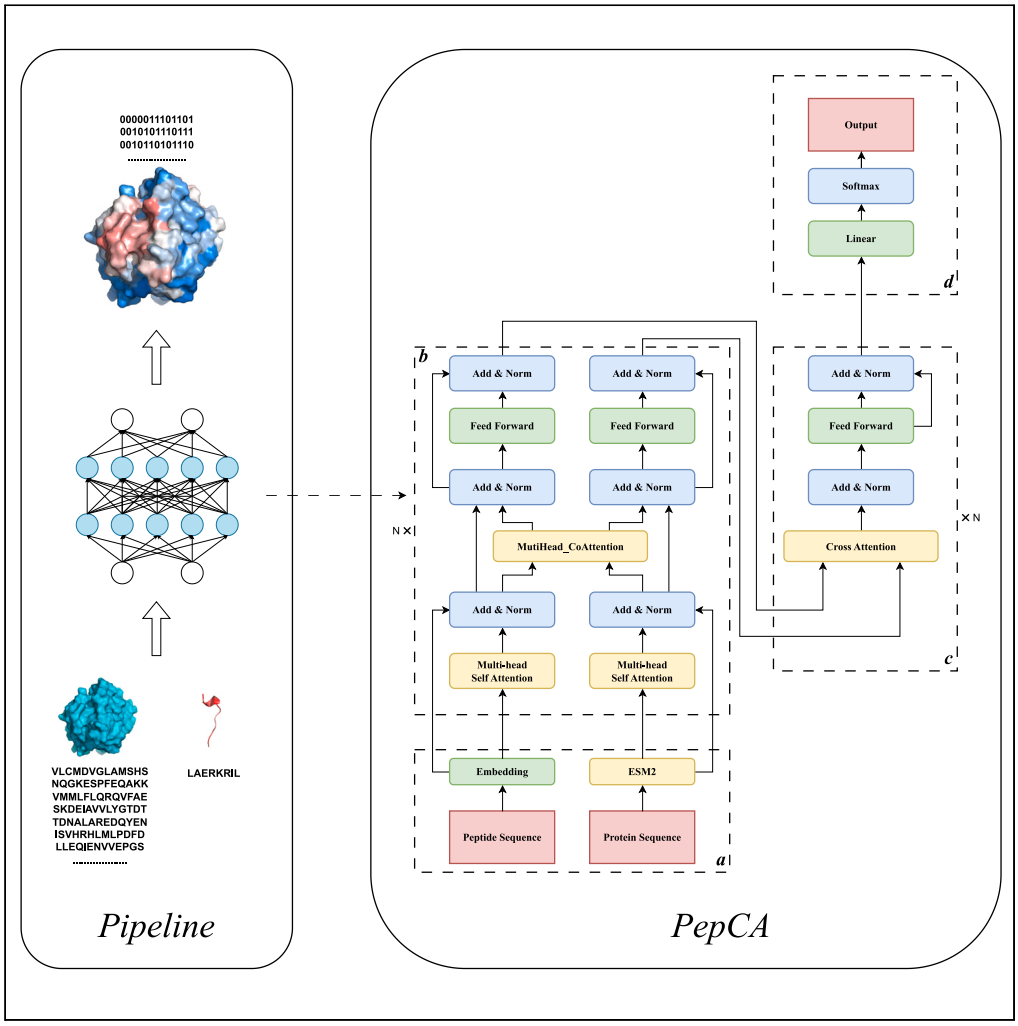
iScience

CellPress
OPEN ACCESS

Article

PepCA: Unveiling protein-peptide interaction sites with a multi-input neural network model



Junxiong Huang, Weikang Li, Bin Xiao, Chunqing Zhao, Hancheng Zheng, Yingrui Li, Jun Wang

wangjun@icarbonx.com (J.W.)
liyr@icarbonx.com (Y.L.)

Highlights

We have developed a network for predicting peptide-protein binding sites, named PepCA

The use of protein language model has enhanced the performance of the network

The model distinguishes positional variations in peptide binding to the same protein

Experimental results on four datasets demonstrate the effectiveness of PepCA

Huang et al., iScience 27, 110850
October 18, 2024
https://doi.org/10.1016/j.isci.2024.110850

iScience

CellPress
OPEN ACCESS

Article

PepCA: Unveiling protein-peptide interaction sites with a multi-input neural network model

Junxiong Huang,[1,5] Weikang Li,[1,5] Bin Xiao,[1,5] Chunqing Zhao,[1,5] Hancheng Zheng,[1,4] Yingrui Li,[1,3,4,5,*] and Jun Wang[1,2,4,5,6,*]

SUMMARY

The protein-peptide interaction plays a pivotal role in fields such as drug development, yet remains underexplored experimentally and challenging to model computationally. Herein, we introduce PepCA, a sequence-based approach for predicting peptide-binding sites on proteins. A primary obstacle in predicting peptide-protein interactions is the difficulty in acquiring precise protein structures, coupled with the uncertainty of polypeptide configurations. To address this, we first encode protein sequences using the Evolutionary Scale Modeling 2 (ESM-2) pre-trained model to extract latent structural information. Additionally, we have developed a multi-input coattention mechanism to concurrently update the encoding of both peptide and protein residues. PepCA integrates this module within an encoder-decoder structure. This model's high precision in identifying binding sites significantly advances the field of computational biology, offering vital insights for peptide drug development and protein science.

INTRODUCTION

Peptides play important roles in various life processes, including signal transduction, programmed cell apoptosis, gene transcription, etc.[1–3] Due to their exceptional safety, excellent tolerability, and unique structural properties, peptides have become an exceptional choice for drug development.[4–6] Peptides mainly interact with a variety of proteins to perform their functions. Therefore, accurately predicting peptide-protein interactions (PepPIs) is the basis for peptide drug discovery and development. However, experimental identification of PepPIs is a time-consuming and costly method.[1] To address this challenge, computational methods have been widely employed to meet the requirements of peptide drug development. Peptide docking is a computational technique that predicts PepPIs by modeling the structural conformations of peptide-protein complexes. Various docking methods such as FlexPepDock,[7] HADDOCK,[8] Pep-SiteFinder,[9] and ADCP[10] are utilized for peptide drug development. Additionally, there are template-based approaches, such as SPOT-Peptide[11] and Interpep2,[12] which further expand the toolkit available for accurately modeling these complex interactions. While peptide docking methods are capable of readily determining the binding structure of peptides with proteins, most of the docking methods are not well suited for peptides owing to their flexibility and larger molecular size.[13,14] With the increasing use of artificial intelligence (AI) in the pharmaceutical industry, numerous AI-based methods for predicting PepPIs are emerging. These AI-based computational methodologies can be classified into two broad categories: structure-based methods and sequence-based methods. Structure-based approaches utilize protein-peptide complex details for predictions. Notable techniques include PepSite,[15] which uses position-specific scoring matrices and distance constraints to locate binding sites, and Peptimap,[16] which employs molecular probes for binding residue mapping. Machine-learning-based algorithms include SPRINT-Str,[17] a random forest predictor using structural features like accessible surface area and secondary structure. Recently, PepNN-Struct[18] was developed, integrating graph and multihead attention mechanisms with peptide and protein embeddings to refine binding residue identification. In contrast, sequence-based methods employ sequence-centric information and have become essential in predicting peptide-binding residues in proteins. Prominent options in this field include the SPRINT-Seq,[19] the consensus-based PepBind,[20] and based on transformer, supplemented by the innovative PepNN-Seq[18] and PepBCL.[21] These advancements collectively demonstrate the dynamic and ongoing evolution in the field of PepPI prediction, showcasing the integration of various computational strategies and machine learning models.

Despite these methodologies achieving high precision in the prediction of peptide-protein binding sites within their respective domains, the field continues to face distinct challenges in PepPI predictions. Structure-based methods rely on the availability of

[1]iCarbonX (Zhuhai) Company Limited, Zhuhai, Guangdong, China
[2]State Key Laboratory of Quality Research in Chinese Medicine, Macau University of Science and Technology, Taipa, Macau, China
[3]Faculty of Health and Medical Sciences, University of Surrey, Guildford, Surrey, UK
[4]Shenzhen Digital Life Institute, Shenzhen, Guangdong, China
[5]iCarbonX (Shenzhen) Pharmaceutical Technology Co, Shenzhen, Guangdong, China
[6]Lead contact
*Correspondence: wangjun@icarbonx.com (J.W.), liyr@icarbonx.com (Y.L.)
https://doi.org/10.1016/j.isci.2024.110850

high-quality peptide-protein complex structures, which are often limited.[22] Moreover, these methods require considerable computational resources due to the complexity of structural data, thus potentially limiting their scalability and applicability.[23] Meanwhile, methods such as AlphaFold for predicting protein structures heavily depend on evolutionary data and may fail for novel proteins that lack extensive evolutionary background.[24–26] Lastly, the dynamic nature of proteins, with conformational changes under varying circumstances, can result in inaccuracies.[27] Hence, these methods may encounter challenges in generalizing to underrepresented proteins in the training data. Sequence-based methods can avoid most of the issues mentioned by relying solely on amino acid sequences, which may overlook vital structural information and often do not directly account for three-dimensional spatial interactions.[28,29] However, these methods also have their own limitations. Generally, sequence-based approaches are considered to have lower predictive accuracy compared to structure-based methods, especially for complex binding sites, and the presence of biological noise in sequence data can further complicate predictions. In relative terms, if the accuracy of sequence-based methods can be significantly enhanced, despite potentially not matching the sophistication of the latest structure-based methods, they still retain extensive applicability across the entire spectrum of PePIs.

To address the aforementioned challenges, we have developed an end-to-end designed, multi-input model, PepCA, for predicting protein-peptide binding residues. This model is improved from the traditional transformer[30] and coattention.[31] This approach marks the application of a multi-input model in the protein-peptide binding domain, utilizing multi-input by processing peptide and protein sequences as two distinct but integrated inputs. Through this integration, PepCA facilitates synchronous updates of both peptide and protein sequences, enhancing its ability to discern variations in binding sites arising from interactions between the same protein and different peptides.

Crucially, the model requires only the sequences of the protein and peptide for its operation. It synergistically combines these sequences with latent protein structural features extracted from the ESM-2[32] (Evolutionary Scale Modeling 2) pre-trained model, which is derived from protein sequences. This integration facilitates a robust prediction of protein-binding sites. The efficacy of PepCA is empirically validated across four benchmark datasets, wherein it consistently outperforms existing models. This marks a significant advancement in the field of protein-peptide interaction prediction, showcasing the potential of multi-input models in enhancing the accuracy and specificity of binding site identification.

RESULTS

Overview of PepCA

In this study, we introduce a neural network architecture designed for the precise prediction of protein-peptide binding sites, as depicted in Figure 1. Our model's architecture is systematically segmented into four integral modules, each contributing uniquely to the task at hand: (1) sequence embedding module, (2) encoder, (3) decoder, and (4) output module.

The sequence embedding module in our PepCA model utilizes the ESM-2 protein language model (pLM) as its pre-trained framework, effectively translating protein sequences. Peptide sequences undergo integer encoding before being processed through a learnable encoding layer. This dual approach facilitates the conversion of sequences into detailed embedding matrices, essential for subsequent analysis. Additionally, ESMFold,[32] a derivative of ESM-2, has proven its superiority in predicting protein structures without relying on multiple sequence alignment (MSA) results, outperforming AlphaFold2.[33] Therefore, despite being primarily a sequence model, the developers believe it can extract useful structural features from amino acid sequences.[32] By employing this pre-trained model for protein sequence encoding, our model may still be able to capture some structural features of proteins. Consequently, we have integrated ESM-2 into our PepCA for protein encoding, leveraging its advanced capabilities for accurate protein-peptide interactions prediction. A notable innovation in our model is the integration of an attention-based module, derived and refined from the multi-input coattention model. Now termed "multihead coattention," this module is instrumental in synchronously updating the peptide and protein embeddings. Its design ensures equal distribution of unnormalized attention values across both protein and peptide residues, critical for capturing the reciprocal influence and dynamic interplay in the binding process. The encoder module processes the enhanced protein feature matrix, while the decoder module engages in cross-attention mechanisms with the initial input matrix. This interaction facilitates the generation of complex and informative representations of the binding dynamics. Finally, the output module is meticulously crafted to compute residue-level peptide-binding probabilities. Its precision allows for the identification and quantification of specific residues' involvement in peptide binding within the input sequence, underscoring the model's capability in pinpointing binding sites with high accuracy.

Dataset

In this research, we selected four widely recognized benchmark datasets to accurately assess and juxtapose our new method with previous approaches. For ease of reference, we labeled them as TS092, TS251, TS639, and TS125. Specifically, TS092 was introduced through the application of the PepNN.[18] Concurrently, TS251 was developed utilizing the Interpep.[34] Dataset TS639 was generated by leveraging the PepBind[20] technique. Lastly, TS125 was conceived via the SPRINT-Seq.[19] It is important to mention that each dataset was independently processed using software like "blastclust" from the BLAST package[35] or MMseq2[36] to reduce sequence identity to 30%, which helps minimize evaluation bias. Peptide-binding residues are defined as those residues containing at least one atom, which is positioned at a distance of less than 3.5 Å[18–20,34] from any atom within the peptide. A comprehensive overview of these datasets is presented in Table 1. Further specifics on how these datasets were used in model training and evaluation are detailed further.

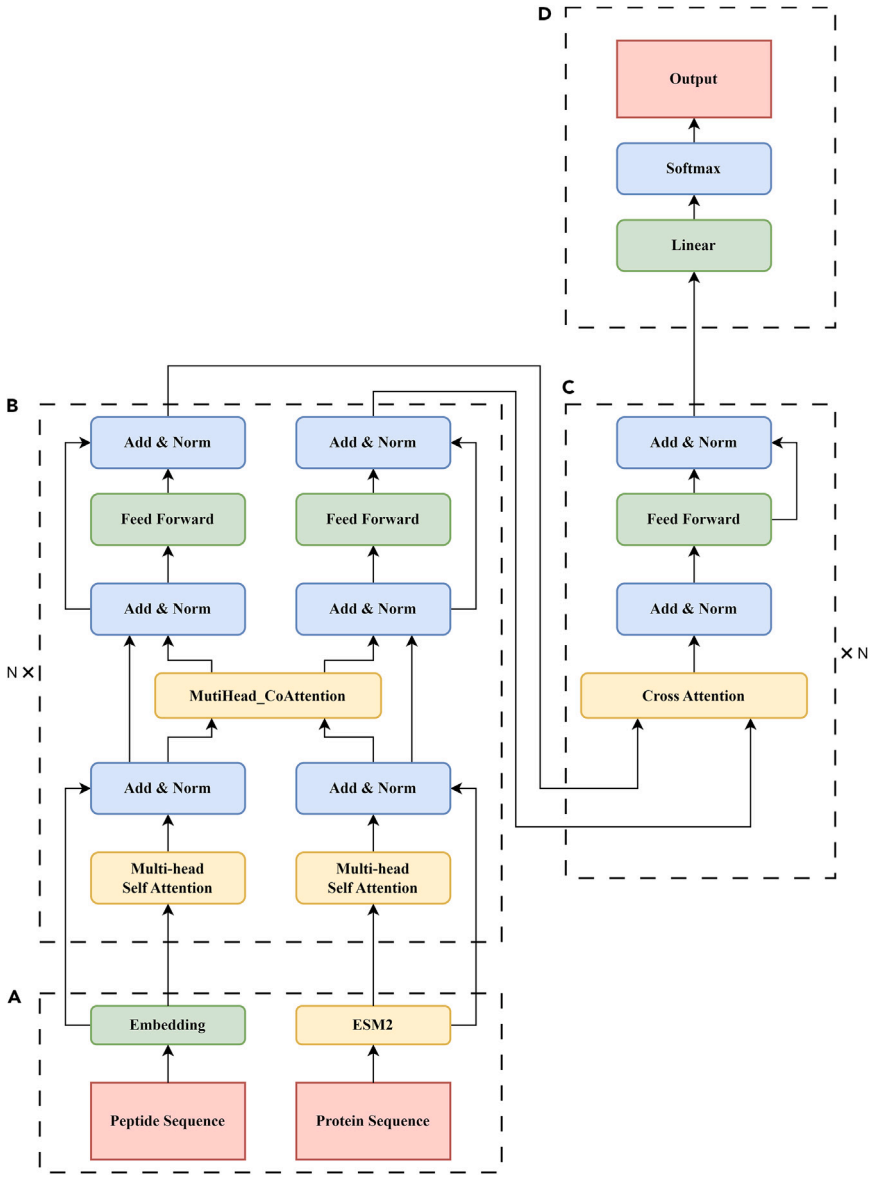


Figure 1. The workflow and architecture of PepCA

(A) Model's embedding layer.

(B) Multi-input multihead coattention component, responsible for concurrently extracting the sequence features of peptides and proteins.

(C) Cross-attention component that amalgamates the characteristics of both peptides and proteins.

(D) Output layer.

PepCA outperforms baseline methods in binding point prediction

To evaluate the performance of our newly proposed PepCA model, we compared it with two state-of-the-art methodologies: PepNN-Seq[18] and PepBCL.[21] These approaches represent the latest developments in end-to-end prediction technologies post-2021. When testing our model on datasets TS251 and TS639, we also included Interpep,[34] PepBind, and PepCNN[37] in our comparison. However, Interpep is a structure-based method, whereas PepBind and PepCNN employ various predictive tools to derive protein features for input into its model. Given these fundamental methodological differences, our study did not engage in an extensive comparison with these additional techniques and focused on methods more similar to our sequence-based, end-to-end predictive framework. Additionally, during our comparisons on the TS125 dataset, we also included traditional machine learning methods such as SPRINT-Seq[19] to broaden the scope of our evaluation.

We conducted performance evaluations of our model on four benchmark datasets: TS092, TS251, TS639, and TS125. To assess the efficacy of all models, we employed five key metrics: sensitivity, specificity, precision, the Matthews correlation coefficient (MCC), and the area under the receiver operating characteristic curve (AUC). Furthermore, since the methods being compared are each trained exclusively on their

Table 1. Summary of datasets

	TS092		TS251		TS639		TS125	
Dataset	Training Set	Testing Set	Training Set	Testing Set	Training Set	Testing Set	Training Set	Testing Set
Number of proteins	2,828	92	251	251	640	639	1,115	125
Number of residues	689,320	23,173	54,488	58,604	157,225	150,328	266,434	30,870
Number of binding residues	85,533	2,523	6,627	6,617	8,259	8,490	14,829	1,716
Number of non-binding residues	603,787	20,650	47,861	51,987	148,966	141,838	251,605	29,154

respective datasets, to ensure a fair comparison of their differences, our model was also trained and tested separately on four distinct datasets. The comparative results are shown in Table 2, respectively. Due to the fact that most methods do not make their trained models publicly available, the results of our comparative analysis are directly extracted from their respective studies.

As depicted in Table 2, our neural network model PepCA significantly outperforms sequence-based methods such as SPRINT-Seq, PepBind, Visual, PepNN-Seq, and PepBCL across four datasets. PepCNN integrates features from pre-trained pLM, evolutionary relationships in the protein sequences using an MSA tool, and the structural attributes in terms of the solvent exposure of the residues in the sequences. PepCA achieves comparable or superior performance metrics, particularly in AUC, MCC, and sensitivity. This underscores the robustness of PepCA, driven solely by features from pre-trained language models, in accurately predicting peptide-protein binding site. Additionally, it is important to note that existing methods are almost exclusively trained on benchmark datasets, which often display a significant imbalance between the number of binding and non-binding sites, resulting in lower sensitivity. However, as can be seen from the table, our method consistently maintains a high level of sensitivity, only falling short of visual in the TS125 dataset. In all other cases, it achieves the highest level of sensitivity. Consequently, our approach is more likely to avoid missing many true protein-peptide binding sites in real-world applications.

In order to assess the specific scenarios for each dataset, we utilized the trained model provided by PepNN-Seq to re-predict the proteins in the four test sets. Additionally, to maintain consistent comparisons and due to the failure in loading the pre-trained PepBCL models, we used the source code from PepBCL to train models on all four datasets: TS092, TS251, TS639, and TS125. It is worth noting that the performance metrics of the models retrained using PepBCL on the four datasets differed slightly from those reported in the publications. Additionally, due to issues with precision, Figure 2 for the PepNN-Seq method shown in the graphs may slightly differ from those in Table 2. In Figures 2A–2D, it is observable that within the datasets TS092, TS251, TS639, and TS125, the receiver operating characteristic (ROC) curves of PepCA consistently surpass those of other existing methods, culminating in the achievement of the highest AUC scores. As illustrated in Figure 2E, for the comparison of individual protein AUC values, we generated boxplots. These plots reveal that, in comparison to the PepNN-Seq and PepBCL models, the PepCA model consistently demonstrated superior median AUC values across all four datasets (TS092, TS251, TS639, and TS125). Notably, the median, indicated by the central line within each blue box (PepCA), is positioned above the corresponding orange box (PepNN-Seq) and green box (PepBCL). This suggests that, at least in terms of central tendency, PepCA demonstrates superior predictive capabilities in identifying protein-ligand binding sites over PepNN-Seq and PepBCL.

Relative to the other two models, the interquartile range of PepCA indicates more consistent performance. Additionally, the consistency of PepCA's AUC values across all datasets is exceptional, with its lower quartile never falling below the highest lower quartile of the other models. Moreover, despite the presence of outliers and some overlap within the quartile ranges, the overall distribution of the PepCA model's predictions tends to skew toward higher AUC values. This tendency is moderately evident in the TS092 and TS251 datasets, where the median of PepCA nearly matches the 75th percentile of PepNN-Seq. These observations underscore the robustness and reliability of the PepCA model, which maintains higher or competitive median AUC values compared to PepNN-Seq and PepBCL, suggesting a potentially more accurate predictive performance in protein analysis. To more clearly articulate our assertion that individual proteins' AUC values in our model markedly exceed those in the other two models, we conducted a t test analysis on these values across four datasets. Our observations revealed that only in the TS092 and TS125 dataset was there no statistically significant difference. This lack of marked significance in the TS092 and TS125 dataset is likely attributable to its constrained sample scope. Conversely, in the other datasets, the p values were consistently below 0.05, denoting significant disparities in the performance of our model across these more expansive and varied datasets.

The performance of our model in datasets other than TS092 and TS125 was markedly superior. The p values less than 0.05 in these datasets demonstrate a clear statistical advantage of our model over the PepBCL model. This suggests that our model is particularly effective in dealing with a wide range of protein samples, exhibiting robustness and adaptability in varied analytical scenarios. In Figure 2F, we randomly selected a protein-peptide complex (PDB ID: 6ICA) from the TS092 test set, where it is evident that our model's predictions are closer to the actual scenario compared to the other two models. Overall, PepCA demonstrated superior performance, underscoring its efficacy in this domain.

Simultaneously, to further validate the universality of our model, we tested it on four different datasets using the models trained on the respective datasets. The results show little variation in the AUC values. Using the results from the same dataset for training and testing as a benchmark, the AUC values when tested on other datasets fall mostly within 0.01 below the benchmark, with the lowest drop being

Table 2. Comparison of the proposed PepCA and other methods

Test dataset	Model	Sensitivity	Specificity	Precision	MCC	AUC
TS092	PepNN-Seq[18]	–	–	–	0.272	0.781
	PepCA (ours)	0.405	0.935	0.434	0.351	0.817
TS251	PepNN-Seq[18]	–	–	–	0.277	0.769
	Interpep[34]	–	–	–	–	0.793
	PepCA (ours)	0.471	0.902	0 380	0.340	0.796
TS639	PepBind[20]	0.317	–	0.450	0.348	0.767
	PepNN-Seq[18]	–	–	–	0.251	0.792
	PepBCL[21]	0.252	0.983	0.470	0.312	0.804
	PepCNN[37]	0.217	0.986	0.479	0.297	0.826
	PepCA (ours)	0.399	0.945	0.343	0.302	0.826
TS125	PepSite[15]	0.180	0.970	–	0.200	0.610
	Peptimap[16]	0.320	0.950	–	0.270	0.630
	SPRINT-Seq[19]	0.210	0.960	–	0.200	0.680
	PepBind[20]	0.344	–	0.469	0.372	0.793
	Visual[38]	0.670	0.680	–	0.170	0.730
	PepNN-Seq[18]	–	–	–	0.278	0.794
	PepBCL[21]	0.315	0.984	0.540	0.383	0.815
	PepCNN[37]	0.254	0.988	0.55	0.350	0.843
	PepCA (ours)	0.386	0.967	0.405	0.360	0.848

0.019 in one instance. In cases where the AUC values were above the benchmark, the highest increase observed was 0.04. Moreover, the AUC value was consistently superior to the PepNN-Seq model on all datasets. Refer to Table S1 for detailed data.

Performance enhancement in modeling through pre-trained pLMs

In our advanced PepCA model, we utilized ESM-2, a state-of-the-art pLM, for pre-training to encode protein sequences into high-dimensional matrices. This integration is pivotal for capturing the complex biological nuances inherent in protein structures.[39–41] Known for its deep learning prowess in handling large-scale protein datasets, the model infuses our system with a comprehensive understanding of protein dynamics and interactions. As illustrated in Figure 3, the performance of the PepCA model using the ESM-2 pre-trained model significantly surpasses that of its counterparts without this enhancement. Ablation studies comparing models with and without this pre-training revealed a substantial performance enhancement, attributable to the model's robust ability to distill and incorporate rich biological information. Although excellent in encoding protein sequences, a decline in overall performance was observed when the model was used to encode peptide sequences as well. This may be due to its training primarily on protein databases, and the inherent differences between protein and peptide sequences hindering its effectiveness in capturing peptide features.

The advent of PepCA and its superiority over single protein sequence prediction methods

One of the improvements of our model is the ability to input protein and peptide sequences simultaneously, which enables the model to identify the binding sites when the same protein interacts with different peptides. This advancement effectively addresses the limitations observed in single-protein sequence models, including the previously utilized PepBCL framework. As illustrated in Figure S1, the model's capability is exemplified by analyzing the protein 6J8F_B (chain B in 6J8F) in complex with peptides 6J8F_A (chain A in 6J8F) and 6J8F_C (chain C in 6J8F). Notably, while there are overlapping regions in the binding sites, the majority of interactions are distinct, predominantly occurring at the respective binding sites of each peptide. In the protein-peptide complex designated as 6TYT, analogous outcomes were observed. These results illustrate the effectiveness of our model in dealing with the interaction dynamics between single proteins and multiple peptides.

Evaluation of PepCA model performance using random and natural peptide sequences

In this study, we assessed the capability of the PepCA model to predict peptide-binding sites using peptides of random lengths less than 30 amino acids (based on TS125 and TS639) as input sequences. Our findings demonstrate that the model exhibits enhanced performance when provided with natural peptide sequences as opposed to random sequences. Furthermore, comparative analysis with the PepBCL model, which solely outputs protein sequences, indicated that PepCA's performance is comparable or superior when random peptides are input,

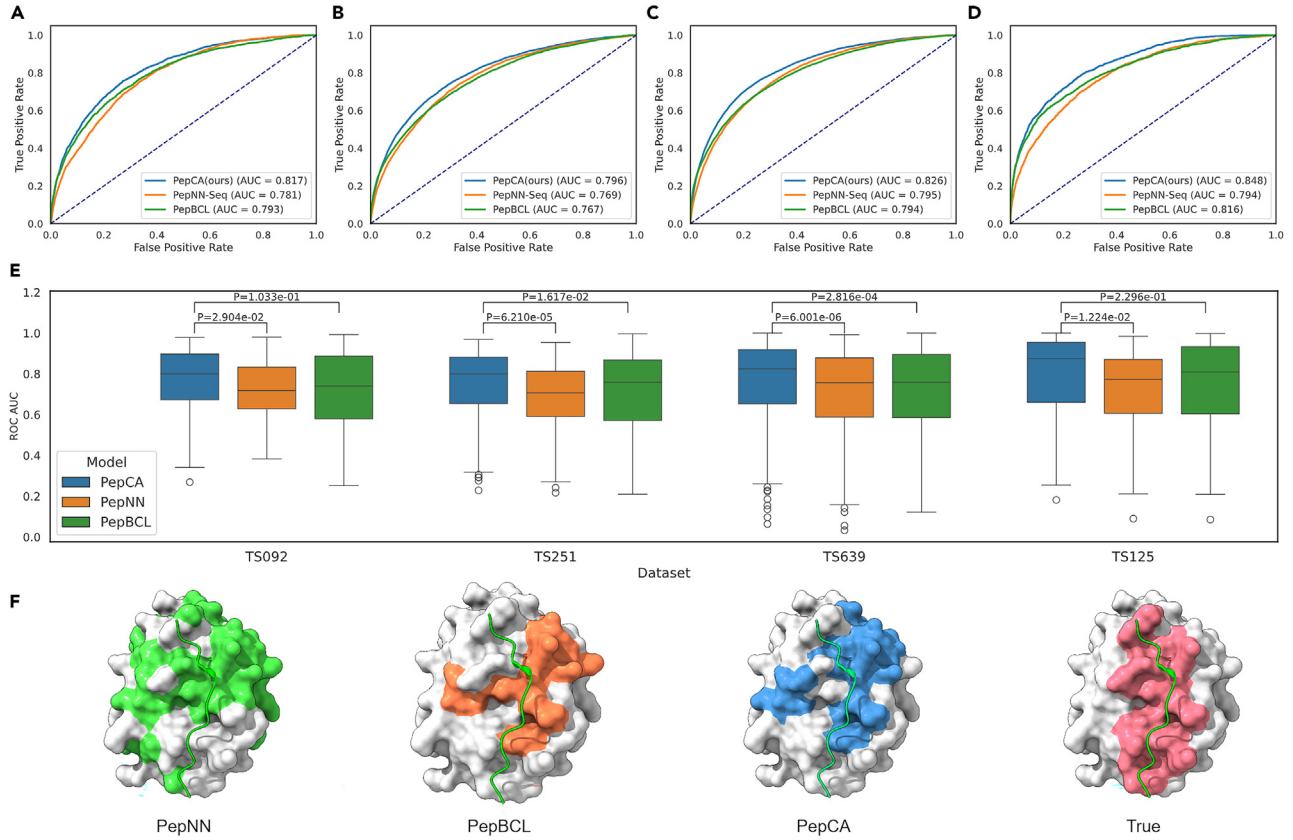


Figure 2. Comparative analysis of PepCA and two similar models across four benchmark datasets

(A) ROC curves on all residues in the dataset using predictions from PepCA, PepNN-Seq, and PepBCL trained on dataset TS092.

(B) ROC curves on all residues in the dataset using predictions from PepCA, PepNN-Seq, and PepBCL trained on dataset TS251.

(C) ROC curves on all residues in the dataset using predictions from PepCA, PepNN-Seq, and PepBCL trained on dataset TS639.

(D) ROC curves on all residues in the dataset using predictions from PepCA, PepNN-Seq, and PepBCL trained on dataset TS125.

(E) Comparison of the distribution of AUCs on different input proteins using predictions from PepCA, PepNN-Seq, and PepBCL trained on four different datasets.

(F) Comparative analysis of the PepCA, PepNN-Seq, and PepBCL effects on a specific protein randomly selected from the TS092 test set, PDB ID is 6ICA.

as evidenced by results on the TS639 and TS125 datasets. Collectively, these results suggest that while PepCA maintains a commendable performance advantage in scenarios where peptide sequences binding to proteins are unknown, the accuracy of the model is improved by using defined natural peptide sequences. For detailed results, refer to Table 3.

Model interpretability through attention mechanism comparative analysis with molecular dynamics

The experimental findings demonstrate that the proposed model is highly effective in predicting protein-peptide binding residues, as evidenced by its commendable 70% overlap in the top 10 residues identified through attention scores for 6RMV_A_C and those obtained from molecular dynamics simulations (Table S2). Molecular dynamics simulations reveal key residues based on their dynamic interactions and stability within the protein-peptide complex, indicating their central role in maintaining the structural integrity and facilitating the binding process. This significant congruence highlights the model's ability to not only capture the abstract patterns of sequence but also to align its learning with biologically relevant interactions such as hydrogen bonds, electrostatic interactions, and hydrophobic contacts, which are critical for the biological function of the complex. By assigning high attention scores to many of the same residues pinpointed as crucial by molecular dynamics, our model underscores its utility in practical applications like drug design and protein engineering. Furthermore, this overlap acts as a validation of our model's effectiveness and provides an insightful glimpse into the "black box" of deep learning, illustrating how the attention mechanism within the model prioritizes biologically significant regions, thus offering a clearer understanding of how such models can be interpretatively and reliably applied to complex biological data.

DISCUSSION

In our study, we present PepCA, a multi-input neural network for accurately predicting protein-peptide binding sites. PepCA processes both protein and peptide sequences through integrated inputs, outperforming existing models like PepNN-Seq[18] and PepBCL.[21] Our model

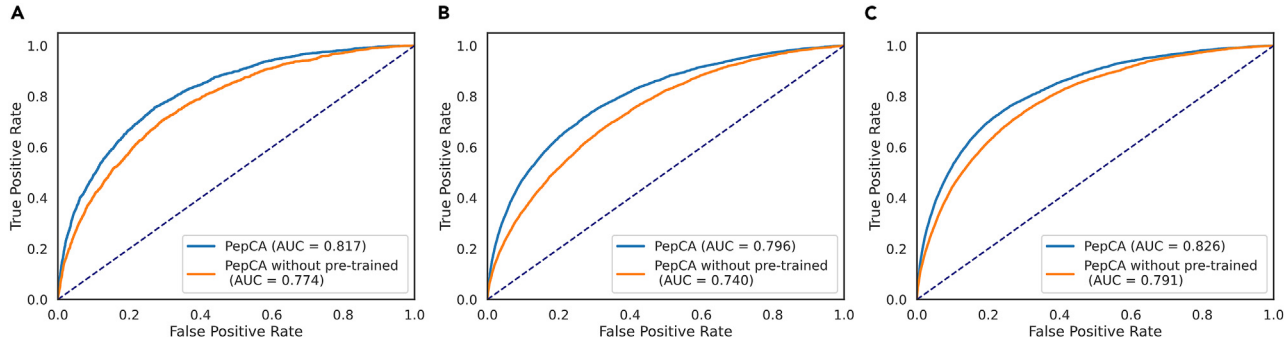


Figure 3. Comparison of model effects with and without ESM-2 pre-trained model in three datasets

(A) ROC curves on all residues in the dataset TS092.

(B) ROC curves on all residues in the dataset TS251.

(C) ROC curves on all residues in the dataset TS639.

employs ESM-2[32] for feature extraction, which is trained on protein structure data. As such, ESM-2 is capable of effectively extracting structural features directly from protein sequences. This enables PepCA to utilize structural features of proteins solely from sequence information. Our model combines the advantages of structure-based and sequence-based methods, overcoming their individual limitations. However, it is imperative to note that, despite our model's commendable performance in sequence-based end-to-end methodologies, it exhibits certain limitations when compared to structurally focused approaches like PepNN-Struct[18] and prediction methods that utilize tools like DSSP,[42] such as DeepProSite.[43] Specifically, in the context of the TS639 dataset, the respective AUC values for these two methods stand at 0.838 and 0.861, underscoring a comparative disadvantage of our model. This comparative shortfall in our model's performance can be attributed to its reliance on extracting features solely from protein and peptide sequences. Currently, this method of feature extraction is not the most accurate, as it overlooks the intricacies of protein and peptide structures. On the other hand, approaches that obtain detailed protein and peptide structures and employ tools like DSSP, though demonstrably more precise, are markedly expensive and time-consuming. This trade-off between accuracy and efficiency is a significant consideration in the development and application of computational models in proteomics. PepCA's design, featuring sequence embedding, encoder, decoder, and the multihead coattention module compute precise residue-level binding probabilities, enhanced by pre-training with ESM-2. This integration of advanced deep learning and multi-input data processing marks a significant progression in computational biology and bioinformatics. PepCA's interpretability is a key aspect, aligning its attention scores with molecular dynamics results to understand binding mechanisms. This approach moves beyond the "black box" nature of many deep learning models, offering insights into protein-peptide interactions. PepCA not only predicts binding sites with high accuracy but also advances our understanding of these complex interactions.

The model's potential in drug development and biomolecular research is notable, particularly in designing peptide-based therapeutics. PepCA's interpretability, demonstrated through attention score visualization and alignment with molecular dynamics, sheds light on binding patterns, enhancing the field of bioinformatics.

In conclusion, PepCA sets a new standard in protein-peptide interaction prediction by integrating sequence and structural data with advanced deep learning techniques. Future research could extend this model to accommodate more peptide and protein types, further advancing protein science.

Limitations of the study

This study introduces an innovative model for predicting protein-peptide interactions. However, it faces limitations, such as potential generalizability issues with atypical protein structures; dependence on high-quality sequence data; computational resource intensity; lack of consideration for dynamic protein conformational changes; and challenges in model interpretability and scalability to large datasets or

Table 3. Comparison of natural peptides and random peptides

Test dataset	Model	Sensitivity	Specificity	Precision	MCC	AUC
TS639	PepBCL[21]	0.252	0.983	0.470	0.312	0.804
	PepCA (random peptides)	0.391	0.939	0.279	0.282	0.805
	PepCA	0.399	0.945	0.343	0.302	0.826
TS125	PepBCL[21]	0.315	0.984	0.540	0.383	0.815
	PepCA (random peptides)	0.307	0.977	0.439	0.337	0.820
	PepCA	0.386	0.967	0.405	0.360	0.848

proteome-wide analyses. Additionally, the reliance on sequences sourced from the PDB rather than full-length native sequences in this study, due to some sequences without structure information in the PDB, needs to be considered in the future study.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Jun Wang (wangjun@icarbonx.com).

Materials availability

This study did not generate new unique reagents.

Data and code availability

All original code and data have been deposited at GitHub (https://github.com/cloudaner115/PepCA).

- All raw data have been uploaded to GitHub (https://github.com/cloudaner115/PepCA).
- Our source code is available at GitHub (https://github.com/cloudaner115/PepCA).
- Any additional information required is available from the lead contact upon request.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China (U23A6012, 82072818) and the Science and Technology Development Fund Macau SAR (file no. 006/2023/SKL).

AUTHOR CONTRIBUTIONS

J.W., Y.L., and J.H. conceived the concept. Y.L. and J.H. designed methodology and performed the experiments. J.H. and H.Z. conducted the coding and trained the models. J.H., W.L., B.X., and C.Z. analyzed the results. J.W. and Y.L. supervised the entire project. J.H. and W.L. wrote the initial draft of the paper. All authors contributed to the revision of the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Trainning
- METHOD DETAILS
  - Peptide sequence embedding module
  - Protein sequence embedding module with ESM-2
  - Transformer-based encoder with multihead coattention
  - Transformer-based decoder with crossattention
  - MD simulations
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.110850.

Received: April 16, 2024
Revised: June 13, 2024
Accepted: August 27, 2024
Published: August 30, 2024

REFERENCES

1. Rubinstein, M., and Niv, M.Y. (2009). Peptidic modulators of protein-protein interactions: progress and challenges in computational design. Biopolymers *91*, 505–513.
2. Pawson, T., and Nash, P. (2003). Assembly of cell regulatory systems through protein interaction domains. Science *300*, 445–452.
3. Zhang, Q.C., Petrey, D., Deng, L., Qiang, L., Shi, Y., Thu, C.A., Bisikirska, B., Lefebvre, C., Accili, D., Hunter, T., et al. (2012). Structure-based prediction of protein–protein interactions on a genome-wide scale. Nature *490*, 556–560.
4. Lau, J.L., and Dunn, M.K. (2018). Therapeutic peptides: Historical perspectives, current development trends, and future directions. Bioorg. Med. Chem. *26*, 2700–2707.
5. Lee, A.C.-L., Harris, J.L., Khanna, K.K., and Hong, J.-H. (2019). A comprehensive review on current advances in peptide drug development and design. Int. J. Mol. Sci. *20*, 2383.
6. Muttenthaler, M., King, G.F., Adams, D.J., and Alewood, P.F. (2021). Trends in peptide drug discovery. Nat. Rev. Drug Discov. *20*, 309–325.
7. Raveh, B., London, N., Zimmerman, L., and Schueler-Furman, O. (2011). Rosetta flexpepdock ab-initio: simultaneous folding,

docking and refinement of peptides onto their receptors. PLoS One *6*, e18934.
8. Dominguez, C., Boelens, R., and Bonvin, A.M.J.J. (2003). HADDOCK: a protein- protein docking approach based on biochemical or biophysical information. J. Am. Chem. Soc. *125*, 1731–1737.
9. Saladin, A., Rey, J., Thévenet, P., Zacharias, M., Moroy, G., and Tufféry, P. (2014). Pepsitefinder: a tool for the blind identification of peptide binding sites on protein surfaces. Nucleic Acids Res. *42*, W221–W226.
10. Zhang, Y., and Sanner, M.F. (2019). AutoDock CrankPep: combining folding and docking to predict protein–peptide complexes. Bioinformatics *35*, 5121–5127.
11. Litfin, T., Yang, Y., and Zhou, Y. (2019). Spotpeptide: template-based prediction of peptide-binding proteins and peptide-binding sites. J. Chem. Inf. Model. *59*, 924–930.
12. Johansson-Åkhe, I., Mirabello, C., and Wallner, B. (2020). Interpep2: global peptide–protein docking using interaction surface templates. Bioinformatics *36*, 2458–2465.
13. Agrawal, P., Singh, H., Srivastava, H.K., Singh, S., Kishore, G., and Raghava, G.P.S. (2019). Benchmarking of different molecular docking methods for protein-peptide docking. BMC Bioinf. *19*, 426.
14. Weng, G., Gao, J., Wang, Z., Wang, E., Hu, X., Yao, X., Cao, D., and Hou, T. (2020). Comprehensive evaluation of fourteen docking programs on protein–peptide complexes. J. Chem. Theor. Comput. *16*, 3959–3969.
15. Petsalaki, E., Stark, A., García-Urdiales, E., and Russell, R.B. (2009). Accurate prediction of peptide binding sites on protein surfaces. PLoS Comput. Biol. *5*, e1000335.
16. Lavi, A., Ngan, C.H., Movshovitz-Attias, D., Bohnuud, T., Yueh, C., Beglov, D., Schueler-Furman, O., and Kozakov, D. (2013). Detection of peptide-binding sites on protein surfaces: The first step toward the modeling and targeting of peptide-mediated interactions. Proteins *81*, 2096–2105.
17. Taherzadeh, G., Zhou, Y., Liew, A.W.-C., and Yang, Y. (2018). Structure-based prediction of protein–peptide binding regions using random forest. Bioinformatics *34*, 477–484.
18. Abdin, O., Nim, S., Wen, H., and Kim, P.M. (2022). PepNN: a deep attention model for the identification of peptide binding sites. Commun. Biol. *5*, 503.
19. Taherzadeh, G., Yang, Y., Zhang, T., Liew, A.W.-C., and Zhou, Y. (2016a). Sequence-based prediction of protein–peptide binding sites using support vector machine. J. Comput. Chem. *37*, 1223–1229.
20. Zhao, Z., Peng, Z., and Yang, J. (2018). Improving sequence-based prediction of protein–peptide binding residues by introducing intrinsic disorder and a consensus method. J. Chem. Inf. Model. *58*, 1459–1468.
21. Wang, R., Jin, J., Zou, Q., Nakai, K., and Wei, L. (2022). Predicting protein–peptide binding residues via interpretable deep learning. Bioinformatics *38*, 3351–3360.
22. Ciemny, M., Kurcinski, M., Kamel, K., Kolinski, A., Alam, N., Schueler-Furman, O., and Kmiecik, S. (2018). Protein–peptide docking: opportunities and challenges. Drug Discov. Today *23*, 1530–1537.
23. Scardino, V., Di Filippo, J.I., and Cavasotto, C.N. (2023). How good are alphafold models for docking-based virtual screening? iScience *26*, 105920.
24. Ruff, K.M., and Pappu, R.V. (2021). Alphafold and implications for intrinsically disordered proteins. J. Mol. Biol. *433*, 167208.
25. Stevens, A.O., and He, Y. (2022). Benchmarking the accuracy of alphafold 2 in loop structure prediction. Biomolecules *12*, 985.
26. Bertoline, L.M.F., Lima, A.N., Krieger, J.E., and Teixeira, S.K. (2023). Before and after alphafold2: An overview of protein structure prediction. Front. Bioinform. *3*, 1120370.
27. Buel, G.R., and Walters, K.J. (2022). Can alphafold2 predict the impact of missense mutations on structure? Nat. Struct. Mol. Biol. *29*, 1–2.
28. Chen, J., Xie, Z.-R., and Wu, Y. (2017). Understand protein functions by comparing the similarity of local structural environments. Biochim. Biophys. Acta, Proteins Proteomics *1865*, 142–152.
29. Rigden, D.J., and Rigden, D.J. (2009). From Protein Structure to Function with Bioinformatics (Springer).
30. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. Adv. Neural Inf. Process. Syst. *30*.
31. Xiong, C., Zhong, V., and Socher, R. (2016). Dynamic coattention networks for question answering. Preprint at arXiv. https://doi.org/10.48550/arXiv.1611.01604.
32. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. Science *379*, 1123–1130.
33. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with alphafold. Nature *596*, 583–589.
34. Johansson-Åkhe, I., Mirabello, C., and Wallner, B. (2019). Predicting protein-peptide interaction sites using distant protein complexes as structural templates. Sci. Rep. *9*, 4267.
35. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.
36. Steinegger, M., and Söding, J. (2017). Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat. Biotechnol. *35*, 1026–1028.
37. Chandra, A., Sharma, A., Dehzangi, I., Tsunoda, T., and Sattar, A. (2023). PepCNN deep learning tool for predicting peptide binding residues in proteins using sequence, structural, and language model features. Sci. Rep. *13*, 20882.
38. Wardah, W., Dehzangi, A., Taherzadeh, G., Rashid, M.A., Khan, M.G.M., Tsunoda, T., and Sharma, A. (2020). Predicting protein-peptide binding sites with a deep convolutional neural network. J. Theor. Biol. *496*, 110278.
39. Gong, J., Jiang, L., Chen, Y., Zhang, Y., Li, X., Ma, Z., Fu, Z., He, F., Sun, P., Ren, Z., and Tian, M. (2023). THPLM: a sequence-based deep learning framework for protein stability changes prediction upon point variations using pretrained protein language model. Bioinformatics *39*, btad646.
40. Chen, T., Hong, L., Yudistyra, V., Vincoff, S., and Chatterjee, P. (2023). Generative design of therapeutics that bind and modulate protein states. Curr. Opin. Biomed. Eng. *28*, 100496.
41. Brixi, G., Ye, T., Hong, L., Wang, T., Monticello, C., Lopez-Barbosa, N., Vincoff, S., Yudistyra, V., Zhao, L., Haarer, E., et al. (2023). SaLT&epPr is an interface-predicting language model for designing peptide-guided protein degraders. Commun. Biol. *6*, 1081.
42. Touw, W.G., Baakman, C., Black, J., te Beek, T.A.H., Krieger, E., Joosten, R.P., and Vriend, G. (2015). A series of pdb-related databanks for everyday needs. Nucleic Acids Res. *43*, D364–D368.
43. Fang, Y., Jiang, Y., Wei, L., Ma, Q., Ren, Z., Yuan, Q., and Wei, D.-Q. (2023). DeepProSite: Structure-aware protein binding site prediction using esmfold and pretrained language model. Bioinformatics *39*, btad718. https://api.semanticscholar.org/CorpusID:265497196.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Dataset in this paper	This paper	https://github.com/cloudaner115/PepCA
Software and algorithms		
Python	Python Software Foundation	https://www.python.org/
PyTorch	PyTorch Foundation	https://pytorch.org/
Pandas	AQR Capital Management	https://pandas.pydata.org/
PepCA	This paper	https://github.com/cloudaner115/PepCA

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Trainning

Training was done using an AdamW optimizer with a learning rate of 1e-6 during the pre-training and fine-tuning of PepCA, on an Nvidia A10 GPU. A weighted cross-entropy loss was optimized to take into account the fact that the training dataset is skewed towards non-binding residues.

METHOD DETAILS

Peptide sequence embedding module

In computational proteomics, converting peptide sequences into a numerical format, known as 'integer encoding', is crucial. This involves assigning a unique integer to each amino acid, typically in alphabetical order. For example, in the sequence 'ACD', 'A' (Alanine) is encoded as 1, 'C' (Cysteine) as 2, and 'D' (Aspartic acid) as 3. This method transforms complex peptide structures into a standardized numerical format, enabling the application of advanced computational models, like machine learning, to analyze and predict peptide structures and functions more effectively.

Protein sequence embedding module with ESM-2

Our protein coding approach uses the 650 million parameter ESM-2[32] protein language model (pLM) developed at Meta AI. This architecture facilitates the characterization of protein sequences, eliminating the need to generate MSAs. ESM-2 represents a cutting-edge deep learning model specifically tailored for protein sequence encoding and interpretation. Trained on a vast dataset of protein sequences, it captures evolutionary relationships and structural features between proteins. In protein encoding, ESM-2 transforms protein sequences into rich, high-dimensional representations that encapsulate biological characteristics and functional information of the sequences.

Transformer-based encoder with multihead coattention

In this study, we introduce a approach for processing protein and peptide sequences, distinct from traditional attention mechanisms, termed multihead coattention. This method enables parallel processing of two different types of input data: protein sequences and peptide sequences. The core architecture of multihead coattention is based on the coattention module. Coattention[31] is a widely used technique in deep learning for handling multiple input sources, representing a variant of the attention mechanism designed to model two or more different inputs or data sources concurrently. This approach has demonstrated significant potential in fields such as natural language processing, computer vision, and multi-input learning.

In our research, the multihead coattention method is specifically tailored for handling interactions between protein and peptide sequences. The method encompasses the following key steps: Firstly, features are extracted from each input source, namely protein and peptide sequences. Subsequently, the similarity or association between the features of these different input sources is computed. Finally, based on the computed similarity matrix, attention weights are generated to weight the input features. This weighting mechanism enables us to effectively focus on those sequence regions that are particularly important for the interaction between proteins and peptides, thereby enhancing the accuracy of predicting protein-peptide interactions. The coattention mechanism is described as follows:

$$\begin{cases} Prot & = Prot\_seqW^{Prot} \\ Pep & = Pep\_seqW^{Pep} \\ V_{prot} & = Prot\_seqW^{V_{prot}} \\ V_{pep} & = Pep\_seqW^{V_{pep}} \end{cases} \quad \text{(Equation 1)}$$

$$Similarity = Prot \cdot Pep \quad \text{(Equation 2)}$$

$$\text{Attention}_{\text{prot}} = \text{softmax}\left(\frac{Similarity}{\sqrt{d_k}}\right) V_{\text{pep}} \quad \text{(Equation 3)}$$

$$\text{Attention}_{\text{pep}} = \text{softmax}\left(\frac{Similarity^T}{\sqrt{d_k}}\right) V_{\text{prot}} \quad \text{(Equation 4)}$$

In this framework, $Prot\_seq \in \mathbb{R}^{L_{\text{prot}} \times d_{\text{model}}}$ represents the protein matrix encoded by a pre-trained model, and $Pep\_seq \in \mathbb{R}^{L_{\text{pep}} \times d_{\text{model}}}$ denotes the peptide matrix encoded by the encoding layer. Subsequently, they are individually transformed through the linear layers W^{Prot} and W^{Pep}, yielding new matrices $Prot \in \mathbb{R}^{L_{\text{prot}} \times d_k}$ and $Pep \in \mathbb{R}^{L_{\text{pep}} \times d_k}$ for proteins and peptides, respectively. Concurrently, they are processed through the linear layers $W^{V_{\text{prot}}}$ and $W^{V_{\text{pep}}}$ respectively, resulting in the protein value matrix $V_{\text{prot}} \in \mathbb{R}^{L_{\text{prot}} \times d_v}$ and the peptide value matrix $V_{\text{pep}} \in \mathbb{R}^{L_{\text{pep}} \times d_v}$. Using the multihead coattention method, we can fuse the embedding vectors of protein and peptide to get better results.

Transformer-based decoder with crossattention

The cross-attention mechanism plays a pivotal role in our model, facilitating the interaction between the protein sequence matrix, denoted as *Q*, and the peptide sequence matrix, referred to as *KV*. This mechanism is integral to capturing the interdependencies between these two distinct biological sequences. The process can be mathematically formulated as follows:

$$\text{Crossattention}(Prot, Pep, Pep) = \text{softmax}\left(\frac{ProtPep^T}{\sqrt{d_k}}\right) Pep \quad \text{(Equation 5)}$$

Here, the matrix *Prot* represents the encoded representations of the protein sequences, whereas *Pep* embodies the combined key (*Pep*) and value (*Pep*) matrices, encoding the peptide sequences. The attention mechanism computes the dot product of the query *Prot* with the transpose of the key *Pep*, followed by scaling the result with the square root of the dimension of the key space ($\sqrt{d_k}$). This scaling is crucial for stabilizing the gradients during training, particularly when the dimensionality is high.

The resultant matrix is then passed through a softmax layer to obtain the attention weights, ensuring that they sum to one and highlight the most relevant features in the peptide sequences with respect to the protein queries. Finally, these weights are applied to the value matrix *Pep*, producing an output that is a weighted sum of the values based on the alignment between the protein and peptide sequences.

This cross-attention module is a critical component of our architecture, enabling the model to dynamically focus on specific parts of the peptide sequences that are most relevant to a given protein sequence. Such a mechanism is instrumental in learning the complex relationships and interactions between proteins and peptides, which is a fundamental aspect of our study.

MD simulations

The X-ray co-crystal structures of system in this study was extracted from PDB database (PDB ID: 6RMV). The LEaP program form AMBERTOOLS package was used to complement all absent hydrogen atoms of initial structures, and counter ions were added to insure the entire neutrality of this system. Finally, a truncated octahedral TIP3P water box was centered around the protein with 10 Ådistance. MD simulations for the studied system was performed using the GROMACS2021 program. Prior to MD simulations, this system was subjected to energy minimization process that contains a 5000 steps of steepest descent minimization followed by 5000 steps of conjugate gradient minimization. In the second, this system was heated gradually from 0 to 300 K in 500 ps. Third, each system was subjected to an equilibrium process for 1 ns under NPT ensemble. Then, the MD simulations for the studied system was preformed under NPT ensembles with periodic boundary conditions and 2 fs step time for 1000 ns. The data obtained from MD simulation was analyzed by CPPTRAJ module embedded in AMBERTOOLS. Figures in this study were generated with PyMOL and Chimera. The binding free energies ($\Delta Gbind$) for this system was computed using MM-PBSA method implemented in AMBERTOOLS with the following equations:

$$\Delta G_{\text{bind}} = G_{\text{complex}} - G_{\text{protein}} - G_{\text{ligand}} \quad \text{(Equation 6)}$$

$$G = E_{\text{MM}} + G_{\text{sol}} - TS \quad \text{(Equation 7)}$$

$$E_{\text{MM}} = E_{\text{int}} + E_{\text{ele}} + E_{\text{vdw}} \quad \text{(Equation 8)}$$

$$\Delta G_{\text{sol}} = G_{\text{PB}} + G_{\text{SA}} \quad \text{(Equation 9)}$$

Gprotein, Gligand, and Gcomplex in Equation 6 represent the free energy of the protein, ligand, and the protein-ligand complex, respectively. EMM, Gsol, and TS in Equation 7 represent the components of molecular mechanics in gas phase, the stabilization energy on account of solvation, and a vibrational entropy term, respectively. Eint, Eele, and Evdw are on behalf of the internal, Coulomb, and van der Waals interaction term, respectively. EMM is the sum of the terms of Eint, Eele, and Evdw. GPB and GSA are the polar and nonpolar contributions of the solvation free energy ($\Delta Gsol$). GPB computed using the Poisson-Boltzman (PB) model, whereas GSA is calculated on account of the solvent accessible surface area (SASA). In order to gain a detailed understanding of the interactions between protein with the ligand peptide,

the MM-PBSA decomposition analysis was applied to decompose the total binding free energy into ligand-residue pairs. Based on the decomposition energy results, all residues contributing smaller than -1 kcal/mol were considered as key residues.

QUANTIFICATION AND STATISTICAL ANALYSIS

In Figure 2E, we employed the T-test to conduct a differential analysis between PepCA and two other comparative methods. The sample sizes, denoted as 'n', for these tests were 92, 251, 639, and 125, respectively, corresponding to the number of complexes in the test set. We considered the differences to be statistically significant when the p-values were less than 0.05.