A sequence-based deep learning framework (PepInter) for protein–peptide interaction representation learning with pretrained protein language models

Xinke Zhan[1], Silong Zhai[1], Tiantao Liu[1], Shaolong Lin[1], Tao Bi[2], Bingwen Zhu[3], Shirley W.I. Siu[1*]

[1] Centre for Artificial Intelligence Driven Drug Discovery, Faculty of Applied Sciences, Macao Polytechnic University, Macau SAR, China

[2] Drug Research Center of Integrated Traditional Chinese and Western Medicine, Southwest Medical University, Sichuan, China

[3] Institute of Infection and Immunity, Henan Academy of Innovations in Medical Science, Zhengzhou, China

Corresponding author: shirleysiu@mpu.edu.mo

Content

S1 The parameter of the Rosetta Peptiderive protocol

Table S1 Hyperparameter setting in the Rosetta Peptiderive protocol.

Option name	Type	Setting
pep_lengths	List of numbers	**5-50**
skip_zero_isc	True/False	True
dump_peptide_pose	True/False	True
dump_cyclic_poses	True/False	False
dump_prepared_pose	True/False	True
dump_report_file	True/False	True
restrict_receptors_to_chains	List of characters	-
restrict_partners_to_chains	List of characters	-
do_minimize	True/False	True
optimize_cyclic_threshold	Real number	0.35
report_format	String	Markdown

S2 Pretraining Loss

Figure S2 shows the training loss curve during the pre-training stage of PepInter on the ProtFragDB dataset. The training loss decreases rapidly in the early phase and gradually converges as training proceeds, indicating stable optimization and effective learning of protein–peptide interaction representations. The smooth convergence behavior suggests that the large-scale pseudo protein–peptide data provide informative training signals for pre-training the model.

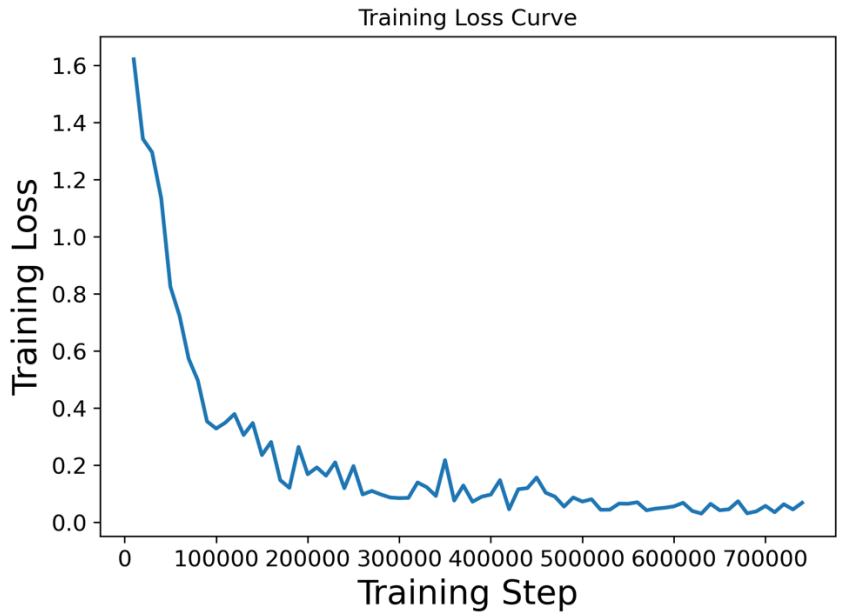

Fig. S1 Pre-training loss curve of PepInter on ProtFragDB.

S3 Baseline Methods

In order to better evaluate our proposed method, we make a comparison with several following competitive methods for predicting PpI and affinity. To ensure the fairness of comparison, all the baseline models are trained and evaluated on the same datasets and most hyper-parameters are set based on the original works.

The comparison models for predicting PpI are list as follows:

- **XGBoost** [1]: XGBoost was an efficient and powerful ensemble decision tree model used to learn patterns from features and make predictions. For the XGBoost baseline, protein and peptide sequences were first encoded using the pretrained ESM2 model, and the resulting embeddings were concatenated to form joint representations, which were then used as input to an XGBoost model with hyperparameters optimized using Optuna [] for binary classification.
- **MAARDTI** [2]: MAARDTI is used to predict drug-target interaction. It utilizes a multi-layer CNN block to learn low-dimensional representation features. The multi-perspective attention aggregating module and bi-contextual refocusing module are adopted to learn the interaction relationship between proteins and drugs, the combined features are finally fed into interaction prediction block. In this study, drug embedding is replaced with a peptide embedding while keeping the rest of the framework unchanged.
- **Rep-ConvDTI** [3]: Rep-ConvDTI is a deep learning framework for drug-target interaction prediction that models large-scale sequence patterns using large-kernel convolutions, while capturing fine-grained information through reparameterization and a gated attention mechanism to better characterize drug–target interactions. In this study, we modify the target encoding to peptide encoding, without altering the overall framework of Rep-ConvDTI.
- **HyperAttentionDTI** [4]: HyperAttentionDTI utilizes the attention mechanism to capture attention vector of each amino acid and atom. This model use CNN block to extract protein and drug features, and mixed original feature and HyperAttention matrix. The modified drug-protein pair feature vector is fed into FCN for final predictions. In this model, we replace the drug encoding with a peptide encoding, while keeping all other hyperparameter settings consistent with the original work.
- **PIPR** [5]: PIPR is an end-to-end sequence-based protein-protein interaction prediction model that uses a Siamese residual RCNN to automatically learn both local and contextual features from protein sequences, achieving strong performance without manual feature engineering.

The comparison models for predicting affinity are list as follows:

- **XGBoost:** For the XGBoost baseline, protein and peptide sequences were first encoded

using the pretrained ESM2 model, and the resulting embeddings were concatenated to form joint representations. For affinity prediction, the model was adapted to a regression setting by modifying only the output layer, while keeping all other components unchanged, with hyperparameters optimized using Optuna.

- **DeepDTA** [6]: DeepDTA utilizes two parallel 3-layer CNNs and fully connected layers for predicting DTI. It originally extracts drug and protein feature maps by CNN block and then drug and protein feature maps are concatenated. Finally, the model obtained prediction result by fed drug-target pairs into fully connected layers.
- **TEFDTA** [7]: TEFDTA is an attention-based framework for drug-target binding affinity prediction that integrates a Transformer encoder with molecular fingerprints to model both non-covalent and covalent interactions. By employing distinct representations for proteins and drugs and leveraging fine-tuning from large non-covalent datasets to smaller covalent datasets, TEFDTA effectively improves affinity prediction accuracy across different interaction types.
- **AttentionDTA** [8]: AttentionDTA is an end-to-end deep learning framework for drug-target binding affinity prediction that integrates attention mechanisms with one-dimensional convolutional neural networks to jointly model drug and protein sequences. By explicitly capturing the mutual importance of subsequences between drugs and targets, AttentionDTA learns more informative representations and improves affinity prediction performance.

For the fairness of the experiment, all hyperparameters are kept consistent with those used in the original work.

S4 Cluster-based cross-validation

Due to a peptide can likewise bind to multiple similar proteins and a protein can also interact with multiple similar peptides, this pattern often contains substantial redundancy. To eliminate the potential bias introduced by such redundancy and to ensure a fair comparison, we followed the same strategy in CAMP [9] and MONN [10], which employed the cluster-based cross-validation settings to evaluate the model performance. The clustering threshold means the minimal distance between any two clusters by a single-linkage clustering algorithm [11]. More specifically, the distance between two peptides (proteins) p_i and p_j is defined as:

$$\mathrm{d_{ij}} = 1 - \frac{\mathrm{SW}(\mathrm{p_i}, \mathrm{p_j})}{\sqrt{\mathrm{SW}(\mathrm{p_i}, \mathrm{p_i})\mathrm{SW}(\mathrm{p_j}, \mathrm{p_j})}} \tag{1}$$

where $SW(\cdot)$ means the Smith-Waterman alignment score (https://github.com/mengyao/Complete-Striped-Smith-Waterman-Library) between two sequences.

We selected similarity thresholds from 0.3 to 0.6 for cluster-based data splitting because this range provides a reasonable balance between overly stringent and overly permissive clustering. As

shown in Tables S4.1 and S4.2, thresholds of 0.1 and 0.2 produced a large number of small clusters, resulting in excessively fragmented partitions, whereas thresholds of 0.7 to 0.9 led to severe cluster collapse, with the largest cluster covering nearly all sequences. Therefore, thresholds of 0.3, 0.4, 0.5, and 0.6 were used to represent different levels of sequence novelty while maintaining meaningful and feasible cluster partitions.

Table S2 Clustering results of peptides under different similarity thresholds.

Threshold	Number of clusters	max cluster size
0.1	14,193	83
0.2	13,016	180
0.3	12,062	188
0.4	10,246	2,939
0.5	6,543	7,885
0.6	2,273	12,797
0.7	37	15,190
0.8	35	15,192
0.9	35	15,192

Table S3 Clustering results of proteins under different similarity thresholds.

Threshold	Number of clusters	max cluster size
0.1	11,683	73
0.2	10,732	153
0.3	9,976	159
0.4	8,619	2,229
0.5	5,594	6,316
0.6	2,047	10,330
0.7	38	12,466
0.8	35	12,469
0.9	35	12,469

For the “Novel Protein” and “Novel Peptide” settings, we performed five-fold cross-validation at the cluster level rather than directly splitting individual protein or peptide sequences. Specifically, protein clusters or peptide clusters were partitioned into five folds, and one-fold was used as the validation set in each round, while the remaining folds were used for training. Because the number of samples within each cluster was not evenly distributed, the validation set accounted for approximately, but not exactly, 20% of the data. For the “Novel Pair” setting, cross-validation was performed jointly based on protein and peptide clusters. Specifically, protein clusters were first divided into three groups, and peptide clusters were also divided into three groups. The combination of protein-cluster groups and peptide-cluster groups resulted in a 3x3 grid, yielding nine non-

overlapping protein-peptide cluster blocks. In each round, one grid block was selected as the validation set. The training set was constructed using only the grid blocks that did not share either protein clusters or peptide clusters with the validation block, thereby ensuring that no protein or peptide cluster was shared between the training and validation sets. This procedure resulted in a nine-fold cross-validation scheme for the “Novel Pair” setting.

S5 Comparative performance of PepInter against baseline methods under different clustering settings

In Supplmentary Data.xlsx (Fig 2)

S6 Evaluation of Generalization Ability on Imbalanced Datasets

In Supplmentary Data.xlsx (Fig4-A,B)

S7 UMAP Visualization of Semaglutide Analog–GLP-1R Interaction EmbeddingsU

MAP visualization was performed to further examine the embedding distribution learned by PepInter in the high-dimensional PpI space. As shown in Figure S3, all candidate protein–peptide pairs were projected into a two-dimensional embedding space. The seven semaglutide-analog peptide–GLP-1R pairs (red) formed a compact cluster that was spatially separated from the negative pairs (blue), indicating that PepInter generated distinguishable representations for interacting and non-interacting candidates.

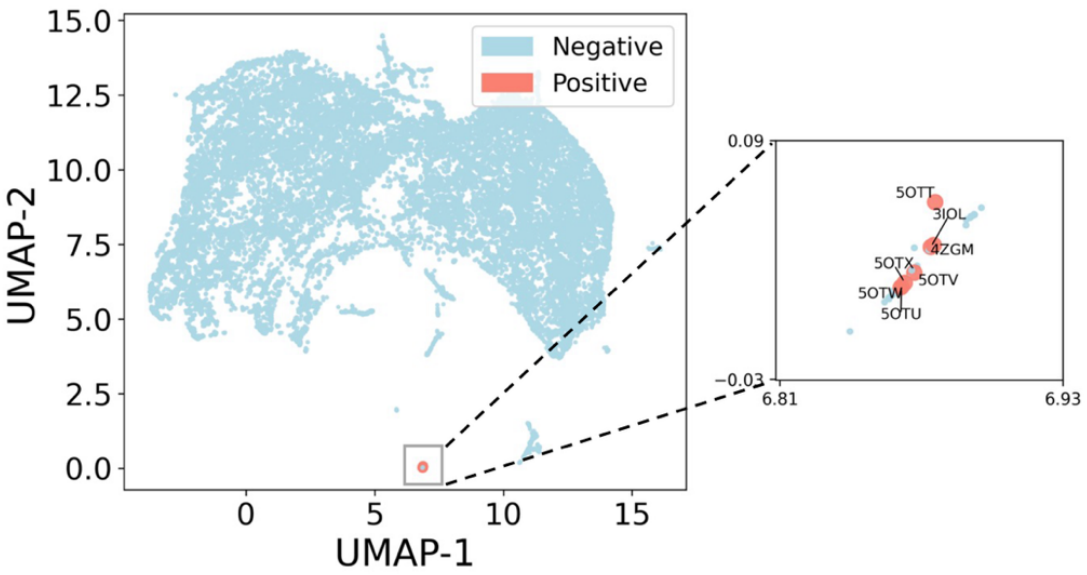

Fig. S2. UMAP visualization of predicted interactions between Semaglutide-analogous peptides and GLP-1R. (Light blue and salmon points indicate negative and positive samples, respectively.)

S7 Case-level comparison with HPEPDOCK under blind and known binding-site docking scenarios

Table S6 and S7 report the HPEPDOCK docking scores for the Case 1 and 2 protein-peptide system. Positive pairs correspond to experimentally supported protein-peptide binders, whereas negative pairs were generated by docking the same peptide sequences to non-cognate receptor proteins. In the blind docking setting, no binding-site information was provided. In the known binding-site setting, docking was guided using binding-site information derived from the experimentally resolved complex and was therefore applied only to positive protein-peptide pairs. More negative HPEPDOCK scores indicate more favorable predicted docking. “N/A” indicates that known binding-site docking was not performed for negative pairs because no experimentally supported binding site was available for these non-cognate receptor–peptide combinations.

Table S6. HPEPDOCK scores for Case 1 under blind and known binding-site docking scenarios.

In Supplmentary Data.xlsx (Case 1)

Table S7. HPEPDOCK scores for Case 2 under blind and known binding-site docking scenarios.

In Supplmentary Data.xlsx (Case 2)

Supplementary References

1. Chen, T. XGBoost: A Scalable Tree Boosting System. *Cornell University*, 2016.
2. Zhan, X.; Liu, T.; Yu, C.; Huang, Y.A.; You, Z.; Siu, S.W. MAARDTI: a multi-perspective attention aggregation model for the prediction of drug-target interactions. *Digi. Discov.* 2025, 4, 2994-3007.
3. Chen, M.; Ju, C.; Zhou G.; Chen X.; Zhang T.; Chang K.i; Zaniolo C.; Wang, W., Multifaceted protein-protein interaction prediction based on Siamese residual RCNN, *Bioinfor.* 2019, 35, i305-i314,
4. Zhao, Q.; Zhao, H.; Zheng, K.; Wang, J. HyperAttentionDTI: improving drug-protein interaction prediction by sequence-based deep learning with attention mechanism. *Bioinfor.* 2022, 38, 655-662.
5. Deng, M.; Wang, J.; Zhao, Y.; Zhao, Y.; Cao, H.; Wang, Z. Predicting drug and target interaction with dilated reparameterize convolution. *Sci. Rep.* 2025, 15, 2579.
6. Öztürk, H.; Özgür, A.; Ozkirimli, E. DeepDTA: deep drug-target binding affinity prediction. *Bioinfor.* 2018, 34, i821-i829.
7. Li, Z.; Ren, P.; Yang, H.; Zheng, J.; Bai, F. TEFDTA: a transformer encoder and fingerprint representation combined prediction method for bonded and non-bonded drug-target affinities. *Bioinfor.* 2024, 40, btad778.
8. Zhao, Q.; Duan, G.; Yang, M.; Cheng, Z.; Li, Y.; Wang, J. AttentionDTA: Drug-target binding affinity prediction by sequence-based deep learning with attention mechanism. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 2022, 20, 852-863.
9. Lei, Y.; Li, S.; Liu, Z.; Wan, F.; Tian, T.; Li, S.; Zhao, D.; Zeng, J. A Deep-Learning Framework for Multi-Level Peptide-Protein Interaction Prediction. *Nat. Commun.* 2021, 12.
10. Li, S.; Wan, F.; Shu, H.; Jiang, T.; Zhao, D.; Zeng, J. MONN: a multi-objective neural network for predicting compound-protein interactions and affinities. *Cell sys*. 2020, 10, 308-22.
11. Gower, JC.; Ross, GJ.; Minimum spanning trees and single linkage cluster analysis. *J. R. Stat. Soc. Ser. C Appl. Stat.* 1969, 18, 54-64.