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

Tong Chen
Department of Biomedical Engineering
Duke University
Durham, NC 27701
tong.chen2@duke.edu

Yinuo Zhang
Department of Biomedical Engineering
Duke University
Durham, NC 27701
yzhang@u.duke.nus.edu

Zachary Quinn
Department of Biomedical Engineering
Duke University
Durham, NC 27701
zachary.quinn@duke.edu

Pranam Chatterjee
Department of Biomedical Engineering
Duke University
Durham, NC 27701
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.

1 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 [1]. 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

Submitted to the NeurIPS 2024 Machine Learning in Structural Biology Workshop.

approach in cancers where p53 is inactivated by mutations [2]. 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 [3]. 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 [4]. Allosteric domains present another important target, exemplified by the potential to modulate G protein-coupled receptor (GPCR) function by binding to their allosteric sites [5]. For intrinsically disordered proteins, targeting specific regions of the tau protein involved in pathological aggregation could provide new avenues for treating tauopathies [6]. 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 [7, 8].

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 [9]. Advances including AlphaFold and RFDiffusion, have shown promise in various protein design tasks, including motif-specific binder design [10–13]. 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 [14], 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 [15–18]. These pLMs have demonstrated utility in various protein design tasks, including our recent work on designing target-specific peptide binders from target sequence alone [19–21]. 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 [15] 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 [20], 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 [22] and PeptiDerive, a Rosetta-based algorithm for identifying key binding residues [23], 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.

2 Methods

BindEvaluator Model Architecture To enable motif-specific peptide binder generation, we first developed a BindEvaluator model to predict peptide-protein interaction binding sites (Figure 1A). BindEvaluator takes a binder sequence and a target sequence as inputs to predict the binding residues on the target protein. Both binder and target sequences are first passed into a pre-trained ESM-2-650M model to obtain their embeddings [15]. For the target sequence embedding, a dilated convolutional neural network (CNN) module captures the local features of adjacent residues. The processed embeddings are then passed through multi-head attention modules to capture global dependencies for

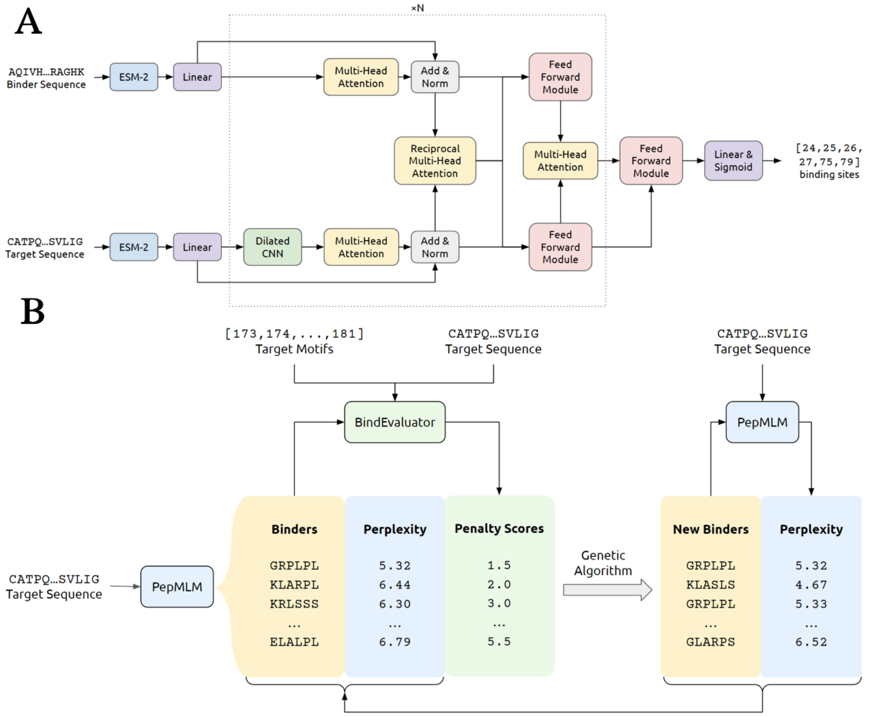

Figure 1: **(A)** Overview of the architecture of BindEvaluator. **(B)** Schematic of moPPIt.

each residue. In the reciprocal attention modules, the target and binder sequence representations are integrated to capture binder-target interaction information. 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.

Motif-Specific Binder Design Algorithm Our **mo**tif-specific **PPI t**argeting algorithm (**moPPIt**) aims to generate motif-specific peptide binders based on the target protein sequence (Figure 1B). The algorithm begins with PepMLM generating a pool of candidate peptide binders of a defined length for a given target protein sequence. PepMLM is a state-of-the-art ESM-2-based model that generates peptide binders conditioned on a target protein sequence alone [20]. 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. Additionally, PepMLM computes the perplexity (PPL) 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. The 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.

3 Results

We initially trained BindEvaluator without dilated CNN modules on a large protein-protein interaction (PPI) dataset containing over 500,000 entries with annotated interface residues [24] to provide foundational knowledge of protein interaction information. The model's performance on the test data confirmed its efficacy in distinguishing between binding and non-binding residues (Table 1). We hypothesized that incorporating dilated CNN modules into BindEvaluator would enhance its performance by effectively extracting local features relevant to binding site information. To test this hypothesis, we trained a version of BindEvaluator with dilated CNN modules on the same PPI

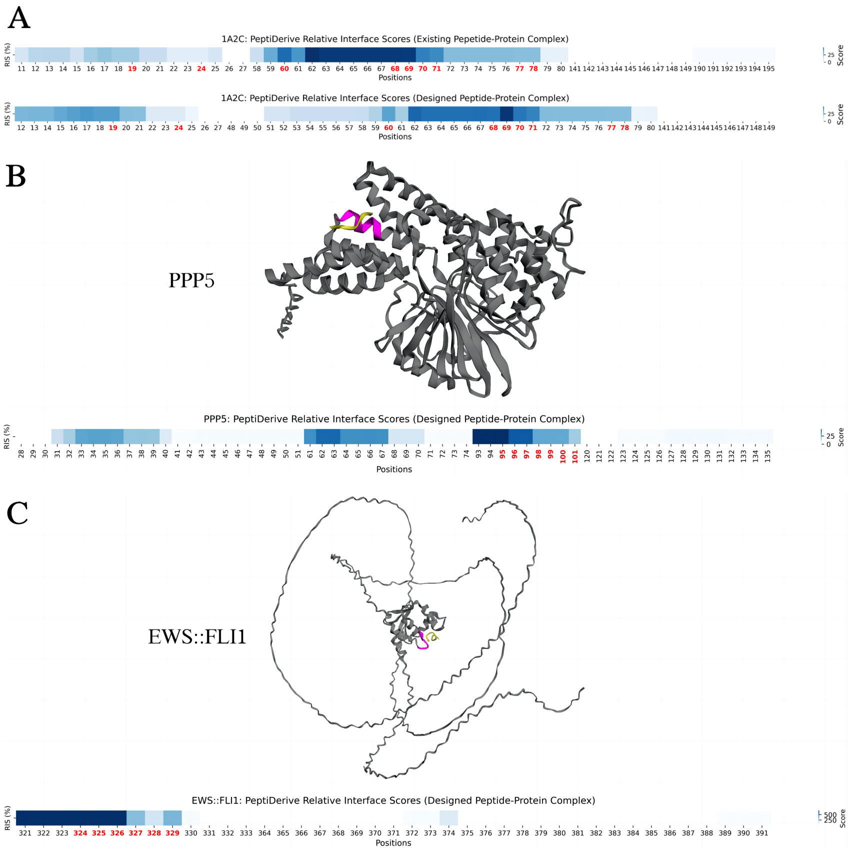

Figure 2: **(A)** PeptiDerive relative interface scores (RIS) for existing and designed peptide-protein complexes were computed and visualized for an example protein with PDB ID: 1A2C. The peptide-complex structure is visualized for **(B)** a protein without known binders (PPP5) and **(C)** a protein with disordered regions (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.

dataset with almost identical training settings except for slightly different gradient accumulation schedules. The inclusion of these CNN modules led to observable improvements across several metrics (Table 1). To adapt our model for peptide-protein binding site prediction, the pre-trained BindEvaluator model with dilated CNN modules was further fine-tuned on over 12,000 structurally validated, non-redundant peptide-protein sequence pairs, which also achieved strong test metrics, indicating high precision in peptide-protein binding site prediction (Table 1).

Table 1: Test Performance metrics of BindEvaluator Across Different Training Configurations

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

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 analyzed the relative interface scores (RIS) of both existing and designed peptide-protein complexes using PeptiDerive [23], which evaluates the energy contribution of specific residues to the overall free energy of the binder-target complex structure (Figure 2A, 7, 8). 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.

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. The potential binding sites are identified by PepMLM and BindEvaluator. We evaluated the epitope specificity of designed binders to corresponding targets (Figure 2B, 5, 9). 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.

To demonstrate moPPIt's ability to design binders targeting intrinsically disordered proteins, we selected three proteins with structurally disordered domains (UCHL5, 4E-BP2, and EWS::FLI1) and designed binders for them using moPPIt. The PeptiDerive scores align with the specified binding motifs, showing high predicted RIS (Figure 2C, 6). 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.

4 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 [12, 14], 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 [19–21]. 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 [25]. 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 [26]. 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.

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.

References

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

[2] K. D. Sullivan, M. D. Galbraith, Z. Andrysik, and J. M. Espinosa, "Mechanisms of transcriptional regulation by p53," *Cell Death amp; Differentiation*, vol. 25, p. 133–143, Nov. 2017.

[3] S. Kitazume, Y. Tachida, R. Oka, K. Shirotani, T. C. Saido, and Y. Hashimoto, "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*, vol. 98, p. 13554–13559, Nov. 2001.

[4] G. Castellani, M. Buccarelli, M. B. Arasi, S. Rossi, M. E. Pisanu, M. Bellenghi, C. Lintas, and C. Tabolacci, "Braf mutations in melanoma: Biological aspects, therapeutic implications, and circulating biomarkers," *Cancers*, vol. 15, p. 4026, Aug. 2023.

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

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

[7] C. M. Linardic, "Pax3–foxo1 fusion gene in rhabdomyosarcoma," *Cancer Letters*, vol. 270, p. 10–18, Oct. 2008.

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

[9] T. Chen, L. Hong, V. Yudistyra, S. Vincoff, and P. Chatterjee, "Generative design of therapeutics that bind and modulate protein states," *Current Opinion in Biomedical Engineering*, vol. 28, p. 100496, Dec. 2023.

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

[11] J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, S. W. Bodenstein, D. A. Evans, C.-C. Hung, M. O'Neill, D. Reiman, K. Tunyasuvunakool, Z. Wu, A. Žemgulytė, E. Arvaniti, C. Beattie, O. Bertolli, A. Bridgland, A. Cherepanov, M. Congreve, A. I. Cowen-Rivers, A. Cowie, M. Figurnov, F. B. Fuchs, H. Gladman, R. Jain, Y. A. Khan, C. M. R. Low, K. Perlin, A. Potapenko, P. Savy, S. Singh, A. Stecula, A. Thillaisundaram, C. Tong, S. Yakneen, E. D. Zhong, M. Zielinski, A. Žídek, V. Bapst, P. Kohli, M. Jaderberg, D. Hassabis, and J. M. Jumper, "Accurate structure prediction of biomolecular interactions with alphafold3," *Nature*, May 2024.

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

[13] P. Bryant and A. Elofsson, “Peptide binder design with inverse folding and protein structure prediction,” *Communications Chemistry*, vol. 6, Oct. 2023.

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

[15] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, *et al.*, “Evolutionary-scale prediction of atomic-level protein structure with a language model,” *Science*, vol. 379, no. 6637, pp. 1123–1130, 2023.

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

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

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

[19] G. Brixi, T. Ye, L. Hong, T. Wang, C. Monticello, N. Lopez-Barbosa, S. Vincoff, V. Yudistyra, L. Zhao, E. Haarer, *et al.*, “Salt&peppr is an interface-predicting language model for designing peptide-guided protein degraders,” *Communications Biology*, vol. 6, no. 1, p. 1081, 2023.

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

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

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

[23] Y. Sedan, O. Marcu, S. Lyskov, and O. Schueler-Furman, “Peptiderive server: derive peptide inhibitors from protein–protein interactions,” *Nucleic Acids Research*, vol. 44, p. W536–W541, May 2016.

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

[25] Z. Peng, B. Schussheim, and P. Chatterjee, “Ptm-mamba: A ptm-aware protein language model with bidirectional gated mamba blocks,” *bioRxiv*, Feb. 2024.

[26] M. H. Abbasian, M. Mahmanzar, K. Rahimian, B. Mahdavi, S. Tokhanbigli, B. Moradi, M. M. Sisakht, and Y. Deng, “Global landscape of sars-cov-2 mutations and conserved regions,” *Journal of Translational Medicine*, vol. 21, Feb. 2023.

[27] O. Abdin, S. Nim, H. Wen, and P. M. Kim, “Pepnn: a deep attention model for the identification of peptide binding sites,” *Communications biology*, vol. 5, no. 1, p. 503, 2022.

[28] C. Zhang, X. Zhang, P. L. Freddolino, and Y. Zhang, “Biolip2: an updated structure database for biologically relevant ligand–protein interactions,” *Nucleic Acids Research*, vol. 52, no. D1, pp. D404–D412, 2024.

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

5 Supplementary Material

5.1 Dataset Curation

The training data for BindEvaluator was curated from the PPIRef dataset, a large and non-redundant databank of PPIs [24]. 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 [27, 28]. 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.

5.2 BindEvaluator Architecture Details

Dilated CNN modules BindEvalutor takes a target sequence and a binder sequence as inputs. Both sequences will first be processed by a pre-trained ESM-2-650 model to generate embeddings [15]. The target sequence embedding will be further processed by a dilated convolutional neural network (CNN) module to capture 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.

Reciprocal Multi-head Attention Both binder embedding and target embedding will be further processed by multiple multi-head self-attention modules and reciprocal multi-head attention modules. In the reciprocal attention 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. In this way, both resulting target embedding and binder embedding will contain binder-target binding information.

5.3 BindEvaluator 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:

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

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.

5.4 moPPIt Details

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 1B.

Penalty scores and Perplexity Given a target protein sequence, PepMLM generates a pool of candidate peptide binders of a specified length. The fine-tuned BindEvaluator then predicts the binding sites on the target protein given each candidate binder. 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.

Genetic Algorithm 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.

5.5 Validation Loss curves for BindEvaluator Training and Fine-tuning

We first trained BindEvaluator without dilated CNN modules on a large protein-protein interaction dataset. During training, we observed a consistent decline in the validation loss, which indicates stable and effective learning (Figure 3A). 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. We then trained BindEvaluator with dilated CNN modules on the same dataset. Both models, with and without dilated CNN modules demonstrated similar declining trends in their loss curves, indicating effective learning (Figure 3B). 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. During fine-tuning on the peptide-protein interaction dataset, we observed validation loss decreasing steadily (Figure 3C), indicating a steady improvement in binding site prediction abilities.

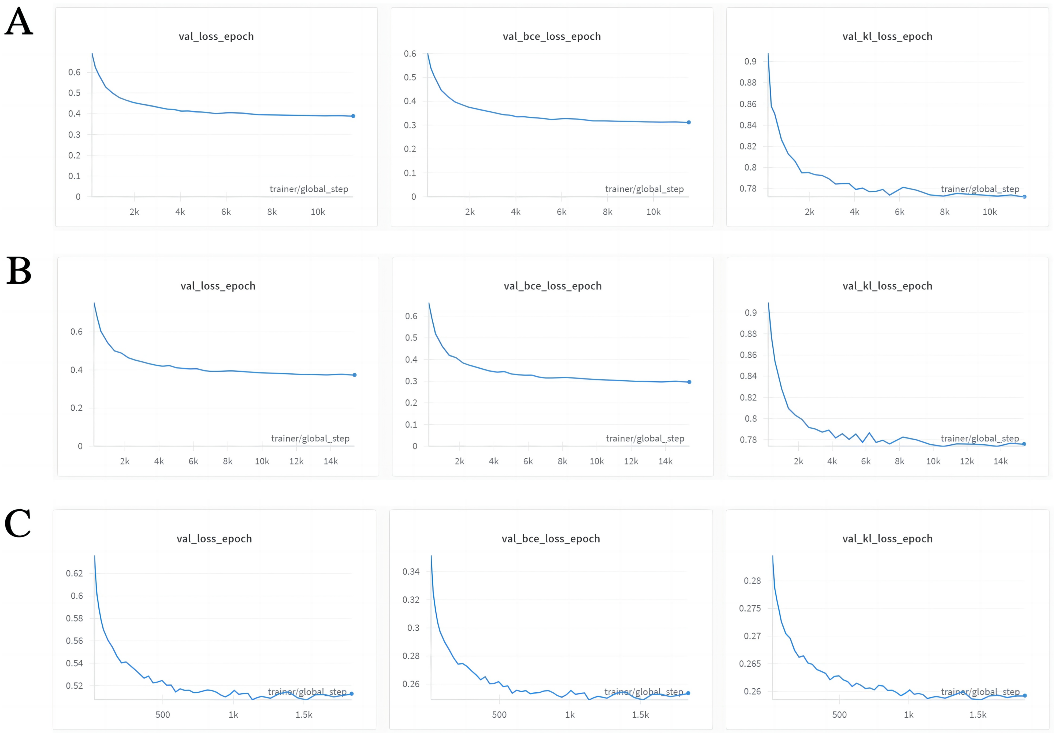

Figure 3: **Validation loss curves for BindEvaluator training and fine-tuning. (A)** Validation loss, binary cross-entropy (BCE) loss, and Kullback-Leibler (KL) divergence loss curves during training of BindEvaluator on the PPI dataset without dilated CNN modules. **(B)** Loss curves for training with 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.

5.6 moPPIt-designed Binders Demonstrate High Binding Capacity

We evaluated the binding potential between moPPIt-designed binders and the target proteins. For the binders designed for 15 structured, unseen proteins with known, pre-existing binders, 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 [22]. These scores were used to compare the performance of the pre-existing binders with those designed by moPPIt. 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 2). 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 4). 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 2). Similar pTM and ipTM results were observed for structured proteins without pre-existing binders (Table 3).

We also assessed the binding capacity of moPPIt-designed binders targeting three intrinsically disordered proteins (UCHL5, 4E-BP2, and EWS::FLI1). Table 4 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 [29]. 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 4).

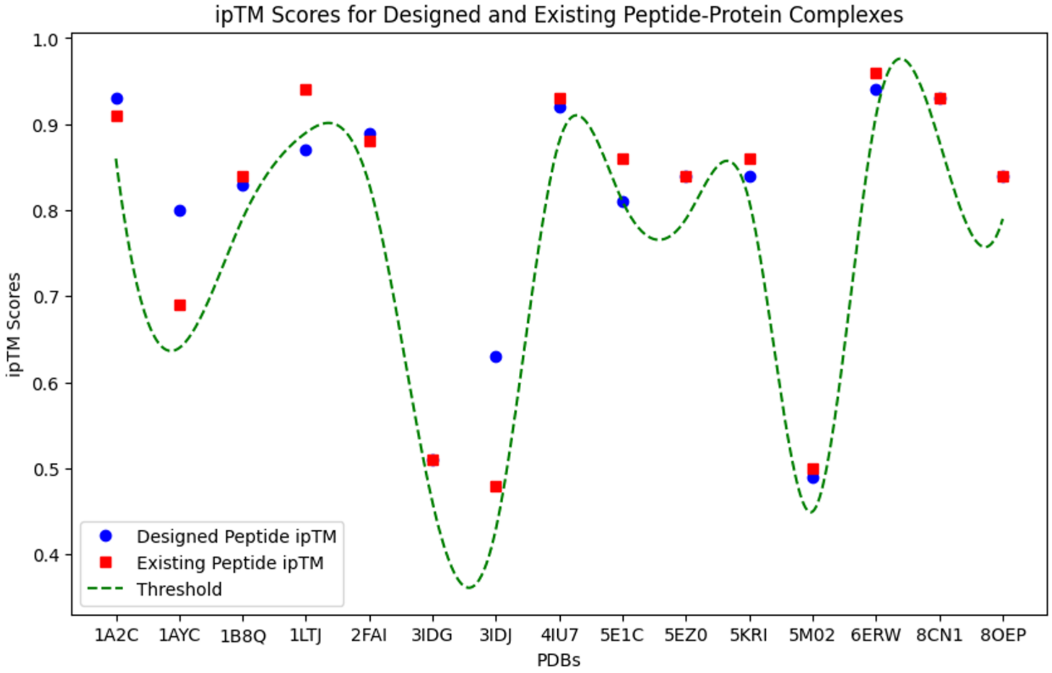

Figure 4: **Hit rate of moPPIt on structured targets with known binders.** The ipTM scores of input peptides, in complex with their target protein, were calculated via AlphaFold-Multimer. The ipTM scores for known peptides (red) from PDB structures were compared to moPPIt-designed peptides (blue) for 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.

5.7 Identify Potential Binding Sites Using PepMLM and BindEvaluator

We leveraged pre-trained PepMLM and fine-tuned BindEvaluator model to identify potential binding sites for structured proteins without pre-existing binders and intrinsically disordered proteins [20]. Specifically, 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.

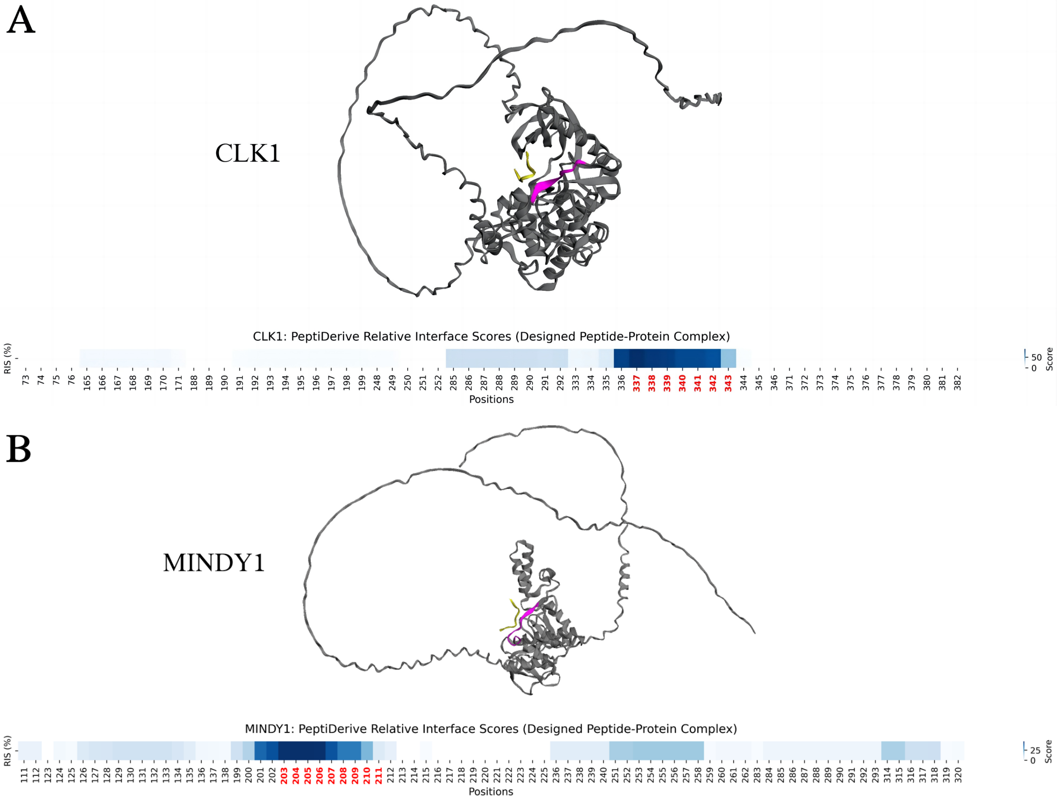

Figure 5: **Structural visualization and PeptiDerive relative interaction scores for designed peptides targeting structured motifs.** The peptide-complex structures are visualized for two proteins without known binders: **(A)** CLK1 and **(B)** MINDY1. The target proteins are depicted in grey, the designed 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.

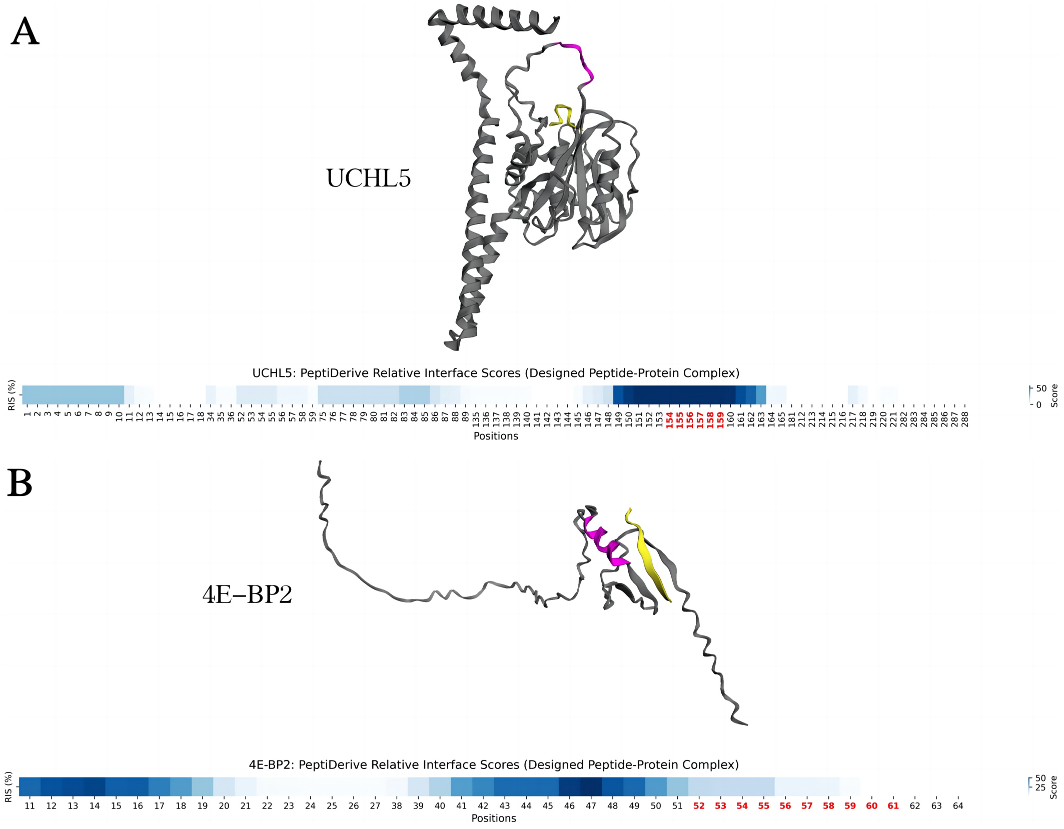

Figure 6: **Structural visualization and PeptiDerive relative interaction scores for designed peptides targeting disordered regions.** The peptide-complex structures are visualized for two proteins with disordered regions: **(A)** UCHL5 and **(B)** 4E-BP2. The target proteins are depicted in grey, the 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.

Table 2: **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 3: **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 4: **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

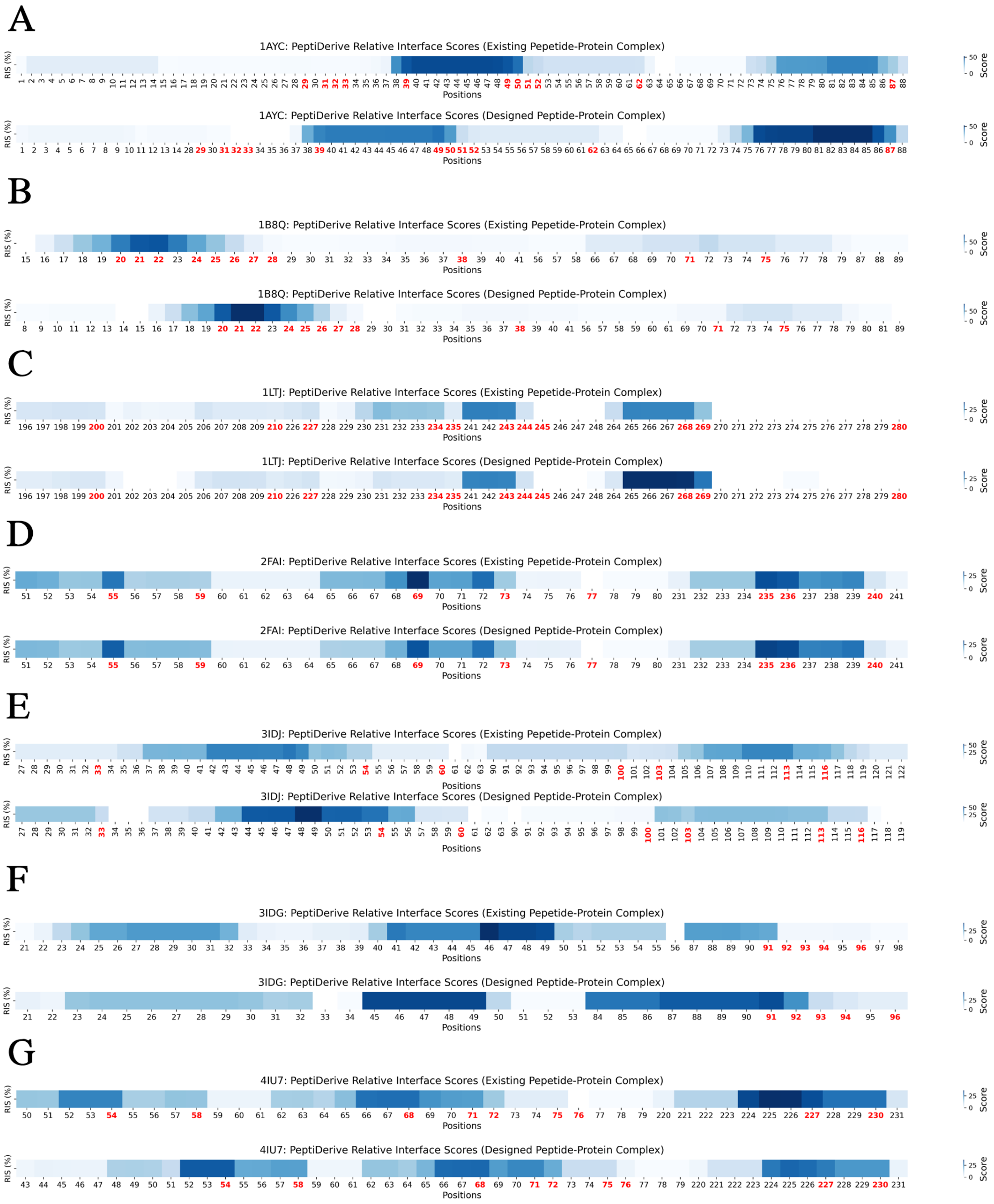

Figure 7: **PeptiDerive relative interface scores for existing and designed peptide-protein complexes.** Heatmaps of PeptiDerive relative interface scores (RIS) are shown for 7 peptide-protein complexes among 15 structured complexes with known binders that were tested: **(A)** 1AYC, **(B)** 1B8Q, **(C)** 1LTJ, **(D)** 2FAI, **(E)** 3IDJ, **(F)** 3IDG, **(G)** 4IU7. 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.

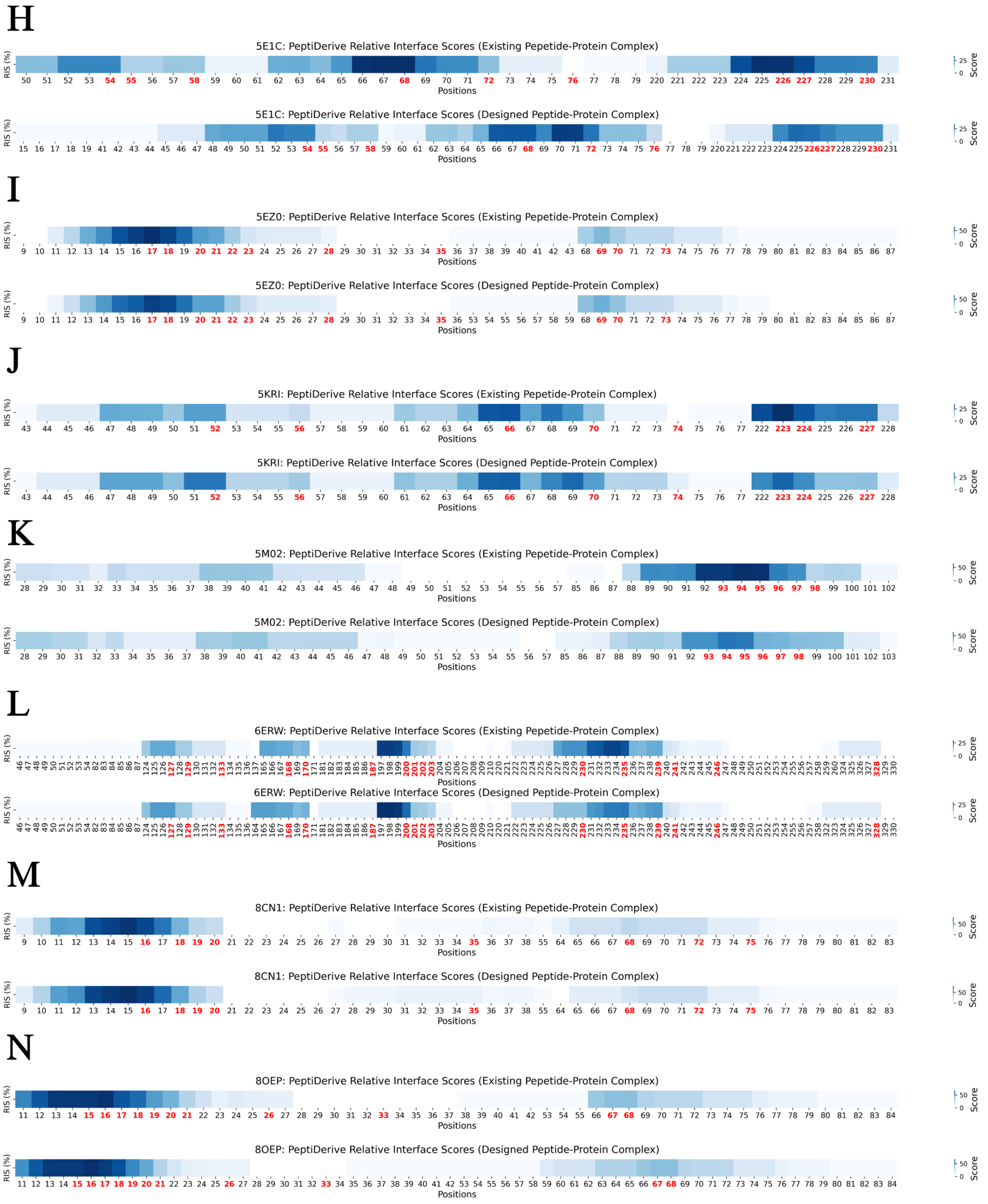

Figure 8: **PeptiDerive relative interface scores for existing and designed peptide-protein complexes.** Heatmaps of PeptiDerive relative interface scores (RIS) are shown for 7 peptide-protein complexes among 15 structured complexes with known binders that were tested:**(H)** 5E1C, **(I)** 5EZ0, **(J)** 5KRI, **(K)** 5M02, **(L)** 6ERW, **(M)** 8CN1, **(N)** 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.

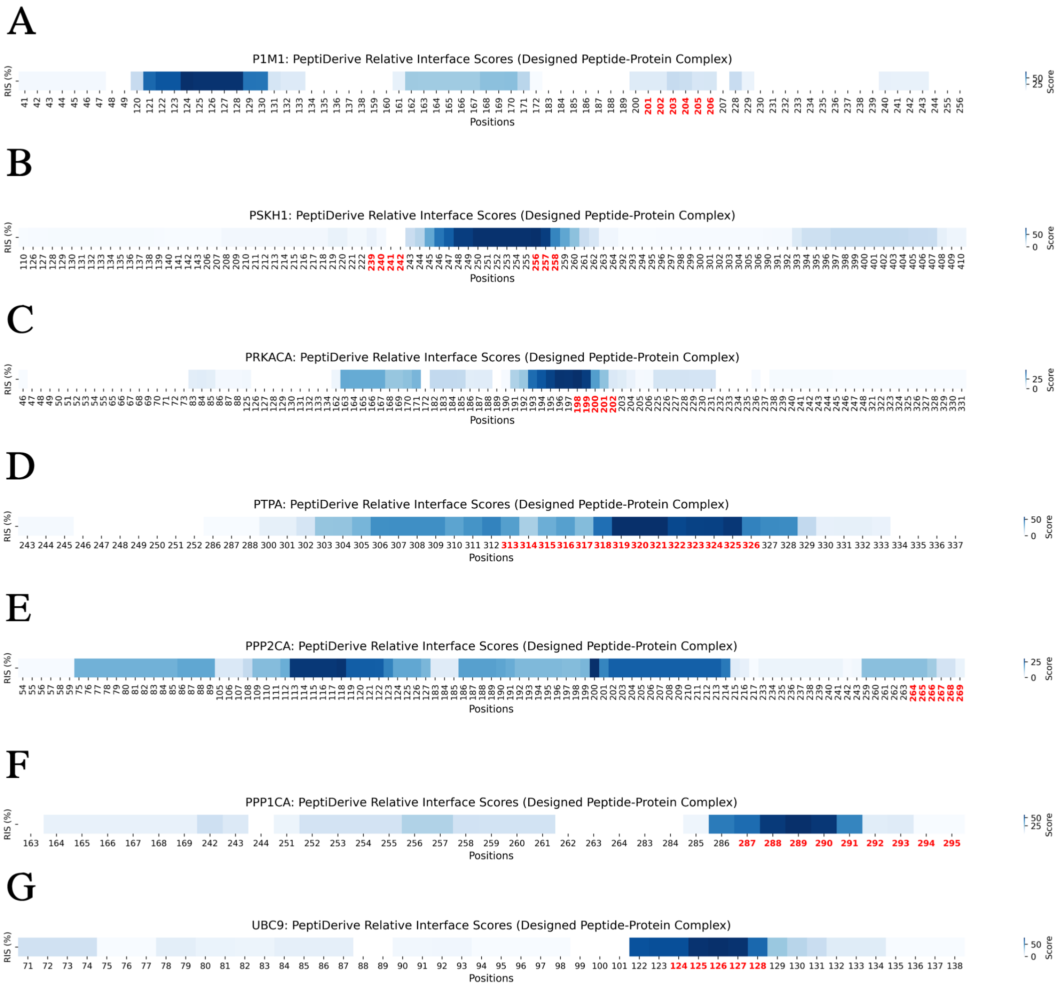

Figure 9: **PeptiDerive relative interface scores for complexes with peptides designed to novel structured targets.** Heatmaps of PeptiDerive relative interface scores (RIS) are shown for 7 peptide-protein complexes from 10 proteins without known binders that were tested: **(A)** P1M1, **(B)** PSKH1, **(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.