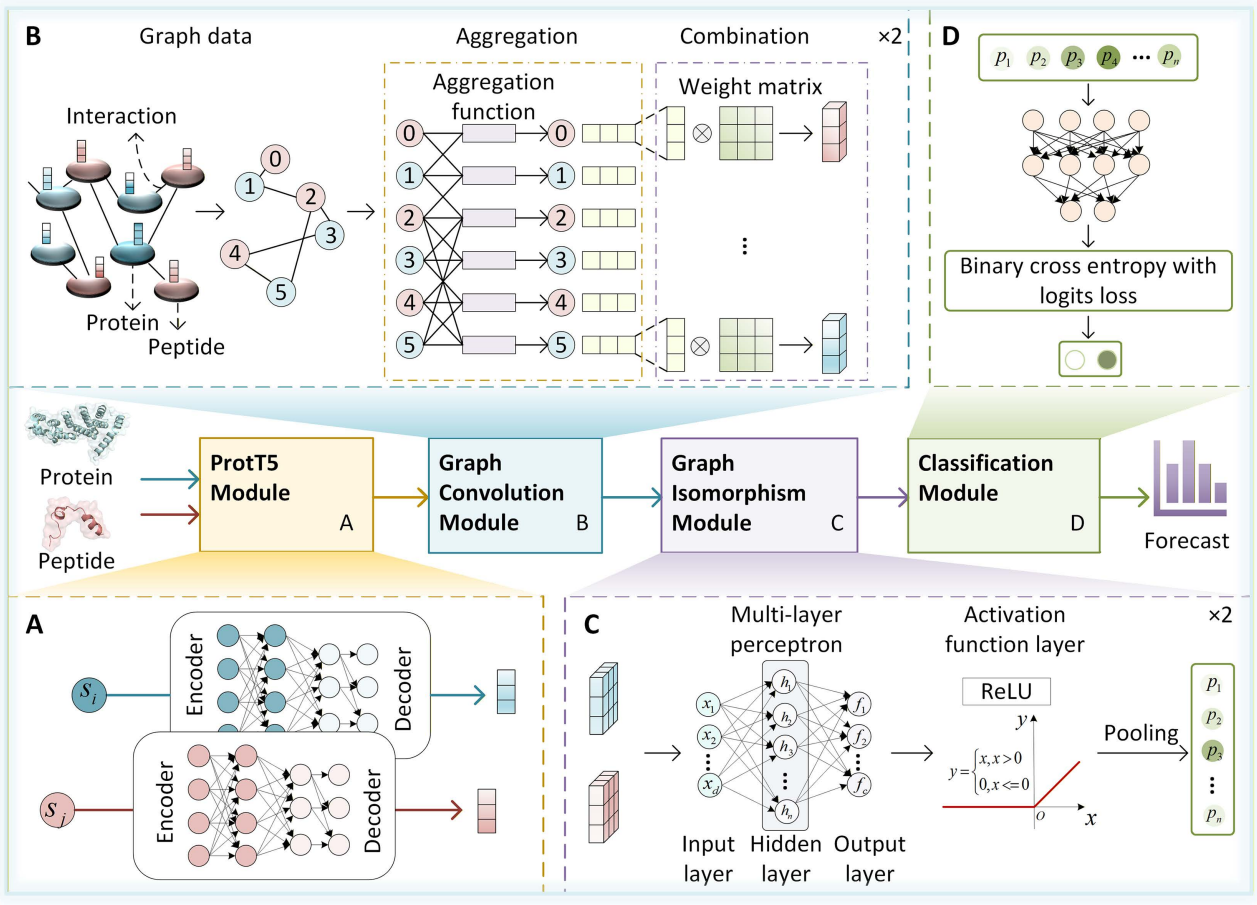

Fig 5. The framework of PepLM-GNN, comprising four modules: ProtT5, graph convolution, graph isomorphism, and classification.

https://doi.org/10.1371/journal.pcbi.1014084.g005

sequence of length L is input into ProtT5, a residue-level feature matrix with dimensions of $L \times 1024$ can be automatically generated. Then, we applied an average pooling operation to these residue-level features to produce a fixed-dimensional 1024-dimensional global representation vector. The calculation formula [60] is:

$$\boldsymbol{v}_{\text{seq}} = \frac{1}{L}\sum_{i=1}^{L}\boldsymbol{v}_i \tag{1}$$

where $\mathbf{v}_i$ is the ProtT5 feature representation. Then the $\mathbf{v}_{\text{seq}}$ was standardised by Z-score normalisation [61] to eliminate the influence of sequence length differences on subsequent modelling, while preserving the overall distribution of features.

During the model construction stage, the feature vectors generated by the pre-trained ProtT5 model are directly assigned as node features in the PepLM-GNN. The semantic dependencies of the sequence are captured through the cross-modal coding mechanism to provide feature input for PepPI prediction.

4.2.2 Hybrid graph network framework. This study designs a hybrid graph network framework that integrates the GCN and GIN to realise collaborative learning. The framework effectively captures the non-Euclidean data characteristics of PepPI, achieving more accurate and robust prediction performance.

The hybrid graph network framework constructs a global peptide-protein interaction network, with node 0 representing the peptide sequence and node 1 representing the protein sequence. The node features adopt the sequence-level embedding vectors extracted by the pre-trained ProtT5 model, which can avoid the defects of traditional manual features, such as "one-sided information and limited dimensions". The edge structure is defined as a bidirectional edge index [[0,1],[1,0]], explicitly simulating the potential interaction between peptides and proteins and avoiding the ambiguity of feature transfer caused by "implicit interaction". The feature matrix is formed by concatenating the ProtT5 embeddings of peptides and proteins, ensuring that each node initially carries complete sequence semantic information of its original sequence.

To accurately extract feature associations between nodes and their direct neighbors, and to meet the micro-modelling requirements of non-Euclidean data in PepPI, we design a GCN local aggregation module. Local neighbourhood feature learning is achieved through two layers of Graph Convolutional Neural Networks (GCNConv) [62]. The first layer maps the input features to a 64-dimensional hidden layer. Instance normalisation (IN) is employed to standardise feature distribution [63], the ReLU activation function introduces non-linearity [64], and a 20% dropout rate suppresses overfitting to avoid micro-interaction misjudgment caused by sample bias [65]. The second layer performs further neighbourhood aggregation and outputs 2×64-dimensional node-level local features. We strengthen the microscopic interaction between peptides and proteins, providing refined support for subsequent global modelling. The formula is as follows [66]:

$$h^{(l+1)} = \text{ReLU}\left(\text{InstanceNorm}\left(\hat{\boldsymbol{D}}^{-1/2}\hat{\boldsymbol{A}}\hat{\boldsymbol{D}}^{-1/2}h^{(l)}\boldsymbol{W}^{(l)}\right)\right) \tag{2}$$

where $\hat{\mathbf{A}}$ is the adjacency matrix with self-loops, $\hat{\mathbf{D}}$ is the corresponding degree matrix, and $\mathbf{W}^{(l)}$ is the weight matrix for layer l.

Furthermore, the proposed method introduces a GIN-based global module to form a hybrid collaboration with GCN. This collaboration effectively mitigates issues of gradient dispersion and over-smoothing in deep GCN layers, thereby preventing the undesirable convergence of node features. In the first Graph Isomorphism Network Convolution (GINConv) layer, a multi-layer perceptron (MLP) serves to non-linearly reconstruct local features, thereby alleviating over-smoothing and enhancing the feature discriminative. In the second GINConv layer, we aggregate node features into 32-dimensional graph-level representations by using the Global Add Pooling [67]. This process captures the global topology of PepPI and significantly improves robustness in cold-start test data. The formula [68] is as follows:

$$h_v^{(l+1)} = f_\Theta\left((1+\epsilon)h_v^{(l)} + \sum_{u\in\mathcal{N}(v)} h_u^{(l)}\right) \tag{3}$$

where f_Θ is the MLP mapping function.

4.2.3 Classification module. Based on the 32-dimensional graph-level representation output by the hybrid graph network framework, this study designs a classification module to achieve binary classification prediction of PepPI. We apply a 20% dropout rate to the graph-level representation to further reduce the model's excessive reliance on the training data and ensure the robustness of the prediction in the cold-start test dataset. The processed features are mapped to one-dimensional logits through a linear transformation layer to reduce parameter redundancy. Use the sigmoid activation function [69] to compress logits to the interval [0,1] to obtain the interaction probability. The model formula is as follows:

$$\text{PPIModel} = \text{FC}\left(\text{Dropout}\left(\text{GIN}\left(\text{GCN}(x, E)\right)\right)\right) \tag{4}$$

where E is the edge index, and FC is the linear transformation layer.

In designing the loss function, we deeply integrate the sigmoid activation function into the calculation process of binary cross-entropy (BCE) loss [70]. It can align the prediction results with the probability attribute of PepPI interactions, thereby improving the model's accuracy in distinguishing between interacting and non-interacting peptide-protein pairs. The formula [36] is as follows:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i log\hat{y}_i + (1-y_i)log(1-\hat{y}_i)\right) \tag{5}$$

where $\hat{y}_i$ is the predicted probability of the model, and y_i is the true label value.

Therefore, the model PepLM-GNN we use constructs a hybrid graph neural network framework that adapts to the non-Euclidean data features of PepPI, efficiently collaborates with local and global information, and enhances the generalisation ability in cold start scenarios.

4.3 Training and optimization of the model

In this study, we implement the PepLM-GNN model using PyTorch [71]. The training process uses the Adam optimiser [72] to optimize the parameters, with a learning rate of 1e-3 and a weight decay of 5e-4. A fixed random seed of 1234 was used throughout the experiments to ensure reproducibility. To avoid overfitting of the network, the Dropout algorithm [73] is introduced during training. To control the training process, we employ the Early Stopping strategy [74], which is regulated by monitoring the model's performance on the validation set, with a patience value of 10. In the model evaluation stage, a dynamic threshold optimization strategy [75] is adopted, and the threshold that maximises the F1 score is selected as the optimal decision boundary of the method. We select BCEWithLogitsLoss as the loss calculation function of the model [70].

4.4 Algorithm performance evaluation criteria

In this study, to comprehensively evaluate the model's performance in the PepPI prediction task, we employed multiple evaluation metrics that reflect the model's performance across different dimensions, including AUC, AUPR, Accuracy (ACC), and F1 score [76,77]. The calculation formulas for the relevant evaluation indicators are as follows [36]:

$$\begin{cases} \text{ACC} & = \frac{TP+TN}{TP+TN+FP+FN} \\ \text{F1} & = \frac{2\times P\times R}{P+R} \end{cases} \tag{6}$$

where TP represents the number of actual positive sequences, TN represents the number of true negative sequences, FP represents the number of false positive sequences, and FN represents the number of false negative sequences. Additionally, P denotes precision (the proportion of correctly predicted positive sequences among all predicted positive sequences) and R denotes recall (the proportion of correctly predicted positive sequences among all actual positive sequences). AUC is the area under the Receiver Operating Characteristic (ROC) curve, which plots the actual positive rate against the false positive rate at different classification thresholds. AUPR is the area under the Precision-Recall curve, which illustrates the relationship between precision and recall under varying threshold settings.

Supporting information

S1 Table. Performance comparison of PepLM-GNN with other baseline methods based on five-fold cross-validation.
(DOCX)

Author contributions

Conceptualization: Ke Yan, Meijing Li, Shutao Chen, Tianyi Liu, Zhen Li.

Funding acquisition: Ke Yan, Bin Liu.

Methodology: Ke Yan, Meijing Li, Shutao Chen, Jing Hao.

Project administration: Bin Liu, Zhen Li.

Software: Meijing Li, Shutao Chen.

Writing – original draft: Ke Yan, Meijing Li, Shutao Chen, Tianyi Liu, Bin Liu.

Writing – review & editing: Ke Yan, Meijing Li, Jing Hao, Bin Liu, Zhen Li.

References

1. Wang Z, Meng J, Dai Q, Li H, Xia S, Yang R, et al. DeepPepPI: a deep cross-dependent framework with information sharing mechanism for predicting plant peptide-protein interactions. Expert Syst Appl. 2024;252:124168. https://doi.org/10.1016/j.eswa.2024.124168
2. Shahid M. Understanding protein-protein interactions: techniques and applications in drug development. Multidiscip J Biochem Technol. 2024;1(2):17–25.
3. Lai L, Liu Y, Song B, Li K, Zeng X. Deep generative models for therapeutic peptide discovery: a comprehensive review. ACM Comput Surv. 2025;57(6):1–29. https://doi.org/10.1145/3714455
4. Li ZL, Feng YH, Jiao J, Ju XY, Yu L, Zhang GL, et al. Why insulin aspart and insulin degludec exhibit distinct release mechanisms. AIChE Journal. 2024;71(1). https://doi.org/10.1002/aic.18609
5. Perrin P, Jongsma ML, Neefjes J, Berlin I. The labyrinth unfolds: architectural rearrangements of the endolysosomal system in antigen-presenting cells. Curr Opin Immunol. 2019;58:1–8. https://doi.org/10.1016/j.coi.2018.12.004 PMID: 30738283
6. Luo J, Zhao K, Chen J, Yang C, Qu F, Liu Y, et al. iMFP-LG: identify novel multi-functional peptides using protein language models and graph-based deep learning. Genomics Proteomics Bioinform. 2025;22(6):qzae084. https://doi.org/10.1093/gpbjnl/qzae084 PMID: 39585308
7. Liu X, Luo J, Wang X, Zhang Y, Chen J. Directed evolution of antimicrobial peptides using multi-objective zeroth-order optimization. Brief Bioinform. 2024;26(1):bbae715. https://doi.org/10.1093/bib/bbae715 PMID: 39800873
8. Ye JH, Li A, Zheng H, Yang BH, Lu YM. Machine learning advances in predicting peptide/protein-protein interactions based on sequence information for lead peptides discovery. Adv Biol. 2023;7(6):2200232.
9. Qi R, Liu S, Hui X, Shaytan AK, Liu B. AI in drug development: advances in response, combination therapy, repositioning, and molecular design. Sci China Inf Sci. 2025;68(7). https://doi.org/10.1007/s11432-024-4461-0
10. Li T, Ren X, Luo X, Wang Z, Li Z, Luo X, et al. A foundation model identifies broad-spectrum antimicrobial peptides against drug-resistant bacterial infection. Nat Commun. 2024;15(1):7538. https://doi.org/10.1038/s41467-024-51933-2 PMID: 39214978
11. Qiao J, Jin J, Wang D, Teng S, Zhang J, Yang X, et al. A self-conformation-aware pre-training framework for molecular property prediction with substructure interpretability. Nat Commun. 2025;16(1):4382. https://doi.org/10.1038/s41467-025-59634-0 PMID: 40355450
12. Mahapatra M, Sahu C, Mohapatra S. Trends of Artificial Intelligence (AI) use in drug targets, discovery and development: current status and future perspectives. Current Drug Targets. 2025;26(4):221–42.

13. Ai C, Yang H, Liu X, Dong R, Ding Y, Guo F. MTMol-GPT: de novo multi-target molecular generation with transformer-based generative adversarial imitation learning. PLoS Comput Biol. 2024;20(6):e1012229. https://doi.org/10.1371/journal.pcbi.1012229 PMID: 38924082
14. Zhang D, Qi R, Lan X, Liu B. A novel multislice framework for precision 3D spatial domain reconstruction and disease pathology analysis. Genome Res. 2025;35(8):1794–808. https://doi.org/10.1101/gr.280281.124 PMID: 40659497
15. Kim J, Park HS, Kang H, Son CY, Kim HK, Choi JH, et al. In silico fragment-based peptide design targeting undruggable proteins for enhanced detection of circulating tumor cells. Chem Eng J. 2025;522:167447. https://doi.org/10.1016/j.cej.2025.167447
16. Sigurdsson EM. Alzheimer's therapy development: a few points to consider. Prog Mol Biol Transl Sci. 2019;168:205–17. https://doi.org/10.1016/bs.pmbts.2019.06.001 PMID: 31699315
17. Yan K, Lv H, Shao J, Chen S, Liu B. TPpred-SC: multi-functional therapeutic peptide prediction based on multi-label supervised contrastive learning. Sci China Inf Sci. 2024;67(11). https://doi.org/10.1007/s11432-024-4147-8
18. Yan K, Chen S, Liu B, Wu H. Accurate prediction of toxicity peptide and its function using multi-view tensor learning and latent semantic learning framework. Bioinformatics. 2025;41(9):btaf489. https://doi.org/10.1093/bioinformatics/btaf489 PMID: 40905623
19. Jiang Y, Wang R, Feng J, Jin J, Liang S, Li Z, et al. Explainable deep hypergraph learning modeling the peptide secondary structure prediction. Adv Sci (Weinh). 2023;10(11):e2206151. https://doi.org/10.1002/advs.202206151 PMID: 36794291
20. Zhao J, Yan W, Yang Y. DeepTP: A Deep Learning Model for Thermophilic Protein Prediction. Int J Mol Sci. 2023;24(3):2217. https://doi.org/10.3390/ijms24032217 PMID: 36768540
21. Shafiee S, Fathi A, Taherzadeh G. Protein-peptide interaction region residues prediction using a generative sampling technique and ensemble deep learning-based models. Appl Soft Comput. 2025;182:113603. https://doi.org/10.1016/j.asoc.2025.113603
22. Hu J, Chen K-X, Rao B, Ni J-Y, Thafar MA, Albaradei S, et al. Protein-peptide binding residue prediction based on protein language models and cross-attention mechanism. Anal Biochem. 2024;694:115637. https://doi.org/10.1016/j.ab.2024.115637 PMID: 39121938
23. Luo YD, Cai JX. Deep learning for the prediction of protein sequence, structure, function, and interaction: applications, challenges, and future directions. Current Proteom. 2024;21(6):561–79.
24. Zhang W, Wei H, Zhang W, Wu H, Liu B. Multiple types of disease-associated RNAs identification for disease prognosis and therapy using heterogeneous graph learning. Sci China Inf Sci. 2024;67(8). https://doi.org/10.1007/s11432-024-4100-7
25. Shao J, Chen J, Liu B. ProFun-SOM: protein function prediction for specific ontology based on multiple sequence alignment reconstruction. IEEE Trans Neural Netw Learn Syst. 2025;36(5):8060–71. https://doi.org/10.1109/TNNLS.2024.3419250 PMID: 38980781
26. Huang Z, Xiao Z, Ao C, Guan L, Yu L. Computational approaches for predicting drug-disease associations: a comprehensive review. Front Comput Sci. 2024;19(5). https://doi.org/10.1007/s11704-024-40072-y
27. Huang Z, Guo X, Qin J, Gao L, Ju F, Zhao C, et al. Accurate RNA velocity estimation based on multibatch network reveals complex lineage in batch scRNA-seq data. BMC Biol. 2024;22(1):290. https://doi.org/10.1186/s12915-024-02085-8 PMID: 39696422
28. Guo X, Huang Z, Ju F, Zhao C, Yu L. Highly Accurate Estimation of Cell Type Abundance in Bulk Tissues Based on Single-Cell Reference and Domain Adaptive Matching. Adv Sci (Weinh). 2024;11(7):e2306329. https://doi.org/10.1002/advs.202306329 PMID: 38072669
29. Pathak D, Narzary S, Nandi S, Som B. Part-of-speech tagger for Bodo language using deep learning approach. Nat lang process. 2024;31(2):215–29. https://doi.org/10.1017/nlp.2024.15
30. Zhang H-Q, Arif M, Thafar MA, Albaradei S, Cai P, Zhang Y, et al. PMPred-AE: a computational model for the detection and interpretation of pathological myopia based on artificial intelligence. Front Med (Lausanne). 2025;12:1529335. https://doi.org/10.3389/fmed.2025.1529335 PMID: 40182849
31. Jin X, Chen Z, Yu D, Jiang Q, Chen Z, Yan B, et al. TPepPro: a deep learning model for predicting peptide-protein interactions. Bioinformatics. 2024;41(1):btae708. https://doi.org/10.1093/bioinformatics/btae708 PMID: 39585721
32. Boadu F, Lee A, Cheng J. Deep learning methods for protein function prediction. Proteomics. 2025;25(1–2):e2300471. https://doi.org/10.1002/pmic.202300471 PMID: 38996351
33. Ma T, Chen Y, Tao W, Zheng D, Lin X, Pang PC-I, et al. Learning to denoise biomedical knowledge graph for robust molecular interaction prediction. IEEE Trans Knowl Data Eng. 2024;36(12):8682–94. https://doi.org/10.1109/tkde.2024.3471508
34. Wei L, Xing P, Zeng J, Chen J, Su R, Guo F. Improved prediction of protein-protein interactions using novel negative samples, features, and an ensemble classifier. Artif Intell Med. 2017;83:67–74. https://doi.org/10.1016/j.artmed.2017.03.001 PMID: 28320624
35. Lei Y, Li S, Liu Z, Wan F, Tian T, Li S, et al. A deep-learning framework for multi-level peptide-protein interaction prediction. Nat Commun. 2021;12(1):5465. https://doi.org/10.1038/s41467-021-25772-4 PMID: 34526500
36. Chen S, Yan K, Li X, Liu B. Protein language pragmatic analysis and progressive transfer learning for profiling peptide-protein interactions. IEEE Trans Neural Netw Learn Syst. 2025;36(8):15385–99. https://doi.org/10.1109/TNNLS.2025.3540291 PMID: 40100664
37. Gao Z, Jiang C, Zhang J, Jiang X, Li L, Zhao P, et al. Hierarchical graph learning for protein-protein interaction. Nat Commun. 2023;14(1):1093. https://doi.org/10.1038/s41467-023-36736-1 PMID: 36841846
38. Bai P, Miljković F, John B, Lu H. Interpretable bilinear attention network with domain adaptation improves drug–target prediction. Nat Mach Intell. 2023;5(2):126–36. https://doi.org/10.1038/s42256-022-00605-1

39. Mentari O, Shujaat M, Tayara H, Chong KT. Toxicity prediction for immune thrombocytopenia caused by drugs based on logistic regression with feature importance. Curr Bioinform. 2024;19(7):641–50. https://doi.org/10.2174/0115748936269606231001140647
40. Ru X, Li L, Zou Q. Incorporating distance-based top-n-gram and random forest to identify electron transport proteins. J Proteome Res. 2019;18(7):2931–9. https://doi.org/10.1021/acs.jproteome.9b00250 PMID: 31136183
41. Wang Y, Zhai Y, Ding Y, Zou Q. SBSM-Pro: support bio-sequence machine for proteins. Sci China Inf Sci. 2024;67(11). https://doi.org/10.1007/s11432-024-4171-9
42. Kumar Meher P, Hati S, Sahu TK, Pradhan U, Gupta A, Rath SN. SVM-Root: identification of root-associated proteins in plants by employing the support vector machine with sequence-derived features. Curr Bioinform. 2024;19(1):91–102. https://doi.org/10.2174/1574893618666230417104543
43. 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
44. 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
45. Xu X, Bonvin AMJJ. DeepRank-GNN-esm: a graph neural network for scoring protein-protein models using protein language model. Bioinform Adv. 2024;4(1):vbad191. https://doi.org/10.1093/bioadv/vbad191 PMID: 38213822
46. Jiang L, Yang X, Guo X, Li D, Li J, Wuchty S, et al. Graph neural network integrated with pretrained protein language model for predicting human-virus protein-protein interactions. Brief Bioinform. 2025;26(5):bbaf461. https://doi.org/10.1093/bib/bbaf461 PMID: 40919914
47. Luo Y, Shi L, Li Y, Zhuang A, Gong Y, Liu L, et al. From intention to implementation: automating biomedical research via LLMs. Sci China Inf Sci. 2025;68(7). https://doi.org/10.1007/s11432-024-4485-0
48. Rao R, Bhattacharya N, Thomas N, Duan Y, Chen X, Canny J, et al. Evaluating protein transfer learning with TAPE. Adv Neural Inf Process Syst. 2019;32:9689–701. PMID: 33390682
49. 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
50. 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
51. Yan K, Yu H, Chen S, Shaytan AK, Liu B, Wang Y. DSCA-HLAII: A dual-stream cross-attention model for predicting peptide-HLA class II interaction and presentation. PLoS Comput Biol. 2026;22(1):e1013836. https://doi.org/10.1371/journal.pcbi.1013836 PMID: 41481588
52. Liu BH, Jobichen C, Chia CSB, Chan THM, Tang JP, Chung TXY, et al. Targeting cancer addiction for SALL4 by shifting its transcriptome with a pharmacologic peptide. Proc Natl Acad Sci U S A. 2018;115(30):E7119–28. https://doi.org/10.1073/pnas.1801253115 PMID: 29976840
53. Ying R, Bourgeois D, You J, Zitnik M, Leskovec J. GNNExplainer: Generating Explanations for Graph Neural Networks. Adv Neural Inf Process Syst. 2019;32:9240–51. PMID: 32265580
54. Piehl DW, Burley SK. Exploring experimentally determined structures and computed structure models from artificial intelligence/machine learning at RCSB Protein Data Bank (RCSB PDB, RCSB.org). Acta Crystallogr A Found Adv. 2023;79(a2):C397–C397. https://doi.org/10.1107/s2053273323092185
55. Hauser AS, Windshügel B. LEADS-PEP: a benchmark data set for assessment of peptide docking performance. J Chem Inf Model. 2016;56(1):188–200. https://doi.org/10.1021/acs.jcim.5b00234 PMID: 26651532
56. Abdin O, Nim S, Wen H, Kim PM. PepNN: a deep attention model for the identification of peptide binding sites. Commun Biol. 2022;5(1):503. https://doi.org/10.1038/s42003-022-03445-2 PMID: 35618814
57. Johansson-Åkhe I, Mirabello C, Wallner B. Predicting protein-peptide interaction sites using distant protein complexes as structural templates. Scientific Reports. 2019;9(1):4267.
58. Chen S, Yan K, Liu B. PDB-BRE: a ligand-protein interaction binding residue extractor based on protein data bank. Proteins. 2024;92(1):145–53. https://doi.org/10.1002/prot.26596 PMID: 37750380
59. Zhang J, Zeng H, Chen J, Zhu Z. INAB: identify nucleic acid binding domain via cross-modal protein language models and multiscale computation. Brief Bioinform. 2025;26(5):bbaf509. https://doi.org/10.1093/bib/bbaf509 PMID: 41020638
60. NaderiAlizadeh N, Singh R. Aggregating residue-level protein language model embeddings with optimal transport. Bioinform Adv. 2025;5(1):vbaf060. https://doi.org/10.1093/bioadv/vbaf060 PMID: 40170888
61. Mahdi WA, Alhowyan A, Obaidullah AJ. Estimation and validation of solubility of recombinant protein in E. coli strains via various advanced machine learning models. Sci Rep. 2025;15(1):12784. https://doi.org/10.1038/s41598-025-97445-x PMID: 40229419
62. Li G, Muller M, Qian G, Delgadillo IC, Abualshour A, Thabet A, et al. DeepGCNs: making GCNs Go as deep as CNNs. IEEE Trans Pattern Anal Mach Intell. 2023;45(6):6923–39. https://doi.org/10.1109/TPAMI.2021.3074057 PMID: 33872143
63. Zhu R, Gao H-H, Wang Y. Joint coordinate attention mechanism and instance normalization for COVID online comments text classification. PeerJ Comput Sci. 2024;10:e2240. https://doi.org/10.7717/peerj-cs.2240 PMID: 39314739
64. Ema RR, Khatun MtA, Adnan MdN, Kabir SkS, Galib SMd, Hossain MdA. Protein Secondary Structure Prediction based on CNN and Machine Learning Algorithms. IJACSA. 2022;13(11). https://doi.org/10.14569/ijacsa.2022.0131108
65. Liu L, Luo YH, Shen X, Sun MZ, Li B. β-Dropout: a unified dropout. IEEE Access. 2019;7:36140–53.

66. 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(1):btac715. https://doi.org/10.1093/bioinformatics/btac715 PMID: 36342186

67. Shen H, Zhang Y, Zheng C, Wang B, Chen P. A cascade graph convolutional network for predicting protein-ligand binding affinity. Int J Mol Sci. 2021;22(8):4023. https://doi.org/10.3390/ijms22084023 PMID: 33919681

68. Wang X, Han T, Feng R, Xia Z, Wang H, Yu W, et al. GTE-PPIS: a protein-protein interaction site predictor based on graph transformer and equivariant graph neural network. Brief Bioinform. 2025;26(3):bbaf290. https://doi.org/10.1093/bib/bbaf290 PMID: 40524427

69. Kumar A, Singh Sodhi S. Classification of data on stacked autoencoder using modified sigmoid activation function. IFS. 2023;44(1):1–18. https://doi.org/10.3233/jifs-212873

70. Si Y, Yan C. Improved protein contact prediction using dimensional hybrid residual networks and singularity enhanced loss function. Brief Bioinform. 2021;22(6):bbab341. https://doi.org/10.1093/bib/bbab341 PMID: 34448830

71. Hu Q, Wang Z, Meng J, Li W, Guo J, Mu Y, et al. OpenDock: a pytorch-based open-source framework for protein-ligand docking and modelling. Bioinformatics. 2024;40(11):btae628. https://doi.org/10.1093/bioinformatics/btae628 PMID: 39432683

72. Kapitan V, Choi M. Adaptive gradient scaling: integrating Adam and landscape modification for protein structure prediction. BMC Bioinformatics. 2025;26(1):161. https://doi.org/10.1186/s12859-025-06185-2 PMID: 40597628

73. Shen X, Tian XM, Liu TL, Xu F, Tao DC. Continuous Dropout. IEEE Transact Neural Network Learn Syst. 2018;29(9):3926–37.

74. Paguada S, Batina L, Buhan I, Armendariz I. Being patient and persistent: optimizing an early stopping strategy for deep learning in profiled attacks. IEEE Trans Comput. 2025;74(3):875–86. https://doi.org/10.1109/tc.2023.3234205

75. Chen L, Wang Z, Wu J, Guo Y, Li F, Li Z. Dynamic threshold strategy optimization for security protection in Internet of Things: An adversarial deep learning-based game-theoretical approach. Concurr Comput. 2022;35(20). https://doi.org/10.1002/cpe.6944

76. Zou X, Ren L, Cai P, Zhang Y, Ding H, Deng K, et al. Accurately identifying hemagglutinin using sequence information and machine learning methods. Front Med (Lausanne). 2023;10:1281880. https://doi.org/10.3389/fmed.2023.1281880 PMID: 38020152

77. Xie X, Changchun W, Fuying D, Kejun D, Dan Y, Jian H. scRiskCell: a single-cell framework for quantifying pancreatic islet risk cells and unravelling their dynamic transcriptional and molecular adaptation in the progression of type 2 diabetes. iMeta. 2025;2025:e70060.