moPPIt: *De Novo* Generation of Motif-Specific Binders with Protein Language Models

Tong Chen,[1] **Yinuo Zhang,**[2] **Pranam Chatterjee**[1,2,3,†]

[1]Department of Biomedical Engineering, Duke University
[2]Department of Biostatistics and Bioinformatics, Duke University
[3]Department of Computer Science, Duke University

[†]Corresponding author: pranam.chatterjee@duke.edu

Abstract

The ability to precisely target specific motifs on disease-related proteins, whether conserved epitopes on viral proteins, intrinsically disordered regions within transcription factors, or breakpoint junctions in fusion oncoproteins, is essential for modulating their function while minimizing off-target effects. Current methods struggle to achieve this specificity without reliable structural information. In this work, we introduce a **mo**tif-specific **PPI t**argeting algorithm, **moPPIt**, for *de novo* generation of motif-specific peptide binders from the target protein sequence alone. At the core of moPPIt is BindEvaluator, a transformer-based model that interpolates protein language model embeddings of two proteins via a series of multi-headed self-attention blocks, with a key focus on local motif features. Trained on over 510,000 annotated PPIs, BindEvaluator accurately predicts target binding sites given protein-protein sequence pairs with a test AUC > 0.94, improving to AUC > 0.96 when fine-tuned on peptide-protein pairs. By combining BindEvaluator with our PepMLM peptide generator and genetic algorithm-based optimization, moPPIt generates peptides that bind specifically to user-defined residues on target proteins. We demonstrate moPPIt's efficacy in computationally designing binders to specific motifs, first on targets with known binding peptides and then extending to structured and disordered targets with no known binders. In total, moPPIt serves as a powerful tool for developing highly specific peptide therapeutics without relying on target structure or structure-dependent latent spaces.

Introduction

Motif-specific targeting of protein-protein interactions (PPIs) offers the potential for highly selective biotherapeutics that can modulate protein function while minimizing off-target effects, an advantage unattainable with traditional small molecule drugs, which typically require well-defined and conserved binding sites for inhibition [Lu et al., 2020]. The importance of targeting specific motifs is evident across a wide range of biological contexts. For instance, in cancer biology, restoring the function of the p53 tumor suppressor by targeting its DNA-binding domain could provide a powerful therapeutic approach in cancers where p53 is inactivated by mutations [Sullivan et al., 2017]. In neurodegenerative disorders like Alzheimer's disease, precise binding to the β-secretase cleavage site of the amyloid precursor protein (APP) could modulate its processing and potentially reduce the formation of toxic amyloid-β peptides [Kitazume et al., 2001]. Targeting active sites of enzymes, such as the catalytic domain of BRAF kinase in melanoma, offers more specific inhibition compared to traditional small molecule inhibitors [Castellani et al., 2023]. Allosteric domains present another important target, exemplified by the potential to modulate G protein-coupled receptor (GPCR) function by binding to their allosteric sites [Shpakov, 2023]. For intrinsically disordered proteins, targeting specific regions of the tau protein involved in pathological aggregation could provide new avenues for treating tauopathies [Chen et al., 2019]. Furthermore, in cancers driven by fusion oncoproteins,

such as PAX3::FOXO1 in alveolar rhabdomyosarcoma, targeting the unique sequence at the fusion breakpoint could offer exquisite specificity for therapeutic interventions [Linardic, 2008, Azorsa et al., 2021].

While experimental methods to generate motif-specific binders, such as animal immunization, phage display, and yeast display, are often prohibitively laborious, computational approaches offer a much more streamlined and efficient design process [Chen et al., 2023a]. Advances including AlphaFold and RFDiffusion, have shown promise in various protein design tasks, including motif-specific binder design [Jumper et al., 2021, Abramson et al., 2024, Watson et al., 2023, Bryant and Elofsson, 2023]. However, these methods operate purely in structure space, making them less suitable for targets lacking stable tertiary conformations, such as intrinsically disordered proteins, which were not present in their training sets. While recent efforts have attempted to extend diffusion-based methods to sample "plausible" conformations of disordered proteins via Gaussian perturbations [Liu et al., 2024], they remain constrained by their reliance on static structural data for training, which biases the underlying latent space, thus precluding accurate conformational sampling. An alternative approach leverages protein language models (pLMs) like ESM-2, ESM3, and ProtT5, which have been trained on vast, diverse protein sequence datasets to capture underlying physicochemical and functional properties of protein sequences, including disorder propensities [Lin et al., 2023, Elnaggar et al., 2022, Hayes et al., 2024, Vincoff et al., 2024]. These pLMs have demonstrated utility in various protein design tasks, including our recent work on designing target-specific peptide binders from target sequence alone [Brixi et al., 2023, Chen et al., 2023b, Bhat et al., 2023]. However, existing pLM-based methods have not yet focused on targeting specific motifs or epitopes on proteins, leaving a significant gap in our ability to design highly specific binders for conformationally and functionally diverse protein targets.

To address this gap, in this work, we develop a **mo**tif-specific **PPI t**argeting algorithm, termed **moPPIt**, that enables the design of motif-specific peptide binders using sequence-only pLM embeddings. To enable moPPIt-based generation, we train BindEvaluator, a transformer interpolating ESM-2 pLM embeddings [Lin et al., 2023] via a series of multi-headed self-attention blocks to capture both global and local interaction properties. Trained on over 510,000 annotated PPI sequence pairs, BindEvaluator accurately predicts binding hotspots between two proteins with a test AUC > 0.94, improving to AUC > 0.96 when fine-tuned on known peptide-protein pairs. moPPIt integrates BindEvaluator with our previous PepMLM peptide generation algorithm [Chen et al., 2023b], via a genetic optimization approach, to generate peptides that bind specifically to user-defined motifs on target proteins. We demonstrate moPPIt's efficacy in designing binders to specific epitopes on a diverse set of targets, including kinases, transcription factors, and even intrinsically disordered regions (IDRs). Using a combination of AlphaFold2-Multimer [Evans et al., 2021] and PeptiDerive, a Rosetta-based algorithm for identifying key binding residues [Sedan et al., 2016], we computationally validate the specificity and binding affinity of our designed peptides on targets with known peptide binders, as well as on novel structured targets and variable disordered domains. Our comprehensive approach allows moPPIt to specifically target motifs on a wide range of targets, including those previously considered "undruggable," potentially aiding drug discovery efforts for diseases driven by aberrant protein interactions.

Results

BindEvaluator accurately predicts target binding sites provided two interacting sequences

To enable motif-specific peptide binder generation, we first developed a BindEvaluator model to predict peptide-protein interaction binding sites (Figure 1A). Specifically, BindEvaluator takes a binder sequence and a target sequence as input and predicts the binding residues on the target protein. Both sequences are first processed by the pre-trained ESM-2-650M model to generate embeddings [Lin et al., 2023]. These embeddings are then refined using multiple multi-head attention modules to capture sequence information. A reciprocal multi-head attention layer then integrates these processed embeddings to learn global interaction dependencies. Finally, feed-forward and linear layers analyze the integrated embeddings to predict interaction hotspots on the target protein (Figure 1A).

We initially trained BindEvaluator on a large protein-protein interaction (PPI) dataset containing over 500,000 entries with annotated interface residues [Bushuiev et al., 2023] to provide foundational knowledge of protein interaction information. During training, we observed a consistent decline in the validation loss, which indicates stable and effective learning (Figure S1A). The steady decrease in binary cross entropy (BCE) loss and Kullback-Leibler (KL) divergence loss suggested that the model improves in distinguishing between

binding and non-binding residues and in understanding the fundamental distribution of binding sites. The model's performance on the test data further confirmed its efficacy, achieving an accuracy of 0.83, an area under the ROC curve (AUC) of 0.93, an F1 score of 0.65, and a Matthews correlation coefficient (MCC) of 0.59 (Figure 1B).

We hypothesized that incorporating dilated convolutional neural network (CNN) modules into BindEvaluator would enhance its performance by effectively extracting local features relevant to binding site information (Figure 1A). To test this hypothesis, we trained a version of BindEvaluator with dilated CNN modules on the same PPI dataset, holding all other training settings constant except for slightly different gradient accumulation schedules compared to the original model without dilated CNN modules. The inclusion of these CNN modules led to observable improvements across several metrics (Figure 1B). Specifically, the adjusted model achieved a lower test loss than the model without these modules. Additionally, the decrease in BCE loss indicates stronger binding site classification performance. Although the KL divergence loss showed a slight increase from 0.773 to 0.776, this did not significantly impact the overall model performance. Both models, with and without dilated CNN modules demonstrated similar declining trends in their loss curves, indicating effective learning (Figure S1B). Notably, the total loss continued to decrease even in the final training epochs, suggesting that the BindEvaluator with dilated CNN modules was more adept at learning subtle features, leading to better performance.

To adapt our model for peptide-protein binding site prediction, the BindEvaluator model with dilated CNN modules, initially trained on PPI data, was further fine-tuned on over 12,000 structurally validated, non-redundant peptide-protein sequence pairs. We observed validation loss curves steadily improving during fine-tuning (Figure S1). The fine-tuned model demonstrated notably lower Kullback-Leibler (KL) divergence loss in peptide-protein binding region predictions compared to its performance on PPI interaction prediction, and achieved strong test metrics, indicating high precision in binding site prediction (Figure 1B).

moPPIt generates epitope-specific binders to target proteins

With the fine-tuned BindEvaluator for peptide-protein binding site prediction in hand, we turned our attention to peptide generation by developing the **mo**tif-specific **PPI t**argeting algorithm (**moPPIt**) to generate motif-specific peptide binders based solely on target protein sequences. Rather than sampling random sequences through BindEvaluator, we decided to employ our recent PepMLM model, a state-of-the-art ESM-2-based model that generates peptide binders conditioned on a target protein sequence alone [Chen et al., 2023b]. As such, we sought to optimize PepMLM-generated peptides to bind to specific motifs on a target protein. To accomplish this, moPPIt leverages a genetic algorithm, integrating the pre-trained PepMLM and fine-tuned BindEvaluator, to produce binders with high specificity to user-defined residues (Figure 2).

The process begins with PepMLM generating a pool of candidate peptide binders of a defined length for a given target protein sequence. BindEvaluator then predicts the interacting residues between each candidate binder and the target protein. A penalty score is assigned to each binder based on these predictions (see Methods). Additionally, PepMLM computes the perplexity of each peptide given the target sequence, which serves as another metric to evaluate the biological relevance of the binders. The candidate binders are then sorted based on penalty scores and perplexity. Subsequently, a new pool of candidate binders is generated from the sorted binders via a genetic algorithm, aiming for lower penalty scores or perplexity (see Methods). This new pool undergoes another round of evaluation by BindEvaluator and PepMLM, and the process is repeated. The iteration continues until the penalty score falls below a set threshold, the maximum number of genetic algorithm rounds is reached, or there is no further improvement in successive rounds. The resulting top binders are expected to exhibit high affinity and specificity for the specified binding motifs on the target protein.

To evaluate moPPIt in a well-controlled setting, we designed binders for 15 structured, unseen proteins with known, pre-existing peptide binders, all derived from the PDB. We calculated the ipTM and pTM scores, which represent confidence in interface formation and overall complex formation, respectively, for the peptide-protein complexes predicted by AlphaFold2-Multimer, comparing the performance of the pre-existing peptides to the ones designed by moPPIt [Evans et al., 2021]. We observed that moPPIt-designed binders can form peptide-protein complexes with similar or superior ipTM and pTM scores compared to the pre-existing ones (Table S1). Notably, only one of the 15 designed peptides fell slightly below the defined ipTM threshold, set at 0.05 below the ipTM score of the existing peptide-protein complex (Figure 2). The superior ipTM scores underscore moPPIt's capability to generate peptides with strong binding to target proteins. Moreover, moPPIt successfully designed binders of varying lengths, demonstrating its overall versatility (Table S1).

We further analyzed the relative interface scores (RIS) of both existing and designed peptide-protein complexes using PeptiDerive [Sedan et al., 2016], which evaluates the energy contribution of specific residues to the overall free energy of the binder-target complex structure (Figure 3, S3). The designed complexes showed similar or higher RIS at specified binding positions compared to existing complexes, indicating similar or stronger binding potential. Additionally, the residues with high RIS were primarily localized in regions adjacent to the binding motifs, showcasing the high specificity of moPPIt-designed binders.

moPPIt generates epitope-specific binders for structured proteins without known binders

To further assess moPPIt's performance, we designed peptide binders for structured proteins without pre-existing binders. We specifically selected proteins from three enzyme classes (kinases, phosphatases, and deubiquitinases) to evaluate moPPIt's versatility in designing binders for diverse structured proteins without pre-identified binding sites. We first identified potential binding sites for these proteins. We utilized PepMLM to generate 50 candidate binders for each protein. BindEvaluator then predicted the binding sites on the target proteins for the top three binders with the lowest perplexity. The binding residues were identified as those present in all three predictions.

We confirm that moPPIt-designed peptides, when co-folded with target proteins, exhibit high pTM and ipTM scores, indicating that the designed peptides can form structured complexes with target proteins (Table S2). Having confirmed binding capacity, we next evaluated the epitope specificity of designed binders to corresponding targets (Figure 4, S4). The structure of three example targets, CLK1, PPP5, and MINDY1, co-folded with their designed binders, are shown in Figure 4. Notably, the residues with the highest RIS predicted by PeptiDerive are at the specified binding motifs, indicating moPPIt's capability to generate highly specific binders. The 3D structures of the peptide-protein complexes show the designed peptides positioned close to the target binding sites, further validating moPPIt's capacity to produce binders with high affinity for the target motifs.

moPPIt generates binders targeting intrinsically disordered proteins

The true utility of using pLMs as opposed to structure-based models such as RFDiffusion, lies in the ability to bind to conformationally diverse targets and regions. To demonstrate this, we selected three proteins with structurally disordered domains (UCHL5, 4E-BP2, and EWS::FLI1) and designed binders for them using moPPIt. Table S3 displays the pTM and ipTM scores for each complex structure formed with the designed binders, along with their binding sites and target proteins' disordered regions predicted by DisorderUnetLM [Kotowski et al., 2024]. For UCHL5, we targeted the binder to one of its disordered regions, achieving a high ipTM score and pTM score, similar to the binder for 4E-BP2. For EWS::FLI, we designed binders to its structured domain so as to demonstrate that moPPIt is able to design binders to the structured regions of intrinsically disordered proteins. The high ipTM scores of the three peptide-protein complexes demonstrate the strong binding capacity of the designed peptides (Table S3).

We next conducted an analysis of binding site specificity. Remarkably, the PeptiDerive scores align with the specified binding motifs, showing high predicted RIS (Figure 5). The visualizations of the 3D predicted structures reveal that the designed peptides are positioned close to the target motifs, suggesting a high affinity for the intrinsically disordered domains. These alignments indicate that moPPIt can design binders targeting both conformationally ordered and disordered regions of structurally disordered proteins. Although BindEvaluator and PepMLM were primarily trained on sequences from solved peptide-protein structures [Chen et al., 2023b], and tools like AlphaFold2-Multimer and PeptiDerive may not be fully reliable for disordered regions, the consistency between the target motifs and PeptiDerive's predictions, as well as 3D complex structures, provides a high degree of confidence in moPPIt-designed binders. Nevertheless, experimental validation is necessary to confirm whether these binders will effectively bind to the specified disordered regions.

Discussion

The challenge of designing highly specific peptide binders, particularly for targets lacking well-defined structural pockets or those with intrinsically disordered regions, has long been a bottleneck in therapeutic development. In this work, we have presented moPPIt, a purely sequence-based approach that addresses this challenge by enabling the design of motif-specific peptide binders without relying or interpolating on structural representations. By integrating feature-rich pLM embeddings, moPPIt demonstrates the ability to

generate peptides that bind to user-defined epitopes across a diverse range of protein targets, those with both structured and conformationally flexible motifs.

We believe moPPIt has the potential to be effective across a broad spectrum of protein targets. To prove this, our next steps will include a comprehensive experimental validation of moPPIt, alongside structure-based methods like RFDiffusion [Watson et al., 2023, Liu et al., 2024], evaluating performance on both structured and disordered regions. This will involve biochemical binding affinity assays and leveraging our chimeric peptide-E3 ubiquitin ligase ubiquibody (uAb) architecture for target degradation studies [Brixi et al., 2023, Chen et al., 2023b, Bhat et al., 2023]. Furthermore, the motif-specific nature of our approach suggests promising applications in developing binders with mutant selectivity and the ability to target specific post-translational modification sites [Peng et al., 2024]. Importantly, moPPIt's capability to target specific epitopes could be particularly valuable in interrogating viral proteins, such as those of SARS-CoV-2 and future pandemic viruses, by enabling the design of binders that target highly conserved regions less prone to escape mutations [Abbasian et al., 2023]. Overall, these capabilities could prove invaluable for both detection and therapeutic applications, potentially enabling more precise modulation of protein function in complex diseases that are driven by aberrant post-translational states. As we move forward with experimental validation, we anticipate that moPPIt will contribute significantly to expanding the repertoire of targetable proteins and advancing the field of precision biotherapeutics.

Methods

Dataset Curation

The training data for BindEvaluator was curated from the PPIRef dataset, a large and non-redundant databank of PPIs [Bushuiev et al., 2023]. To augment the dataset, additional entries were generated by reversing the roles of the target and binder sequences for each original entry. Proteins exceeding 500 amino acids were removed due to GPU constraints. After removing all duplicates, the final dataset comprised 510,804 triplets, each containing target sequence, binder sequence, and binding motifs. This dataset was split at a 60/20/20 ratio into a training set, validation set, and test set.

The peptide-protein interaction data for fine-tuning BindEvaluator was curated from the PepNN and BioLip2 databases [Abdin et al., 2022, Zhang et al., 2024]. Specifically, 3022 PepNN and 9251 BioLip2 non-redundant triplets for peptide-protein binding were collected. Proteins longer than 500 amino acids and peptides longer than 25 amino acids were removed. The dataset was split at a 80/10/10 ratio into a training set, validation set, and test set.

BindEvaluator Model Architecture

The generation algorithm is based on the BindEvaluator model. As shown in Figure 1, BindEvaluator takes a binder sequence and a target sequence as inputs to predict the binding residues on the target protein. The design of this model draws inspiration from the architectures of PepNN and Pseq2Sites, which have demonstrated effectiveness in similar tasks [Abdin et al., 2022, Seo et al., 2024].

Both binder and target sequences are first passed into the pre-trained ESM-2-650M model to obtain their embeddings [Lin et al., 2023]. For the target sequence, a dilated CNN module captures the local features of adjacent residues. Specifically, the module is composed of three stacked CNN blocks with different dilation rates (1, 2, and 3) to extract hierarchical features. Each block consists of three convolutional layers with different kernel widths (3, 5, and 7) to cover different receptive field sizes, accommodating different binding site sizes. Padding is added to each convolutional layer to maintain consistent output and input sizes. Since the focus is to identify binding residues for the target protein, the dilated CNN module is applied only to the target sequence. Given that no binding motifs in the training set contain more than 23 continuous residues, the dilated CNN module is sufficient to capture the binding region features.

The processed embeddings are then passed through multi-head attention modules to capture global dependencies for each residue. In the reciprocal attention modules, the target and binder sequence representations are integrated to capture binder-target interaction information. Specifically, in these modules, the binder representations are projected into a key matrix K and a query matrix Q, while the target representations are projected into a value matrix V, and vice versa. The reciprocal attention is formulated as follows:

$$\text{Attention}_{\text{target}}(Q, K, V_{\text{binder}}) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V_{\text{binder}} \tag{1}$$

$$\text{Attention}_{\text{binder}}(Q, K, V_{\text{target}}) = \text{softmax}\left(\frac{KQ^T}{\sqrt{d_k}}\right) V_{\text{target}} \tag{2}$$

where d_k is the model dimension.

Following several layers of dilated CNN and attention modules, the resulting target sequence representation encapsulates the binder-target binding information. Finally, this representation is processed by feed-forward layers and linear layers to predict the binding sites.

Model Training and Fine-Tuning

BindEvaluator is first trained on a PPI dataset and then fine-tuned using peptide-protein binding data. During training and fine-tuning, the same model architecture is used. The weights of ESM-2-650M are fixed, and all other parameters remain trainable. To accurately capture the intrinsic distribution of binding residues, the loss function L is designed to be the sum of the Binary Cross-Entropy (BCE) loss and the Kullback-Leibler (KL) divergence between the predicted and the true binding motifs. Specifically, letting $\hat{y}$ be the predicted binding motifs and y be the true binding motifs, the loss function is defined as:

$$L(y, \hat{y}) = -\sum_i \left[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)\right] + \lambda \sum_i y_i \log\left(\frac{y_i}{\hat{y}_i}\right) \tag{3}$$

Here, λ is a hyper-parameter that balances the contribution of the KL divergence to the total loss. During training, λ is set to 0.1, while during fine-tuning, λ is set to 1.

BindEvaluator was trained on a 6xA6000 NVIDIA RTX GPU system with 48 GB of VRAM each for 30 epochs. The batch size was set to 32, with a learning rate of 1e-3, a dropout rate of 0.3, and a gradient clipping value of 0.5. The AdamW optimizer was used with weight decay. Fine-tuning was performed on the same six GPUs for 30 epochs, with an increased dropout rate of 0.5. The batch size, learning rate, gradient clipping, and optimizer settings were identical to those used during training.

Motif-Specific Binder Design Algorithm

The moPPIt algorithm aims to generate motif-specific peptide binders based on the target protein sequence, leveraging the PepMLM algorithm: `https://huggingface.co/ChatterjeeLab/PepMLM-650M` The visualization of the moPPIt algorithm is shown in Figure 2.

Given a target protein sequence, PepMLM generates a pool of candidate peptide binders of a specified length. BindEvaluator then predicts the interacting residues of a candidate binder to the target protein. A penalty score is assigned based on these predictions. Specifically, for each amino acid in the specified binding motifs, but not in the predicted binding sites, the penalty score increases by 1. In contrast, for each amino acid in the predicted but not the specified binding motifs, the penalty score increases by 0.5. The scoring system ensures that the generated binders target the specified motifs with high selectivity. Additionally, as a masked language model, PepMLM computes the perplexity of each peptide based on the target protein sequence. The perplexity (PPL) of the peptide sequence given the target protein is defined as:

$$PPL = \exp\left(-\frac{1}{L}\sum_{i=1}^{L} \log P(a_i | T, a_{<i})\right) \tag{4}$$

where L is the number of amino acids in the peptide, a_i is the i-th amino acid in the peptide sequence, T represents the target protein, and $P(a_i|T, a_{<i})$ is the probability of the i-th amino acid given the target protein and the preceding amino acids in the peptide sequence.

After computing the penalty scores and perplexity, candidate binders are sorted based on these metrics. A genetic algorithm is then applied to the sorted binders. Specifically, the top 10% of binder sequences remain unchanged. For the remaining 90%, new binders are created by randomly selecting and mating two sequences from the top half of the original pool. During mating, each position on the binder sequence has a 45% chance

of inheriting the amino acid from one parent sequence, a 45% chance from the other parent, and a 10% chance of being replaced by a new amino acid. This process generates a new pool of candidate binders.

The new pool undergoes another round of evaluation using BindEvaluator and PepMLM to compute penalty scores and perplexity. After sorting based on the metrics, these binders are forwarded to the next round of the genetic algorithm. This iterative process continues until the penalty score falls below a certain threshold, the maximum number of genetic algorithm rounds is reached, or no further improvement is observed in successive rounds. In the benchmarking tasks presented in this paper, the threshold is set to be one-tenth of the number of specified binding positions, rounded to the nearest integer. The resulting top binders are expected to bind to the specified motifs on the target protein with high affinity and specificity.

In Silico Benchmarking

All proteins and PDB IDs used for benchmarking can be found in the supplementary tables. Complex structures of peptide and target proteins, and their associated pTM and ipTM scores, were predicted using the AlphaFold2-Multimer v3 model in a locally installed version of ColabFold [Mirdita et al., 2022]. The PeptiDerive algorithm (`https://rosie.rosettacommons.org/peptiderive`) was used to calculate the relative interaction score (RIS) for each position on the target protein based on the output AlphaFold2-Multimer structure [Sedan et al., 2016].

Declarations

Acknowledgements

Research reported in this manuscript was supported by the National Institute of General Medical Sciences (NIGMS) of the National Institutes of Health (NIH) under award number R35GM155282. We thank Shrey Goel for assistance in visualizing motif interactions.

Author Contributions

T.C. designed and evaluated BindEvaluator and moPPIt, and performed model benchmarking and visualizations. Y.Z. curated and processed the PPIRef dataset for training. P.C. conceived, designed, directed, and supervised the study.

Data and Materials Availability

All sequences and data needed to evaluate the conclusions are presented in the paper and tables. All code to train moPPIt and generate peptides are accessible by the academic community at `https://huggingface.co/ChatterjeeLab/moPPIt`, after signing a free, non-commercial, research-only academic license.

Competing Interests

P.C. is a listed inventor on numerous patents related to peptide binder design with protein language models. P.C.'s interests are reviewed and managed by Duke University in accordance with their conflict-of-interest policies. T.C. and Y.Z. declare no competing interests.

References

[Abbasian et al., 2023] Abbasian, M. H., Mahmanzar, M., Rahimian, K., Mahdavi, B., Tokhanbigli, S., Moradi, B., Sisakht, M. M., and Deng, Y. (2023). Global landscape of sars-cov-2 mutations and conserved regions. *Journal of Translational Medicine*, 21(1).

[Abdin et al., 2022] Abdin, O., Nim, S., Wen, H., and Kim, P. M. (2022). Pepnn: a deep attention model for the identification of peptide binding sites. *Communications biology*, 5(1):503.

[Abramson et al., 2024] Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., Bodenstein, S. W., Evans, D. A., Hung, C.-C., O'Neill, M., Reiman, D., Tunyasuvunakool, K., Wu, Z., Žemgulytė, A., Arvaniti, E., Beattie, C., Bertolli, O., Bridgland, A., Cherepanov, A., Congreve, M., Cowen-Rivers, A. I., Cowie, A., Figurnov, M., Fuchs, F. B., Gladman, H., Jain, R., Khan, Y. A., Low, C. M. R., Perlin, K., Potapenko, A., Savy, P., Singh, S., Stecula, A.,

Thillaisundaram, A., Tong, C., Yakneen, S., Zhong, E. D., Zielinski, M., Žídek, A., Bapst, V., Kohli, P., Jaderberg, M., Hassabis, D., and Jumper, J. M. (2024). Accurate structure prediction of biomolecular interactions with alphafold3. *Nature*.

[Azorsa et al., 2021] Azorsa, D. O., Bode, P. K., Wachtel, M., Cheuk, A. T. C., Meltzer, P. S., Vokuhl, C., Camenisch, U., Khov, H. L., Bode, B., Schäfer, B. W., and Khan, J. (2021). Immunohistochemical detection of pax-foxo1 fusion proteins in alveolar rhabdomyosarcoma using breakpoint specific monoclonal antibodies. *Modern Pathology*, 34(4):748–757.

[Bhat et al., 2023] Bhat, S., Palepu, K., Hong, L., Mao, J., Ye, T., Iyer, R., Zhao, L., Chen, T., Vincoff, S., Watson, R., Wang, T., Srijay, D., Kavirayuni, V. S., Kholina, K., Goel, S., Vure, P., Desphande, A. H., Soderling, S., DeLisa, M., and Chatterjee, P. (2023). De novo design of peptide binders to conformationally diverse targets with contrastive language modeling. *bioRxiv*.

[Brixi et al., 2023] 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&peppr is an interface-predicting language model for designing peptide-guided protein degraders. *Communications Biology*, 6(1):1081.

[Bryant and Elofsson, 2023] Bryant, P. and Elofsson, A. (2023). Peptide binder design with inverse folding and protein structure prediction. *Communications Chemistry*, 6(1).

[Bushuiev et al., 2023] Bushuiev, A., Bushuiev, R., Kouba, P., Filkin, A., Gabrielova, M., Gabriel, M., Sedlar, J., Pluskal, T., Damborsky, J., Mazurenko, S., and Sivic, J. (2023). Learning to design protein-protein interactions with enhanced generalization.

[Castellani et al., 2023] Castellani, G., Buccarelli, M., Arasi, M. B., Rossi, S., Pisanu, M. E., Bellenghi, M., Lintas, C., and Tabolacci, C. (2023). Braf mutations in melanoma: Biological aspects, therapeutic implications, and circulating biomarkers. *Cancers*, 15(16):4026.

[Chen et al., 2019] Chen, D., Drombosky, K. W., Hou, Z., Sari, L., Kashmer, O. M., Ryder, B. D., Perez, V. A., Woodard, D. R., Lin, M. M., Diamond, M. I., and Joachimiak, L. A. (2019). Tau local structure shields an amyloid-forming motif and controls aggregation propensity. *Nature Communications*, 10(1).

[Chen et al., 2023a] Chen, T., Hong, L., Yudistyra, V., Vincoff, S., and Chatterjee, P. (2023a). Generative design of therapeutics that bind and modulate protein states. *Current Opinion in Biomedical Engineering*, 28:100496.

[Chen et al., 2023b] Chen, T., Pertsemlidis, S., Watson, R., Kavirayuni, V. S., Hsu, A., Vure, P., Pulugurta, R., Vincoff, S., Hong, L., Wang, T., et al. (2023b). Pepmlm: Target sequence-conditioned generation of peptide binders via masked language modeling. *ArXiv*.

[Elnaggar et al., 2022] Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., and Rost, B. (2022). Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127.

[Evans et al., 2021] Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Žídek, A., Bates, R., Blackwell, S., Yim, J., Ronneberger, O., Bodenstein, S., Zielinski, M., Bridgland, A., Potapenko, A., Cowie, A., Tunyasuvunakool, K., Jain, R., Clancy, E., Kohli, P., Jumper, J., and Hassabis, D. (2021). Protein complex prediction with alphafold-multimer. *bioRxiv*.

[Hayes et al., 2024] Hayes, T., Rao, R., Akin, H., Sofroniew, N. J., Oktay, D., Lin, Z., Verkuil, R., Tran, V. Q., Deaton, J., Wiggert, M., Badkundri, R., Shafkat, I., Gong, J., Derry, A., Molina, R. S., Thomas, N., Khan, Y., Mishra, C., Kim, C., Bartie, L. J., Nemeth, M., Hsu, P. D., Sercu, T., Candido, S., and Rives, A. (2024). Simulating 500 million years of evolution with a language model. *bioRxiv*.

[Jumper et al., 2021] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. (2021). Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589.

[Kitazume et al., 2001] Kitazume, S., Tachida, Y., Oka, R., Shirotani, K., Saido, T. C., and Hashimoto, Y. (2001). Alzheimer's -secretase, -site amyloid precursor protein-cleaving enzyme, is responsible for cleavage secretion of a golgi-resident sialyltransferase. *Proceedings of the National Academy of Sciences*, 98(24):13554–13559.

[Kotowski et al., 2024] Kotowski, K., Roterman, I., and Stapor, K. (2024). Protein intrinsic disorder prediction using attention u-net and prottrans protein language model. *arXiv preprint arXiv:2404.08108*.

[Lin et al., 2023] 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(6637):1123–1130.

[Linardic, 2008] Linardic, C. M. (2008). Pax3–foxo1 fusion gene in rhabdomyosarcoma. *Cancer Letters*, 270(1):10–18.

[Liu et al., 2024] Liu, C., Wu, K., Choi, H., Han, H., Zhang, X., Watson, J. L., Shijo, S., Bera, A. K., Kang, A., Brackenbrough, E., Coventry, B., Hick, D. R., Hoofnagle, A. N., Zhu, P., Li, X., Decarreau, J., Gerben, S. R., Yang, W., Wang, X., Lamp, M., Murray, A., Bauer, M., and Baker, D. (2024). Diffusing protein binders to intrinsically disordered proteins. *bioRxiv*.

[Lu et al., 2020] Lu, H., Zhou, Q., He, J., Jiang, Z., Peng, C., Tong, R., and Shi, J. (2020). Recent advances in the development of protein–protein interactions modulators: mechanisms and clinical trials. *Signal Transduction and Targeted Therapy*, 5(1).

[Mirdita et al., 2022] Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. (2022). Colabfold: making protein folding accessible to all. *Nature Methods*, 19(6):679–682.

[Peng et al., 2024] Peng, Z., Schussheim, B., and Chatterjee, P. (2024). Ptm-mamba: A ptm-aware protein language model with bidirectional gated mamba blocks. *bioRxiv*.

[Sedan et al., 2016] Sedan, Y., Marcu, O., Lyskov, S., and Schueler-Furman, O. (2016). Peptiderive server: derive peptide inhibitors from protein–protein interactions. *Nucleic Acids Research*, 44(W1):W536–W541.

[Seo et al., 2024] Seo, S., Choi, J., Choi, S., Lee, J., Park, C., and Park, S. (2024). Pseq2sites: Enhancing protein sequence-based ligand binding-site prediction accuracy via the deep convolutional network and attention mechanism. *Engineering Applications of Artificial Intelligence*, 127:107257.

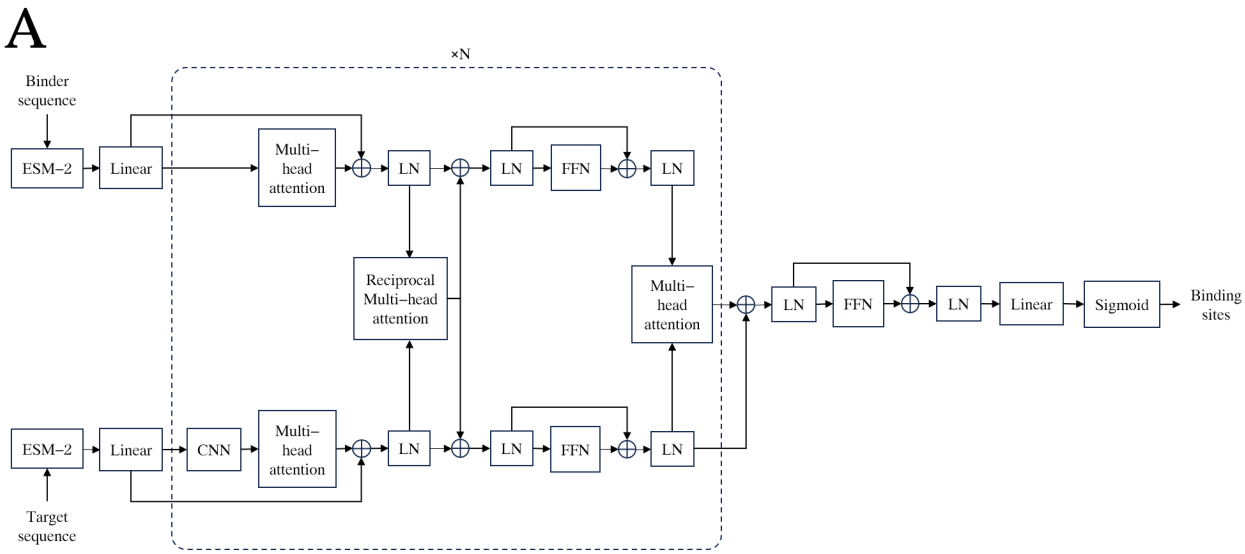
[Shpakov, 2023] Shpakov, A. O. (2023). Allosteric regulation of g-protein-coupled receptors: From diversity of molecular mechanisms to multiple allosteric sites and their ligands. *International Journal of Molecular Sciences*, 24(7):6187.

[Sullivan et al., 2017] Sullivan, K. D., Galbraith, M. D., Andrysik, Z., and Espinosa, J. M. (2017). Mechanisms of transcriptional regulation by p53. *Cell Death amp; Differentiation*, 25(1):133–143.

[Vincoff et al., 2024] Vincoff, S., Goel, S., Kholina, K., Pulugurta, R., Vure, P., and Chatterjee, P. (2024). Fuson-plm: A fusion oncoprotein-specific language model via focused probabilistic masking. *bioRxiv*.

[Watson et al., 2023] Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Hanikel, N., Pellock, S. J., Courbet, A., Sheffler, W., Wang, J., Venkatesh, P., Sappington, I., Torres, S. V., Lauko, A., De Bortoli, V., Mathieu, E., Ovchinnikov, S., Barzilay, R., Jaakkola, T. S., DiMaio, F., Baek, M., and Baker, D. (2023). De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100.

[Zhang et al., 2024] Zhang, C., Zhang, X., Freddolino, P. L., and Zhang, Y. (2024). Biolip2: an updated structure database for biologically relevant ligand–protein interactions. *Nucleic Acids Research*, 52(D1):D404–D412.



B

Test Metric	Train w/o CNN	Train w/ CNN	Fine-tune w/ CNN
Loss	0.388	0.373	0.514
BCE Loss	0.311	0.295	0.580
KL Loss	0.773	0.776	0.254
Accuracy	0.83	0.84	0.91
AUC	0.93	0.94	0.97
F1 Score	0.65	0.66	0.58
MCC	0.59	0.61	0.59

Figure 1: **BindEvaluator. (A)** Overview of the architecture of BindEvaluator. BindEvaluator predicts the binding residues on the target protein given a target sequence and a binder sequence. The binder and target sequences are first processed using a pre-trained ESM-2-650M model to obtain their embeddings. The target sequence embeddings are further refined using a dilated CNN module to capture local features. Both embeddings are then passed through multi-head attention modules to capture global dependencies. Reciprocal multi-head attention modules integrate the representations of the target and binder sequences, allowing for the capture of binder-target interaction information. Feed-forward and linear layers subsequently process the refined embeddings to predict the binding sites. **(B)** Test performance metrics of BindEvaluator across different training configurations. Performance metrics were calculated for BindEvaluator across three configurations: trained without dilated CNN modules, trained with dilated CNN modules, and fine-tuned for peptide-protein binding site prediction. Metrics include overall loss, binary cross-entropy (BCE) loss, KL divergence loss, accuracy, area under the ROC curve (AUC), F1 score, and Matthews correlation coefficient (MCC).

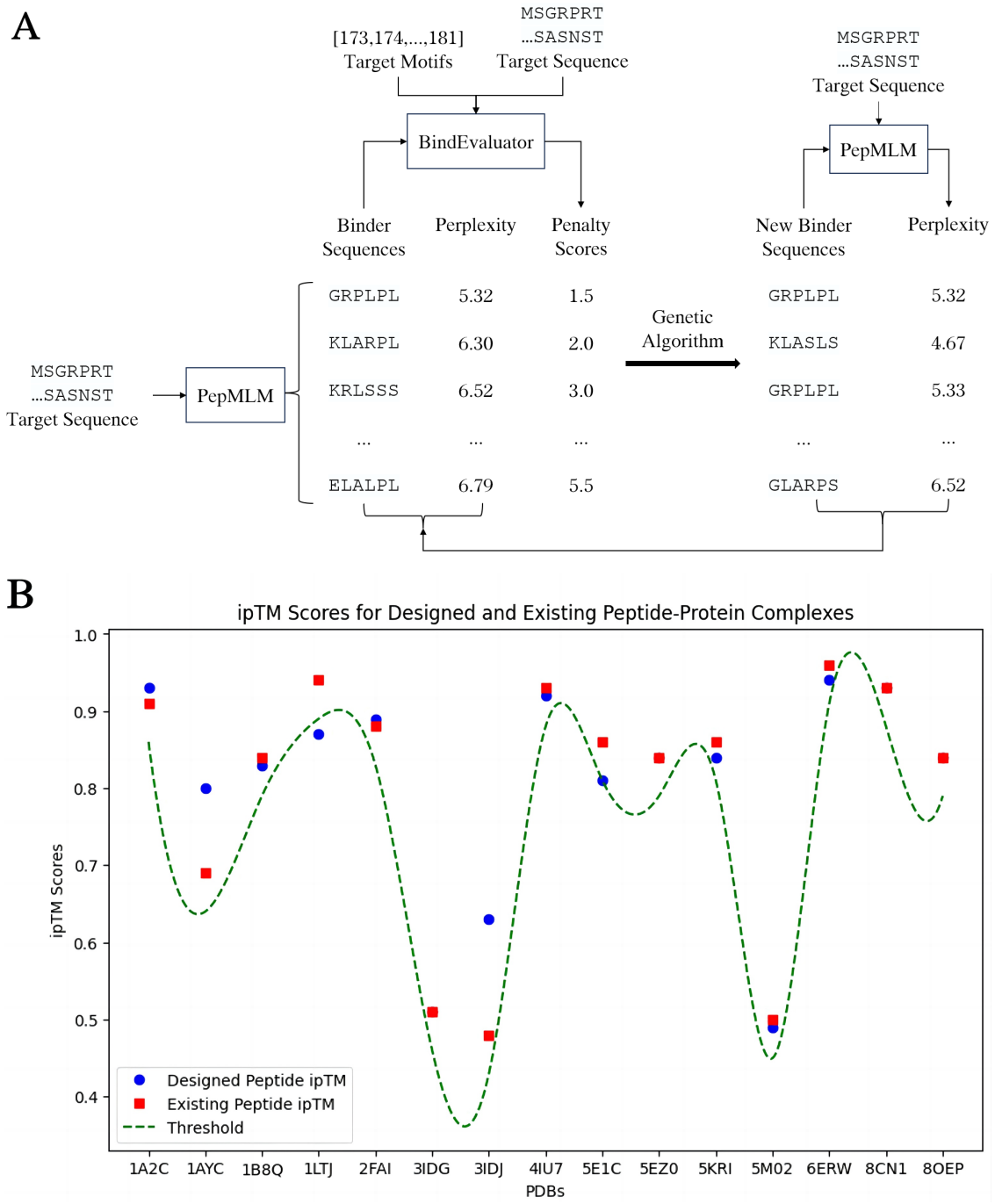


Figure 2: **moPPIT. (A)** Schematic of moPPIt. The algorithm starts with a target protein sequence input, generating initial binder sequences using PepMLM and calculating their perplexity. BindEvaluator predicts binding residues and assigns penalty scores. The binders are refined via a genetic algorithm, iterating until stopping criteria are met, as detailed in the main text. **(B)** Hit rate of moPPIt on structured targets with known binders. The ipTM scores of input peptides, in complex with their target protein, were calculated via AlphaFold-Multimer. The ipTM scores for known peptides (red) from PDB structures were compared to moPPIt-designed peptides (blue) for the same target proteins. An ipTM below 0.05 of the existing peptide for a given target protein (green line) was used as a threshold to call hits.

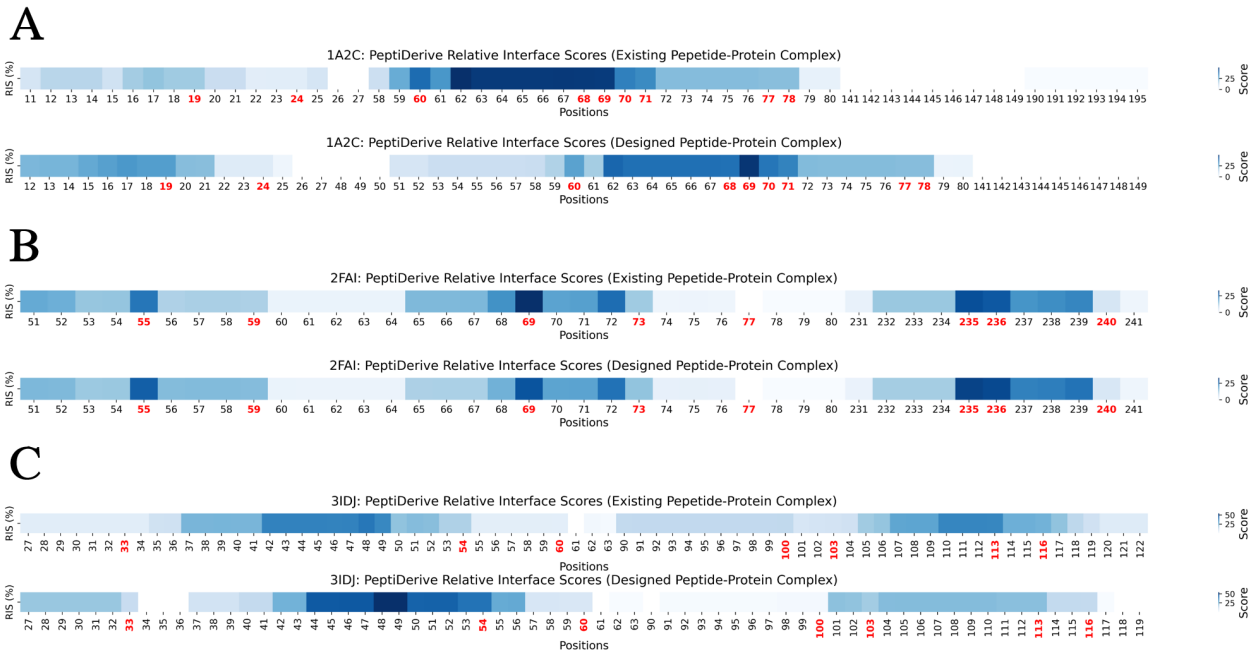


Figure 3: **PeptiDerive relative interface scores for existing and designed peptide-protein complexes.** The PeptiDerive relative interface scores (RIS) for existing and designed peptide-protein complexes were computed and visualized for three example proteins with PDB IDs: **(A)** 1A2C, **(B)** 2FAI, and **(C)** 3IDJ. The first heatmap for each protein shows the RIS of the existing peptide-protein complex, while the second heatmap shows the scores for the designed peptide-protein complex. For each heatmap, the x-axis indicates the residue positions, with highlighted positions in red representing the target binding amino acids that were input into moPPIt. High RIS at these positions indicate strong binding potential.

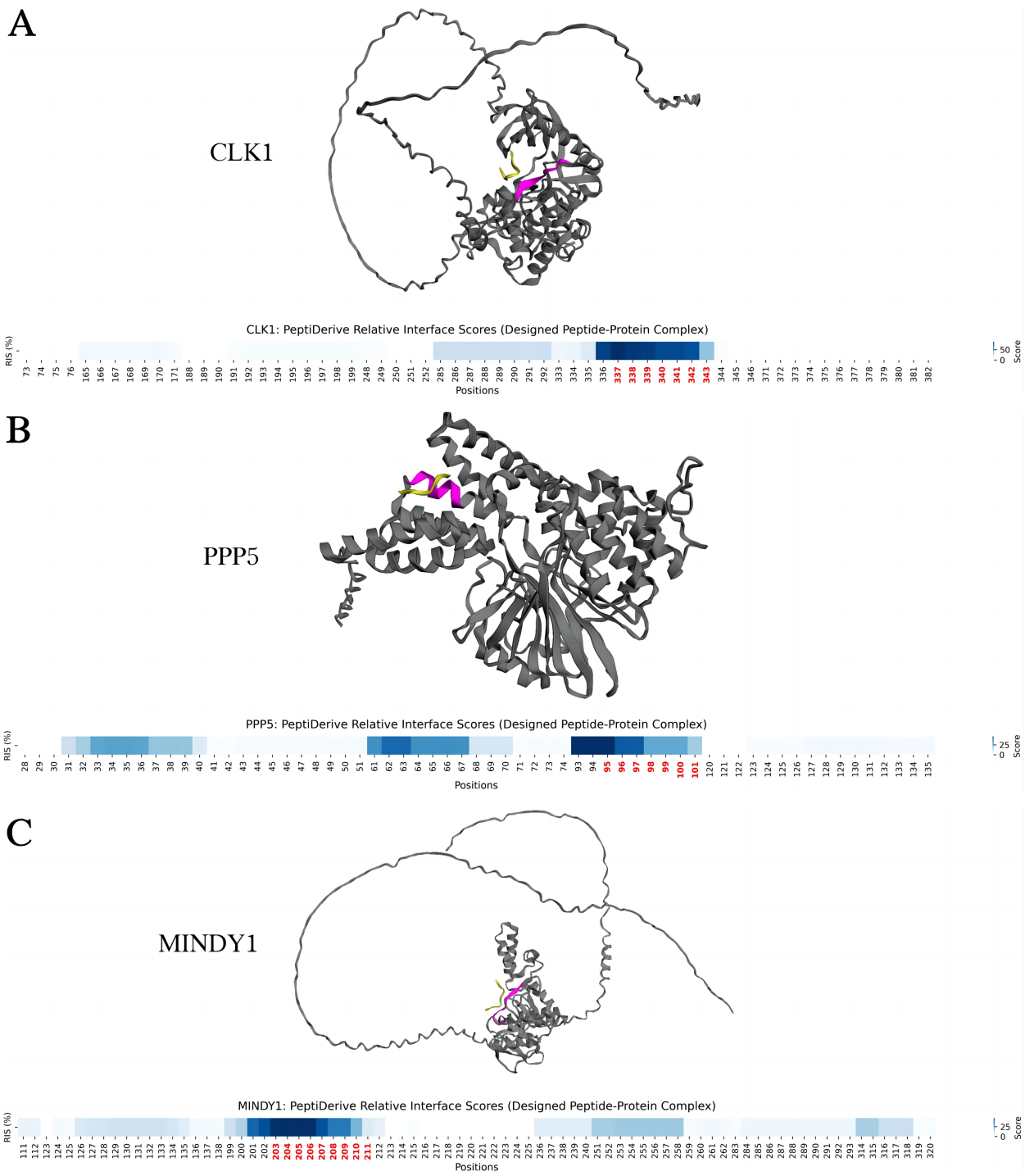


Figure 4: **Structural visualization and PeptiDerive relative interaction scores for designed peptides targeting structured motifs.** The peptide-complex structures are visualized for three proteins without known binders: **(A)** CLK1, **(B)** PPP5, and **(C)** MINDY1. The target proteins are depicted in grey, the designed peptide binders are shown in yellow, and the binding residues specified by the moPPIt algorithm are highlighted in magenta. Below each structure, the relative interaction scores (RIS) computed by PeptiDerive are shown, with high scores indicating strong binding potential. Positions highlighted in red were input into moPPIt as the desired target amino acids.

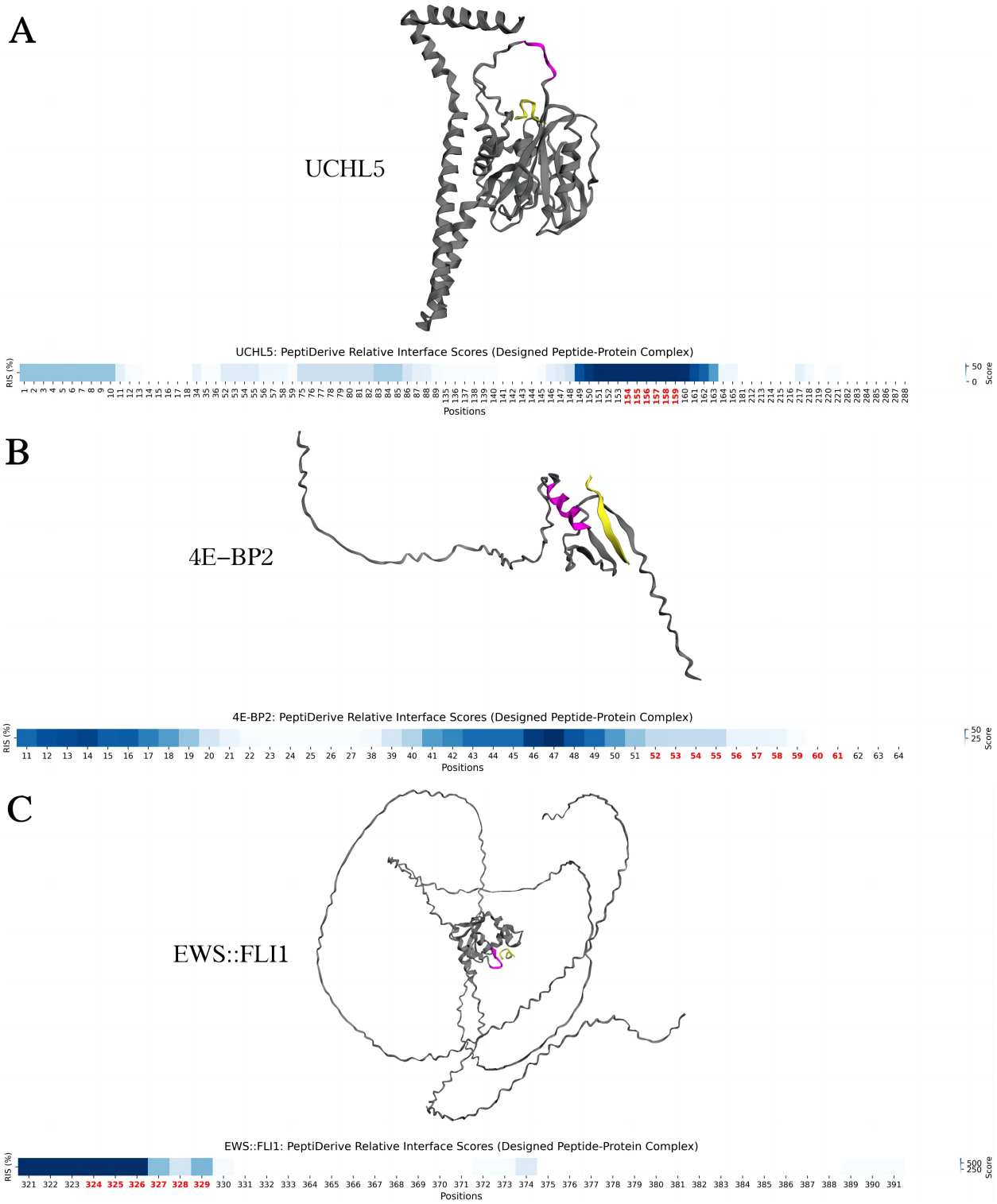


Figure 5: **Structural visualization and PeptiDerive relative interaction scores for designed peptides targeting disordered regions.** The peptide-complex structures are visualized for three proteins with disordered regions: **(A)** UCHL5, **(B)** 4E-BP2, and **(C)** EWS::FLI1. The target proteins are depicted in grey, the designed peptide binders are shown in yellow, and the binding residues specified by the moPPIt algorithm are highlighted in magenta. Below each structure, the relative interaction scores (RIS) computed by PeptiDerive are shown, with high scores indicating strong binding potential. Positions highlighted in red were input into moPPIt as the desired target amino acids.

Supplementary Information

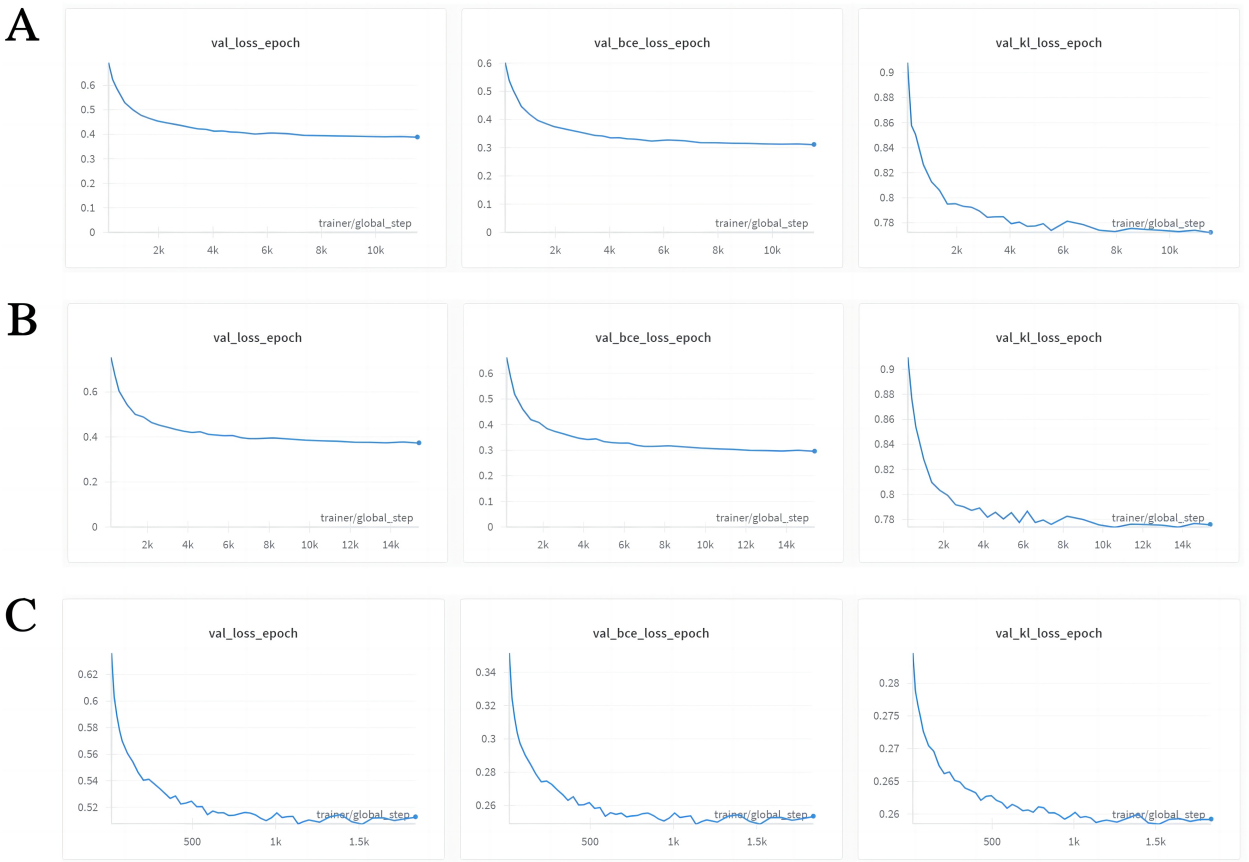


Figure S1: **Validation loss curves for BindEvaluator training and fine-tuning. (A)** Validation loss, binary cross-entropy (BCE) loss, and Kullback-Leibler (KL) divergence loss curves during training of BindEvaluator on the PPI dataset without dilated CNN modules. **(B)** Loss curves for training with dilated CNN modules, showing similar trends to (A) but with noticeable reductions in losses during the final epochs. **(C)** Loss curves during fine-tuning of BindEvaluator with dilated CNN modules on peptide-protein binding data, illustrating further decreases in loss metrics, particularly in KL divergence.

Table S1: **Comparison of ipTM and pTM scores for existing and designed peptide-protein complexes.** The ipTM and pTM scores are calculated by AlphaFold2-Multimer for peptide-protein complexes using both existing peptides and peptides designed by the moPPIt algorithm. The designed binders for each protein are presented.

PDB ID	ipTM score (existing binder)	ipTM score (designed binder)	pTM score (existing binder)	pTM score (designed binder)	Designed Binder
1A2C	0.91	0.93	0.96	0.96	GYEEIPEEYLQ
1AYC	0.69	0.8	0.9	0.88	SSQVVADLQPP
1B8Q	0.84	0.83	0.82	0.81	VVSVDSV
1LTJ	0.94	0.87	0.89	0.89	GHRG
2FAI	0.88	0.89	0.95	0.95	HHKILHRLLQDSS
3IDG	0.51	0.51	0.71	0.72	PRRRGGRR
3IDJ	0.48	0.63	0.79	0.77	LLLELDKWLLS
4IU7	0.93	0.92	0.93	0.92	KKIHHRLLQD
5E1C	0.86	0.81	0.93	0.93	HKKIHHRLLQQQSE
5EZ0	0.84	0.84	0.85	0.85	GWESLKTGKETPL
5KRI	0.86	0.84	0.93	0.92	HKKILHRLLQDSSS
5M02	0.5	0.49	0.86	0.86	KAPANFATM
6ERW	0.96	0.94	0.95	0.96	TTYADIIASGRTGRRAAI
8CN1	0.93	0.93	0.89	0.89	VVTV
8OEP	0.84	0.84	0.84	0.85	RWRDPKARPGRETPL

Table S2: **pTM and ipTM Scores for designed binders targeting proteins without known binders.** This table lists the pTM and ipTM scores for the complex structures of proteins with designed binders targeting proteins without known binders. The proteins are categorized by type, including kinases, phosphatases, and deubiquitinating enzymes (DUBs). The designed binders are provided alongside each protein.

UniProt ID	Protein Name	Type	ipTM score	pTM score	Designed Binder
P49759	CLK1	Kinases	0.76	0.72	PDGDRR
P11309	P1M1	Kinases	0.83	0.84	KKRRRHPS
P11801	PSKH1	Kinases	0.86	0.72	RRPDDIAW
P17612	PRKACA	Kinases	0.88	0.95	TRGRIHI
P53041	PPP5	Phosphatases	0.88	0.89	EDLPA
Q15257	PTPA	Phosphatases	0.88	0.84	PDLFDLFL
P67775	PPP2CA	Phosphatases	0.8	0.92	SELGDRFP
P62136	PPP1CA	Phosphatases	0.82	0.89	PLVVTE
P63279	UBC9	DUBs	0.84	0.92	AQVVPE
Q8N5J2	MINDY1	DUBs	0.87	0.58	SRLSSGK

Table S3: **pTM and ipTM scores for designed binders targeting disordered regions in selected proteins.** The table includes the UniProt ID, protein name, binding sites, disordered regions, pTM and ipTM scores, and the designed binders for each protein. High pTM and ipTM scores indicate the reliability and stability of the predicted complex structures.

UniProt ID	Protein Name	Binding Sites	Disordered Regions	ipTM score	pTM score	Binder
Q9Y5K5	UCHL5	153-158	1-4/148-159/243-253/327-329	0.68	0.81	AQRGRGR
Q13542	4E-BP2	52-67	1-119	0.71	0.36	STTAQAFVQE
B1PRL2	EWS:FLI	324-331	1-260	0.82	0.3	GPSSWYS

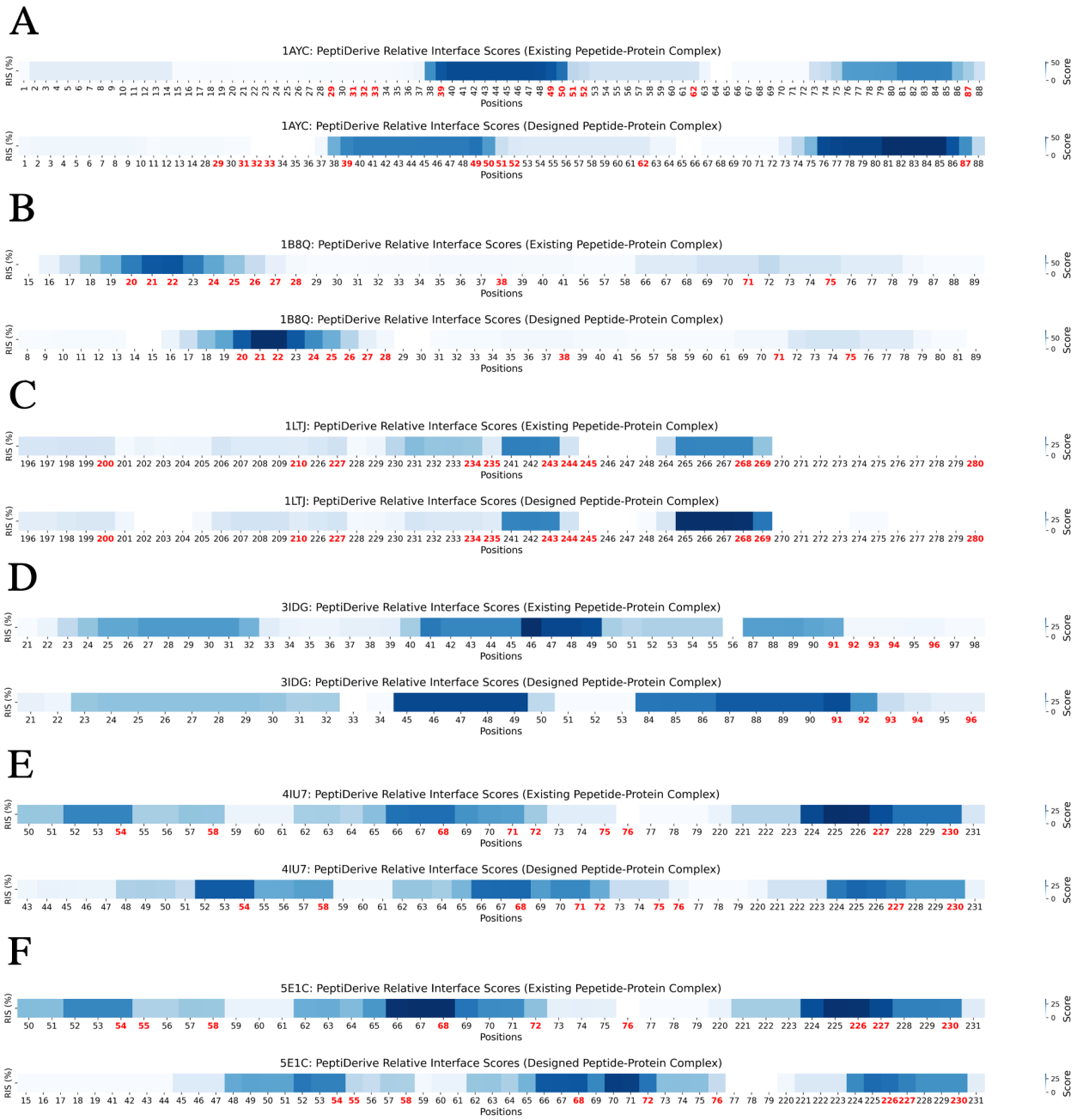


Figure S2: **PeptiDerive relative interface scores for existing and designed peptide-protein complexes.** Heatmaps of PeptiDerive relative interface scores (RIS) are shown for 6 peptide-protein complexes among 15 structured complexes with known binders that were tested: **(A)** 1AYC, **(B)** 1B8Q, **(C)** 1LTJ, **(D)** 3IDG, **(E)** 4IU7, **(F)** 5E1C. The first heatmap for each protein shows the RIS of the existing peptide-protein complex, while the second heatmap shows the scores for the designed peptide-protein complex. For each heatmap, the x-axis indicates the residue positions, with highlighted positions in red representing the target binding amino acid positions that were input into moPPIt. High RIS at these positions indicate strong binding potential.

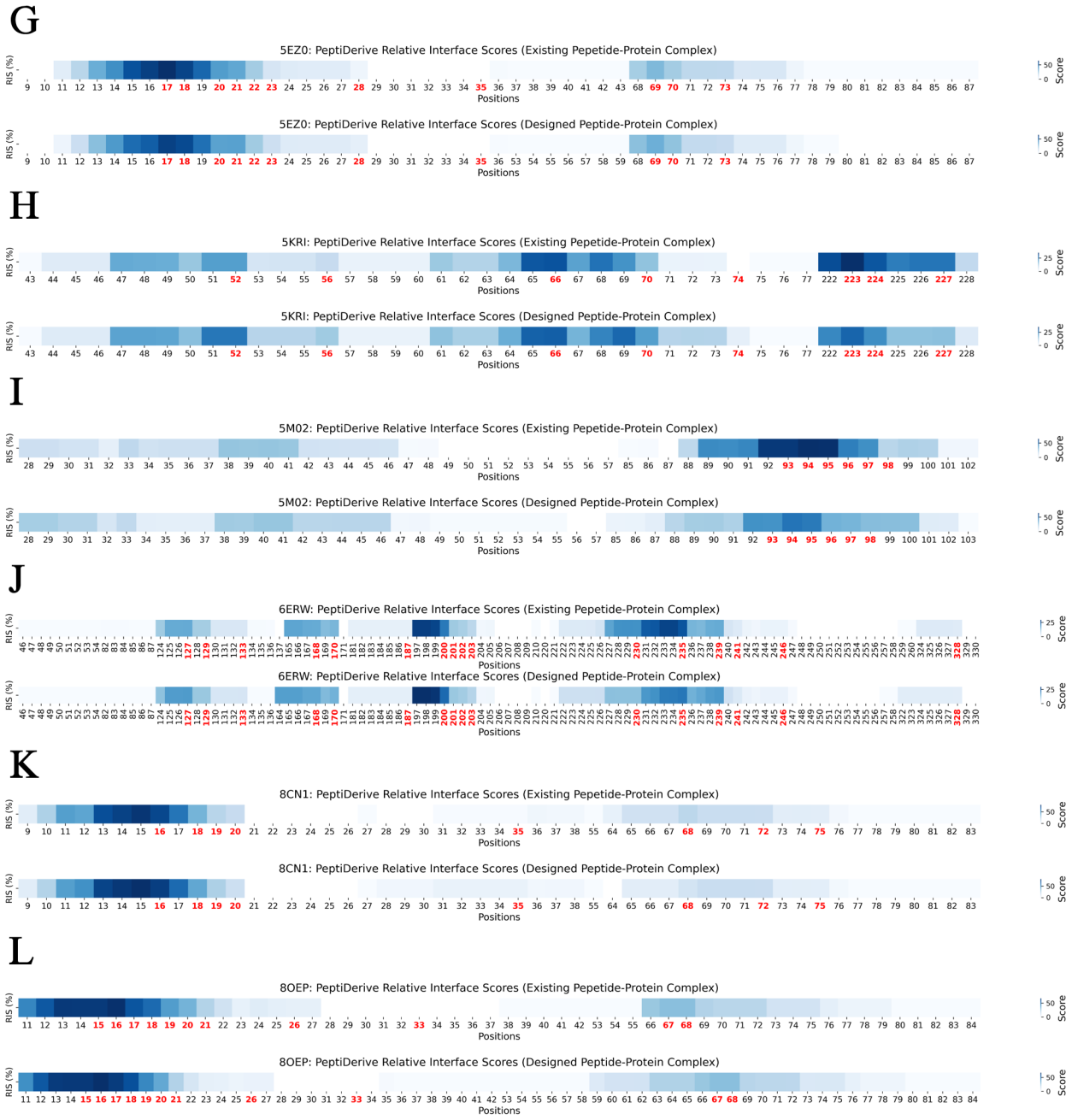


Figure S3: **PeptiDerive relative interface scores for existing and designed peptide-protein complexes.** Heatmaps of PeptiDerive relative interface scores (RIS) are shown for 6 peptide-protein complexes among 15 structured complexes with known binders that were tested: **(G)** 5EZ0, **(H)** 5KRI, **(I)** 5M02, **(J)** 6ERW, **(K)** 8CN1, **(L)** 8OEP. The first heatmap for each protein shows the RIS of the existing peptide-protein complex, while the second heatmap shows the scores for the designed peptide-protein complex. For each heatmap, the x-axis indicates the residue positions, with highlighted positions in red representing the target binding amino acid positions that were input into moPPIt. High RIS at these positions indicate strong binding potential.

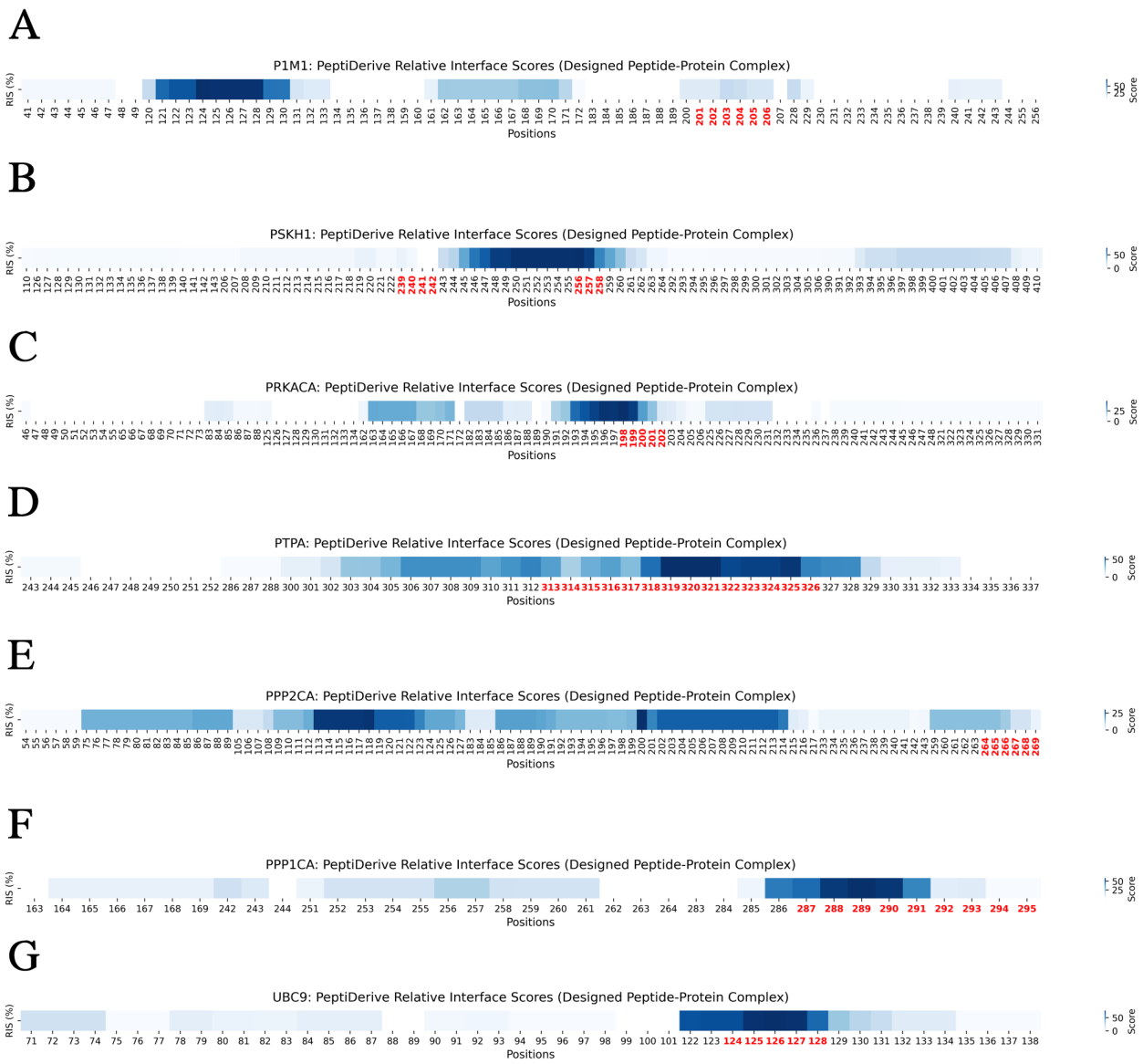


Figure S4: **PeptiDerive relative interface scores for complexes with peptides designed to novel structured targets.** Heatmaps of PeptiDerive relative interface scores (RIS) are shown for 7 peptide-protein complexes from 10 proteins without known binders that were tested: **(A)** P1M1, **(B)** PSKH1, **(C)** PRKACA, **(D)** PTPA, **(E)** PPP2CA, **(F)** PPP1CA, **(G)** UBC9. For each heatmap, the x-axis indicates the residue positions, with highlighted positions in red representing the target binding amino acids positions that were input into moPPIt. High RIS at these positions indicate strong binding potential.