moPPIt: *De Novo* Generation of Motif-Specific Peptide Binders via Conditional Uniform Discrete Diffusion

Tong Chen,[1] **Yinuo Zhang,**[1] **Zachary Quinn,**[1] **Pranam Chatterjee**[1,†]

[1]Duke University, Durham, NC

[†]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 often fall short in achieving this specificity due to a lack of reliable structural information. In this work, we introduce **moPPIt**, a **mo**tif-specific **PPI t**argeting algorithm, 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 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. Additionally, we present **PepUDLM**, a uniform diffusion language model that generates diverse and biologically plausible peptides. By integrating BindEvaluator into PepUDLM's sampling process, 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 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. (2023b). 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. (2024a); Watson et al. (2023); Bryant & 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 diffusion models like masked diffusion language model and uniform diffusion language model, which have been trained on vast, diverse protein sequence datasets to capture underlying physicochemical and functional properties of protein sequences and to support guided generation of novel proteins with specific properties Sahoo et al. (2024); Schiff et al. (2024). However, existing diffusion-based methods have not yet been focused on generating binders with a specific motif-targeting property, leaving a significant gap in our ability to design motif-specific therapeutics.

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 protein language model (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 peptide-protein pairs. We further trained PepUDLM that generates diverse and biologically plausible peptides, a uniform diffusion language model trained on a custom dataset, comprising peptides from the PepNN, BioLip2, and PPIRef dataset Abdin et al. (2022); Zhang et al. (2024); Bushuiev et al. (2023). moPPIt integrates BindEvaluator into PepUDLM's sampling process, where BindEvaluator's predictions guide PepUDLM to generate binders specifically targeting user-defined motifs. We demonstrate moPPIt's efficacy in designing binders to specific epitopes on a diverse set of targets, including kinases, transcription factors, GPCRs, and even intrinsically disordered regions (IDRs). Using a combination of AlphaFold3, AutoDock VINA, and PeptiDerive, a Rosetta-based algorithm for identifying key binding residues Abramson et al. (2024b); Sedan et al. (2016); Eberhardt et al. (2021), 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 and Results

2.1 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 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 Lin et al. (2023). 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 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

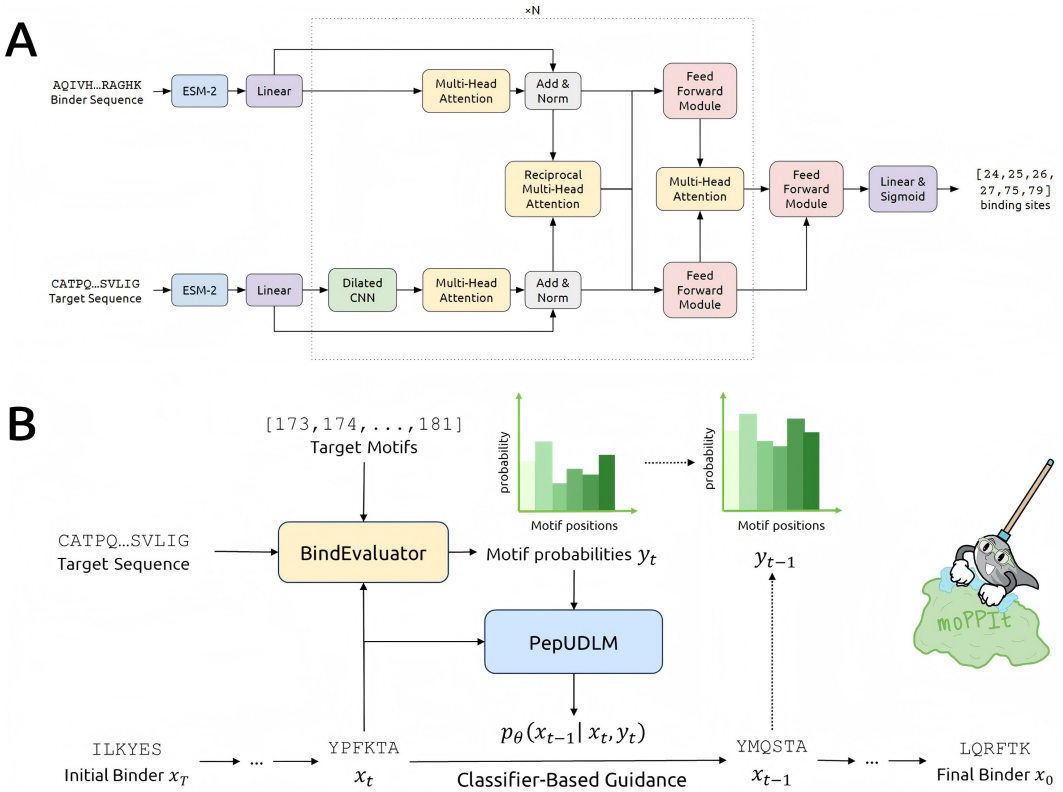

Figure 1: **(A)** Overview of the architecture of BindEvaluator. **(B)** Schematic of moPPIt.

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.

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 Bushuiev et al. (2023) 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 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).

2.2 PepUDLM Generates Diverse and Biologically Plausible Peptides

To enable the efficient generation of peptide binders, we developed an unconditional peptide generator, **PepUDLM**, based on the Uniform Diffusion Language Model (UDLM) Schiff et al. (2024). UDLMs can reverse random token perturbations and continuously edit discrete data, making them highly suitable for guided generation. We trained PepUDLM on a custom dataset that includes all peptides from the PepNN and BioLip2 datasets, as well as sequences from the PPIRef dataset with lengths ranging from 6 to 49 amino acids Abdin et al. (2022); Zhang et al. (2024); Bushuiev et al. (2023). PepUDLM demonstrates superior performance compared to autoregressive generators across multiple evaluation metrics, including lower Bits Per Dimension (BPD), reduced Negative Log-Likelihood (NLL), and significantly improved perplexity (PPL) (Table 2). Furthermore, PepUDLM generates

peptides with substantially high Hamming distances from the test set, indicating a great degree of diversity and novelty in the generated sequences (Figure 4). Additionally, the Shannon entropy of the generated peptides closely matches that of the test set, highlighting the model's capability to produce biologically plausible peptides with diverse sequence lengths (Figure 4).

2.3 moPPIt GENERATES EPITOPE-SPECIFIC BINDERS TO TARGET PROTEINS

With BindEvaluator for peptide-protein binding site prediction and PepUDLM for peptide generation, we developed the **mo**tif-specific **PPI t**argeting algorithm (**moPPIt**) to generate motif-specific peptide binders based solely on target protein sequences. Instead of filtering random sequences through PepUDLM, we adopted a classifier-guided diffusion approach, where binding motifs predicted by BindEvaluator guide PepUDLM to generate binders specific to the target motifs (Figure 1B).

moPPIt begins with a randomly initialized peptide sequence of a defined length. Applying a classifier-guided diffusion process, it iteratively refines the sequence by sampling from a tempered distribution:

$$p^{\gamma}(z_s|z_t, y) \propto p_{\phi}(y|z_s)^{\gamma} p_{\theta}(z_s|z_t), \tag{1}$$

where $p_{\theta}(z_s|z_t)$ represents the pre-trained PepUDLM diffusion prior, and $p_{\phi}(y|z_s)$ is the fine-tuned BindEvaluator providing motif-specific guidance. The parameter γ controls the strength of classifier guidance. From Eq1, the guidance is derived as:

$$\nabla_{\mathbf{z}_s} \log p^{\gamma}(\mathbf{z}_s \mid y, \mathbf{z}_t) = \gamma \nabla_{\mathbf{z}_s} \log p_{\phi}(y \mid \mathbf{z}_s) + \nabla_{\mathbf{z}_s} \log p_{\theta}(\mathbf{z}_s \mid \mathbf{z}_t). \tag{2}$$

BindEvaluator predicts logits for each amino acid in the target sequence, indicating their likelihood of belonging to the binding motifs, but these logits cannot be directly used as guidance for PepUDLM. Instead, we compute the average log probability of amino acids at the desired target motif positions and use it as the classifier guidance term in the first term on the right-hand side of Eq2. Specifically:

$$\log p_{\phi}(y \mid \mathbf{z}_s) = \frac{1}{n} \sum_{m_i \in M} log(softmax(logits))[m_i], \tag{3}$$

where M represents the desired motifs. This ensures that motif specificity is reinforced throughout the diffusion process. By iteratively refining sequences under this framework, PepUDLM generates peptide binders that are highly likely to interact with the specified binding motifs on the target protein.

To evaluate moPPIt in a well-controlled setting, we designed binders for 15 structured proteins from the PDB with pre-existing peptide binders. We calculated the ipTM scores, which represent confidence in interface formation, for the AlphaFold3 peptide-protein complex structures, comparing the performance of the known peptides to those designed by moPPIt Abramson et al. (2024b). We observed that moPPIt-designed binders form peptide-protein complexes with ipTM scores similar to or higher than those of the pre-existing binders (Figure 5, Table 4). The superior ipTM scores highlight moPPIt's ability to generate peptides with strong binding affinity to target proteins. 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 (Figure 2D, 10, 11). The designed complexes exhibited similar or higher RIS at the binding sites compared to existing complexes, indicating comparable or enhanced binding potential. Additionally, residues with high RIS were predominantly localized near the binding motifs, demonstrating the high specificity of moPPIt-designed binders. The AutoDock VINA structure visualizations further demonstrate high motif-specificity of moPPIt-designed binders (Figure 2A). The designed binder also exhibited a distinct structural conformation compared to the existing binder, underscoring moPPIt's ability to generate novel binders that diverge from naturally occurring sequences.

To further assess moPPIt's performance, we designed peptide binders for structured proteins without pre-existing binders. We selected proteins from three enzyme classes (kinases, phosphatases, and deubiquitinases), as well as GPCRs, to evaluate moPPIt's versatility in designing binders for diverse structured proteins without pre-identified binding sites. Potential binding sites are identified using APBS electrostatic analysis. We evaluated the epitope specificity of the designed binders to their respective targets (Figure 2B, 2D, 6, 7, 8, Table 3). Notably, residues at the specified binding motifs exhibited high relative interface scores (RIS) predicted by PeptiDerive, demonstrating moPPIt's ability to generate highly specific binders. Additionally, the high AutoDock VINA docking scores

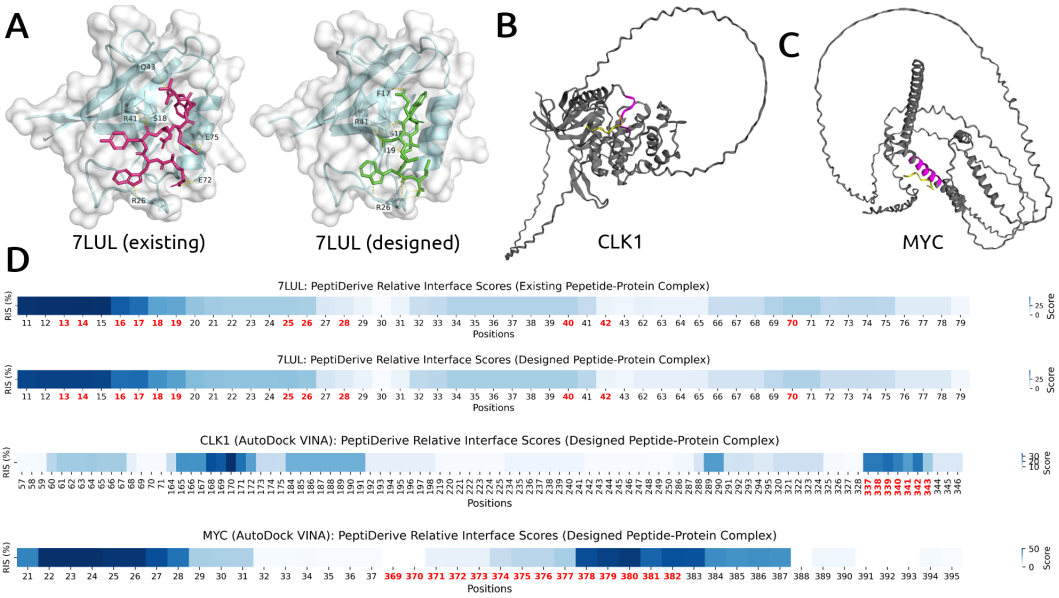

Figure 2: **(A)** AutoDock VINA docking visualization of protein (PDB ID: 7LUL) with existing and designed peptide binders, highlighting interacting residues. **(B), (C)** AutoDock VINA visualizations of a structured and disordered protein without pre-existing binders, with target proteins in grey, designed peptides in yellow, and binding residues from moPPIt in magenta. **(D)** PeptiDerive relative interaction scores (RIS), with high scores indicating strong binding potential. Positions highlighted in red were input into moPPIt as target motifs.The first two heatmaps show RIS for 7LUL with existing and designed binders, while the third and fourth heatmaps show RIS for CLK1 and MYC in (B) and (C), respectively.

and 3D visualizations, which show the designed peptides adjacent to the target binding sites, further validate moPPIt's capacity to produce binders with strong affinity for the target motifs.

To demonstrate moPPIt's capability in designing binders for intrinsically disordered proteins, we selected two proteins with disordered domains (MYC and EWS::FLI1) and designed binders using moPPIt. The PeptiDerive scores align with the specified binding motifs, showing high predicted RIS (Figure 2C, 2D, 9). The 3D predicted structures reveal that the designed peptides are positioned close to the target motifs. High pTM and ipTM scores, and AutoDock VINA docking scores further suggest high binding affinities (Table 3). These results indicate that moPPIt can effectively design binders targeting both ordered and disordered regions of structurally disordered proteins.

3 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 tackles this challenge by enabling the design of motif-specific peptide binders without relying on structural representations. By leveraging pLM embeddings and conditional uniform discrete diffusion, moPPIt demonstrates the ability to generate peptides that bind to user-defined epitopes across a wide 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 a chimeric peptide-E3 ubiquitin ligase ubiquibody (uAb) architecture for target degradation studies Brixi et al. (2023); Chen et al. (2023a); Bhat et al. (2025). 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 hold great promise for both detection and therapeutic applications, enabling precise modulation of protein function in diseases.

REFERENCES

Mohammad Hadi Abbasian, Mohammadamin Mahmanzar, Karim Rahimian, Bahar Mahdavi, Samaneh Tokhanbigli, Bahman Moradi, Mahsa Mollapour Sisakht, and Youping Deng. Global landscape of sars-cov-2 mutations and conserved regions. *Journal of Translational Medicine*, 21(1), February 2023. ISSN 1479-5876. doi: 10.1186/s12967-023-03996-w. URL http://dx.doi.org/10.1186/s12967-023-03996-w.

Osama Abdin, Satra Nim, Han Wen, and Philip M Kim. Pepnn: a deep attention model for the identification of peptide binding sites. *Communications biology*, 5(1):503, 2022.

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper. Accurate structure prediction of biomolecular interactions with alphafold3. *Nature*, May 2024a. ISSN 1476-4687. doi: 10.1038/s41586-024-07487-w. URL http://dx.doi.org/10.1038/s41586-024-07487-w.

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pp. 1–3, 2024b.

David O. Azorsa, Peter K. Bode, Marco Wachtel, Adam Tai Chi Cheuk, Paul S. Meltzer, Christian Vokuhl, Ulrike Camenisch, Huy Leng Khov, Beata Bode, Beat W. Schäfer, and Javed Khan. Immunohistochemical detection of pax-foxo1 fusion proteins in alveolar rhabdomyosarcoma using breakpoint specific monoclonal antibodies. *Modern Pathology*, 34(4):748–757, April 2021. ISSN 0893-3952. doi: 10.1038/s41379-020-00719-0. URL http://dx.doi.org/10.1038/s41379-020-00719-0.

Suhaas Bhat, Kalyan Palepu, Lauren Hong, Joey Mao, Tianzheng Ye, Rema Iyer, Lin Zhao, Tianlai Chen, Sophia Vincoff, Rio Watson, Tian Z. Wang, Divya Srijay, Venkata Srikar Kavirayuni, Kseniia Kholina, Shrey Goel, Pranay Vure, Aniruddha J. Deshpande, Scott H. Soderling, Matthew P. DeLisa, and Pranam Chatterjee. De novo design of peptide binders to conformationally diverse targets with contrastive language modeling. *Science Advances*, 11(4), January 2025. ISSN 2375-2548. doi: 10.1126/sciadv.adr8638. URL http://dx.doi.org/10.1126/sciadv.adr8638.

Garyk Brixi, Tianzheng Ye, Lauren Hong, Tian Wang, Connor Monticello, Natalia Lopez-Barbosa, Sophia Vincoff, Vivian Yudistyra, Lin Zhao, Elena Haarer, et al. Salt&peppr is an interface-predicting language model for designing peptide-guided protein degraders. *Communications Biology*, 6(1):1081, 2023.

Patrick Bryant and Arne Elofsson. Peptide binder design with inverse folding and protein structure prediction. *Communications Chemistry*, 6(1), October 2023. ISSN 2399-3669. doi: 10.1038/s42004-023-01029-7. URL http://dx.doi.org/10.1038/s42004-023-01029-7.

Anton Bushuiev, Roman Bushuiev, Petr Kouba, Anatolii Filkin, Marketa Gabrielova, Michal Gabriel, Jiri Sedlar, Tomas Pluskal, Jiri Damborsky, Stanislav Mazurenko, et al. Learning to design protein-protein interactions with enhanced generalization. *arXiv preprint arXiv:2310.18515*, 2023.

Giorgia Castellani, Mariachiara Buccarelli, Maria Beatrice Arasi, Stefania Rossi, Maria Elena Pisanu, Maria Bellenghi, Carla Lintas, and Claudio Tabolacci. Braf mutations in melanoma: Biological aspects, therapeutic implications, and circulating biomarkers. *Cancers*, 15(16):4026, August 2023. ISSN 2072-6694. doi: 10.3390/cancers15164026. URL http://dx.doi.org/10.3390/cancers15164026.

Dailu Chen, Kenneth W. Drombosky, Zhiqiang Hou, Levent Sari, Omar M. Kashmer, Bryan D. Ryder, Valerie A. Perez, DaNae R. Woodard, Milo M. Lin, Marc I. Diamond, and Lukasz A. Joachimiak. Tau local structure shields an amyloid-forming motif and controls aggregation propensity. *Nature Communications*, 10(1), June 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-10355-1. URL http://dx.doi.org/10.1038/s41467-019-10355-1.

Tianlai Chen, Madeleine Dumas, Rio Watson, Sophia Vincoff, Christina Peng, Lin Zhao, Lauren Hong, Sarah Pertsemlidis, Mayumi Shaepers-Cheu, Tian Zi Wang, Divya Srijay, Connor Monticello, Pranay Vure, Rishab Pulugurta, Kseniia Kholina, Shrey Goel, Matthew P. DeLisa, Ray Truant, Hector C. Aguilar, and Pranam Chatterjee. Pepmlm: Target sequence-conditioned generation of therapeutic peptide binders via span masked language modeling, 2023a. URL https://arxiv.org/abs/2310.03842.

Tianlai Chen, Lauren Hong, Vivian Yudistyra, Sophia Vincoff, and Pranam Chatterjee. Generative design of therapeutics that bind and modulate protein states. *Current Opinion in Biomedical Engineering*, 28:100496, December 2023b. ISSN 2468-4511. doi: 10.1016/j.cobme.2023.100496. URL http://dx.doi.org/10.1016/j.cobme.2023.100496.

Jerome Eberhardt, Diogo Santos-Martins, Andreas F Tillack, and Stefano Forli. Autodock vina 1.2. 0: New docking methods, expanded force field, and python bindings. *Journal of chemical information and modeling*, 61(8):3891–3898, 2021.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, July 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL http://dx.doi.org/10.1038/s41586-021-03819-2.

Shinobu Kitazume, Yuriko Tachida, Ritsuko Oka, Keiro Shirotani, Takaomi C. Saido, and Yasuhiro 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*, 98(24):13554–13559, November 2001. ISSN 1091-6490. doi: 10.1073/pnas.241509198. URL http://dx.doi.org/10.1073/pnas.241509198.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Corinne M. Linardic. Pax3–foxo1 fusion gene in rhabdomyosarcoma. *Cancer Letters*, 270(1):10–18, October 2008. ISSN 0304-3835. doi: 10.1016/j.canlet.2008.03.035. URL http://dx.doi.org/10.1016/j.canlet.2008.03.035.

Caixuan Liu, Kejia Wu, Hojun Choi, Hannah Han, Xulie Zhang, Joseph L. Watson, Sara Shijo, Asim K. Bera, Alex Kang, Evans Brackenbrough, Brian Coventry, Derrick R. Hick, Andrew N. Hoofnagle, Ping Zhu, Xingting Li, Justin Decarreau, Stacey R. Gerben, Wei Yang, Xinru Wang, Mila Lamp, Analisa Murray, Magnus Bauer, and David Baker. Diffusing protein binders to intrinsically disordered proteins. *bioRxiv*, July 2024. doi: 10.1101/2024.07.16.603789. URL http://dx.doi.org/10.1101/2024.07.16.603789.

Haiying Lu, Qiaodan Zhou, Jun He, Zhongliang Jiang, Cheng Peng, Rongsheng Tong, and Jianyou Shi. Recent advances in the development of protein–protein interactions modulators: mechanisms and clinical trials. *Signal Transduction and Targeted Therapy*, 5(1), September 2020. ISSN

2059-3635. doi: 10.1038/s41392-020-00315-3. URL http://dx.doi.org/10.1038/s41392-020-00315-3.

Zhangzhi Peng, Benjamin Schussheim, and Pranam Chatterjee. Ptm-mamba: A ptm-aware protein language model with bidirectional gated mamba blocks. *bioRxiv*, February 2024. doi: 10.1101/2024.02.28.581983. URL http://dx.doi.org/10.1101/2024.02.28.581983.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. *arXiv preprint arXiv:2406.07524*, 2024.

Yair Schiff, Subham Sekhar Sahoo, Hao Phung, Guanghan Wang, Sam Boshar, Hugo Dalla-torre, Bernardo P de Almeida, Alexander Rush, Thomas Pierrot, and Volodymyr Kuleshov. Simple guidance mechanisms for discrete diffusion models. *arXiv preprint arXiv:2412.10193*, 2024.

Yuval Sedan, Orly Marcu, Sergey Lyskov, and Ora Schueler-Furman. Peptiderive server: derive peptide inhibitors from protein–protein interactions. *Nucleic Acids Research*, 44(W1):W536–W541, May 2016. ISSN 1362-4962. doi: 10.1093/nar/gkw385. URL http://dx.doi.org/10.1093/nar/gkw385.

Alexander 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*, 24(7):6187, March 2023. ISSN 1422-0067. doi: 10.3390/ijms24076187. URL http://dx.doi.org/10.3390/ijms24076187.

Kelly D Sullivan, Matthew D Galbraith, Zdenek Andrysik, and Joaquin M Espinosa. Mechanisms of transcriptional regulation by p53. *Cell Death amp; Differentiation*, 25(1):133–143, November 2017. ISSN 1476-5403. doi: 10.1038/cdd.2017.174. URL http://dx.doi.org/10.1038/cdd.2017.174.

Joseph L. Watson, David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, Woody Ahern, Andrew J. Borst, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Nikita Hanikel, Samuel J. Pellock, Alexis Courbet, William Sheffler, Jue Wang, Preetham Venkatesh, Isaac Sappington, Susana Vázquez Torres, Anna Lauko, Valentin De Bortoli, Emile Mathieu, Sergey Ovchinnikov, Regina Barzilay, Tommi S. Jaakkola, Frank DiMaio, Minkyung Baek, and David Baker. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, July 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06415-8. URL http://dx.doi.org/10.1038/s41586-023-06415-8.

Chengxin Zhang, Xi Zhang, Peter L Freddolino, and Yang Zhang. Biolip2: an updated structure database for biologically relevant ligand–protein interactions. *Nucleic Acids Research*, 52(D1): D404–D412, 2024.

ACKNOWLEDGMENTS

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. We also thank Yair Schiff for providing guidance on UDLM design for moPPIt.

A SUPPLEMENTARY MATERIAL

A.1 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. The dataset was divided into training, validation, and test sets at an60/20/20 ratio.

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 divided into training, validation, and test sets at an 80/10/10 ratio.

The dataset for PepUDLM training was curated from the PepNN, BioLip2, and PPIRef dataset Abdin et al. (2022); Zhang et al. (2024); Bushuiev et al. (2023). All peptides from PepNN and BioLip2 were included, along with sequences from PPIRef ranging from 6 to 50 amino acids in length. The dataset was divided into training, validation, and test sets at an 80/10/10 ratio.

A.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 Lin et al. (2023). 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:

$$\text{Attention}_{\text{target}}(Q, K, V_{\text{binder}}) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V_{\text{binder}} \tag{4}$$

$$\text{Attention}_{\text{binder}}(Q, K, V_{\text{target}}) = \text{softmax}\left(\frac{KQ^T}{\sqrt{d_k}}\right)V_{\text{target}} \tag{5}$$

where d_k is the model dimension. In this way, both resulting target embedding and binder embedding will contain binder-target binding information.

A.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:

$$L(y, \hat{y}) = -\sum_i [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] + \lambda \sum_i y_i \log\left(\frac{y_i}{\hat{y}_i}\right) \tag{6}$$

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.

A.4 PepUDLM Training

Dynamic Batching. To enhance computational efficiency and manage variable-length token sequences, we implemented dynamic batching. Drawing inspiration from ESM-2's approach Lin et al. (2023), input peptide sequences were sorted by length to optimize GPU memory utilization, with a maximum token size of 100 per GPU.

PepUDLM employed a DDIT backbone model with a hidden layer size of 768, 12 blocks, 12 attention heads, and a dropout rate of 0.1. Training was conducted on a 2xH100 NVIDIA NVL GPU system with 94 GB of VRAM for 100 epochs. The AdamW optimizer was employed with a learning rate of 1e-5, weight decay of 1e-4, beta1 of 0.9, beta2 of 0.999, and epsilon of 1e-8. Gradient clipping was set to 1, and a learning rate scheduler with 10 warm-up epochs and cosine decay was used, with initial and minimum learning rates of 1e-5 and 1e-6, respectively.

A.5 Sampling Details

Sampling Hyper-parameters. The fine-tuned BindEvaluator and pre-trained PepUDLM were used in moPPIt to sample peptide candidates for in-silico benchmarking. The gamma hyper-parameter that controls the guidance strength was set to 2.0 during sampling. The total sampling steps was set to 32. Random seed was set to 42.

Binding Site identification. For targets without pre-existing binders, we applied APBS Electrostatic analysis to identify potential binding sites. Specifically, we select motifs where highly negative or highly positive electrostatic potentials are concentrated and caves are formed to facilitate binder interaction.

A.6 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.

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

Table 2: Test performance metrics of PepUDLM and an auto-regressive model trained on the same dataset.

Test Metrics	PepUDLM	PepAR
Loss	2.51	4.20
BPD	3.77	6.86
NLL	2.61	4.75
PPL	13.62	116.10

Table 3: **pTM and ipTM scores and VINA docking 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), GPCRs, and intrinsically disordered proteins. The designed binders and AutoDock VINA docking scores are provided alongside each protein.

UniProt ID	Protein Name	Type	ipTM score	pTM score	VINA Docking Score	Designed Binder
Q16671	AMHR2	Kinases	0.74	0.87	-7.8	SSSYPEP
P49759	CLK1		0.56	0.72	-8.6	DELPNEA
P53041	PPP5	Phosphatases	0.83	0.89	-6.5	TNTMNVSC
Q9UNI6	DUSP12		0.47	0.79	-6	AELLMQL
P63279	UBC9	DUBs	0.57	0.92	-5.9	DFLDD
Q9Y5K5	UCHL5		0.5	0.8	-7.2	GDGMTQGV
O43613	OX1R-TM3	GPCRs	0.57	0.72	-8.2	LFPSCMPEMV
	OX1R-TM5		0.58	0.73	-10	VWFDLSPIVS
	OX1R-TM7		0.59	0.73	-9.7	WEPLENAACL
P01106	MYC	Disordered	0.39	0.25	-5.8	EQPEWMDE
B1PRL2	EWS::FLI1		0.6	0.28	-4.9	PSRCREDC

Table 4: **Comparison of ipTM for existing and designed peptide-protein complexes.** The ipTM scores are calculated by AlphaFold3 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)	Designed Binder
1AYC	0.52	0.64	ARLIDDQLLKS
1B8Q	0.72	0.66	EVEFGFG
2Q8Y	0.52	0.51	ALRRELADW
3EQS	0.89	0.82	GDHARQGLLALG
3IDJ	0.66	0.72	LKWWWLL
3NIH	0.86	0.88	KLRIR
4EZN	0.53	0.59	PTSYPYETEPGVGMPYNPASVVP
4GNE	0.89	0.84	ARTKQTA
4IU7	0.94	0.92	KHLHLLLSAS
5E1C	0.85	0.86	HSHHHLRLLLQQSP
5EYZ	0.85	0.88	SSRSRLRKKETRL
5KRI	0.85	0.85	IHHHLLQLLQSEAT
5M02	0.55	0.56	GKPLNGAPV
7LUL	0.94	0.93	SWEDVWI
8CN1	0.94	0.92	SEAV

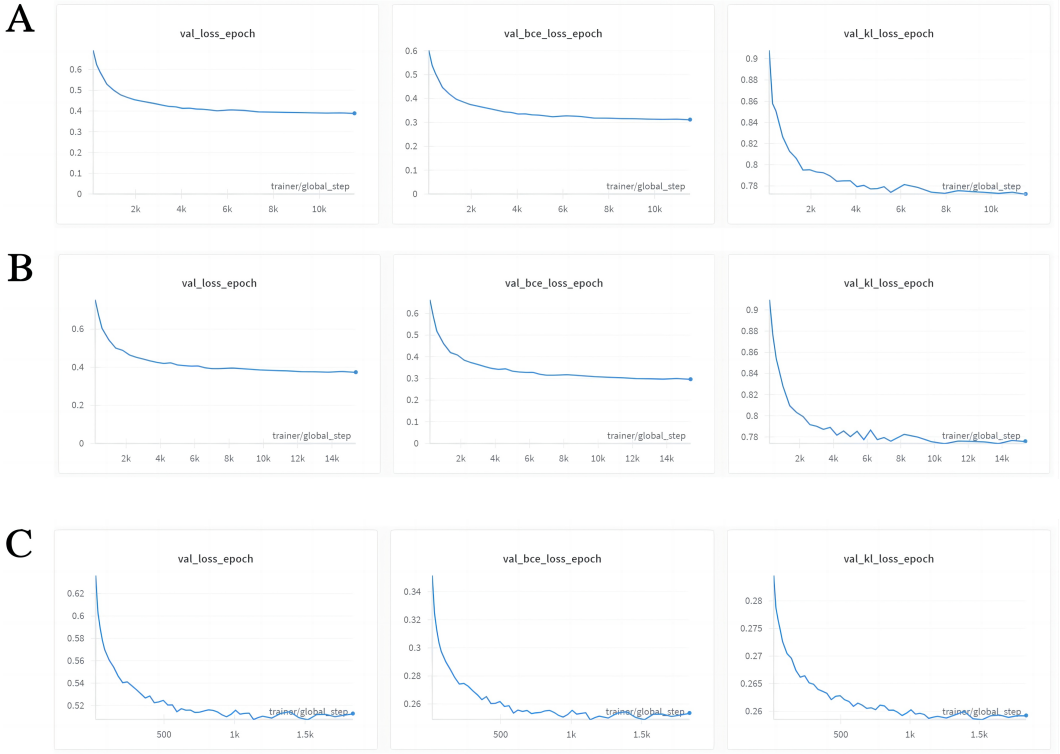

Figure 3: **Validation loss curves for BindEvaluator training and fine-tuning.** (A) Validation loss, binary cross-entropy (BCE) loss, and Kullback-Leibler (KL) divergence loss curves during training of BindEvaluator on the PPI dataset without dilated CNN modules. (B) Loss curves for training with dilated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

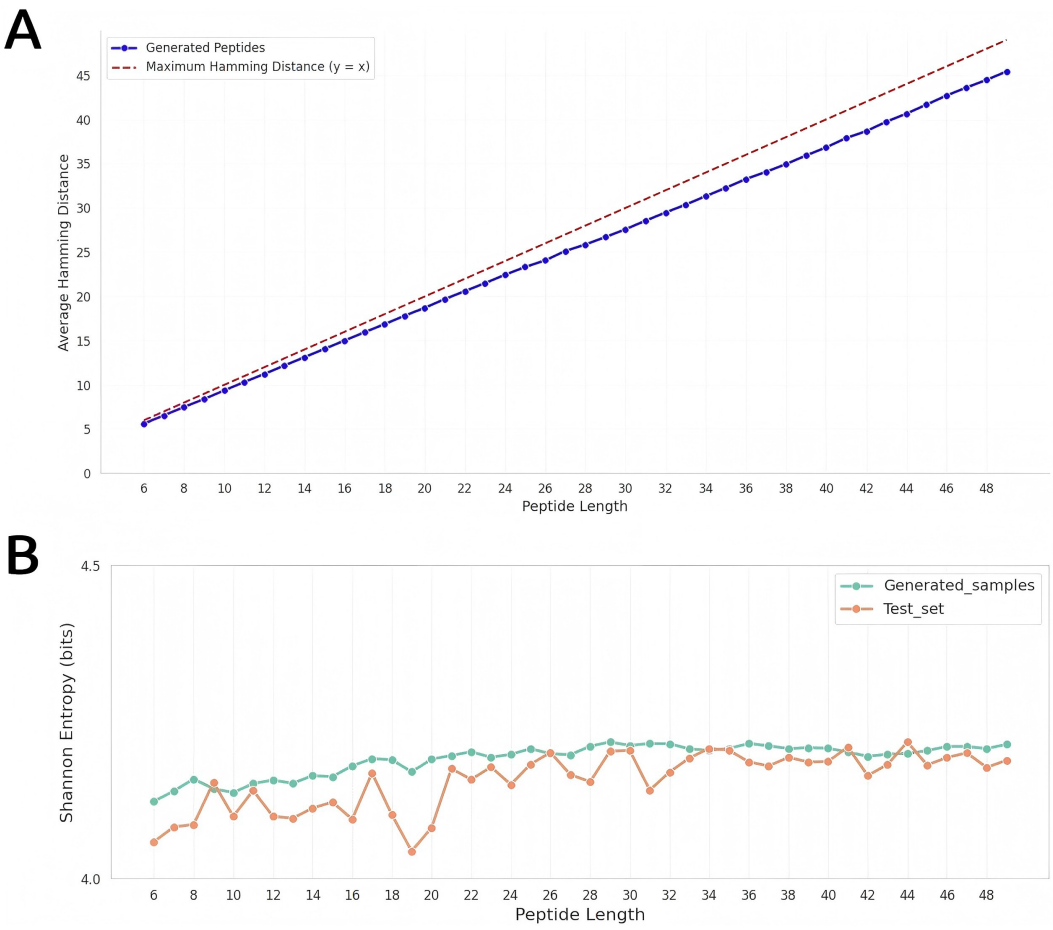

Figure 4: **(A)** The Hamming distance of sampled peptides of different lengths to the peptides of the same length in the test set. **(B)** The Shannon Entropy of sampled peptides of different lengths to the peptides of the same length in the test set.

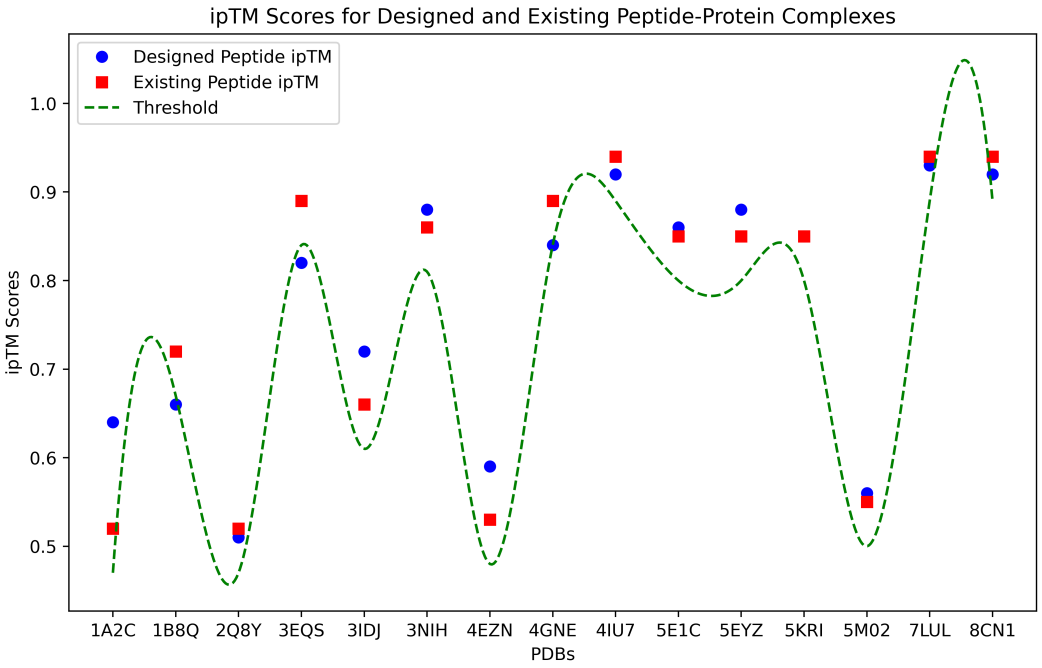

Figure 5: **Hit rate of moPPIt on structured targets with known binders.** The ipTM scores of input peptides, in complex with their target protein, were calculated via AlphaFold-Multimer. The ipTM scores for known peptides (red) from PDB structures were compared to moPPIt-designed peptides (blue) for 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.

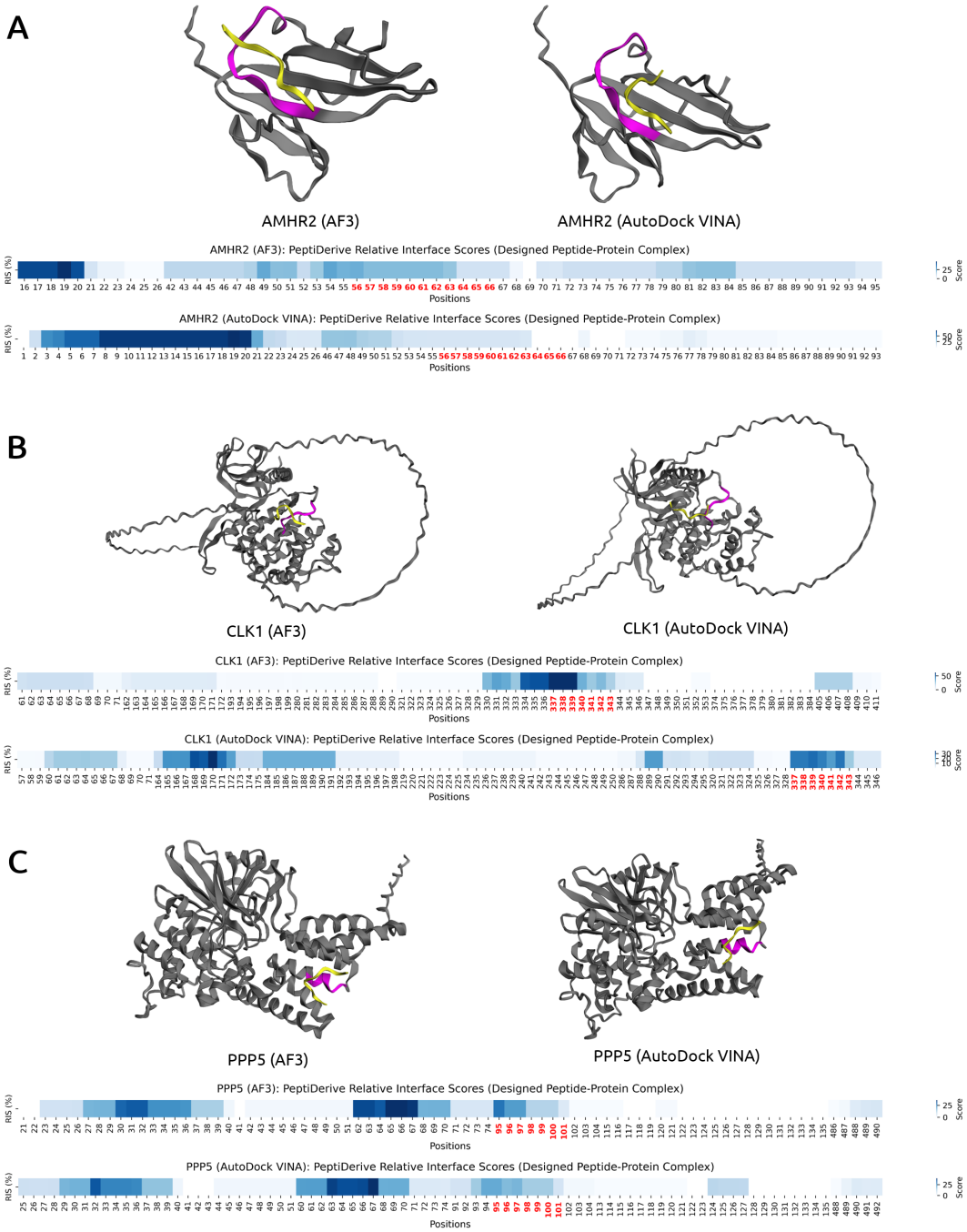

Figure 6: **Structural visualization and PeptiDerive relative interaction scores for designed peptides targeting structured motifs.** The peptide-protein complex structures are visualized for three proteins without known binders: **(A)** AMHR2, **(B)** CLK1, **(C)** PPP5 using AlphaFold3 and AutoDock VINA. The target proteins are depicted in grey, the designed peptide binders are shown in yellow, and the binding residues specified by 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.

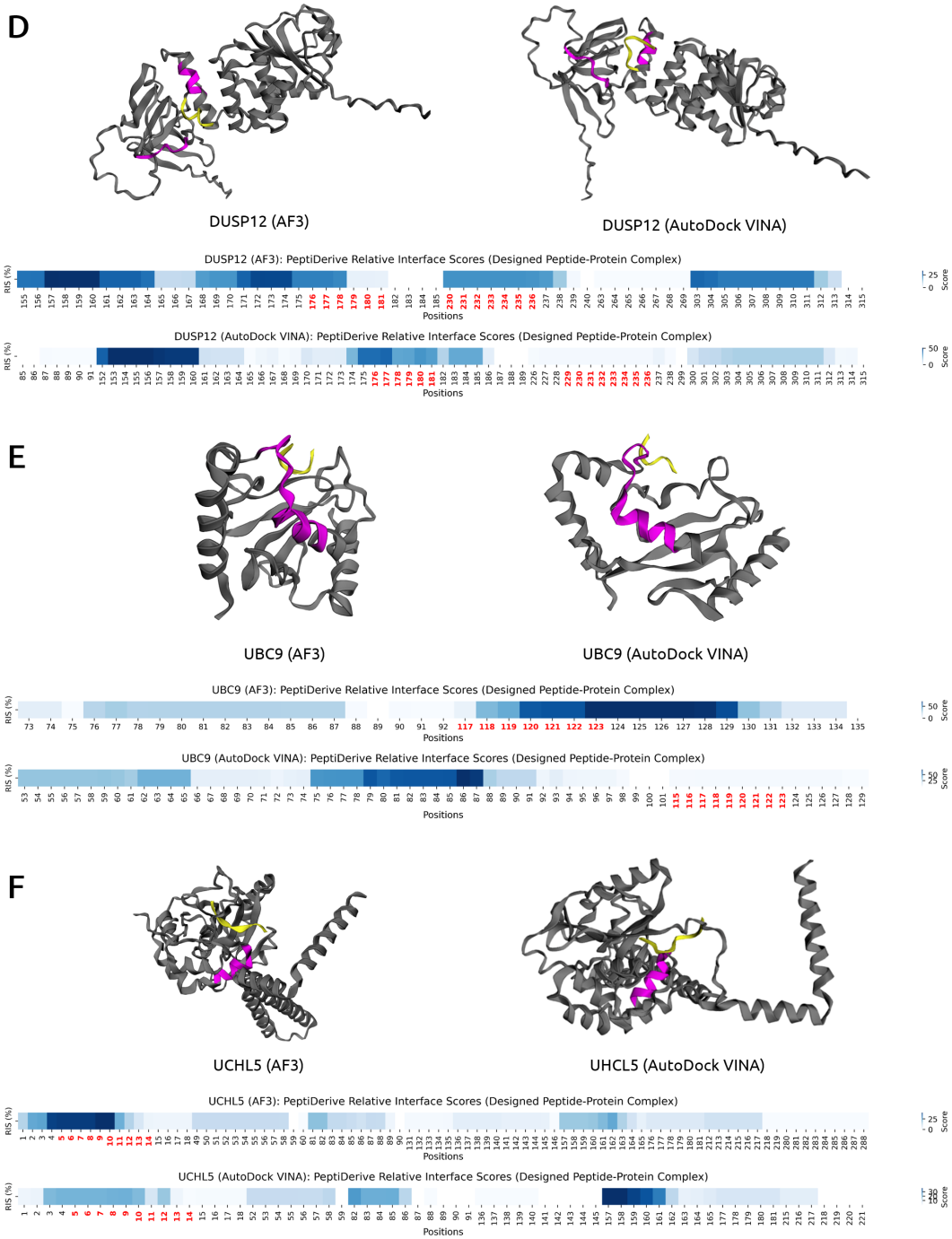

Figure 7: **Structural visualization and PeptiDerive relative interaction scores for designed peptides targeting structured motifs.** The peptide-protein complex structures are visualized for three proteins without known binders: **(D)** DUSP12, **(E)** UBC9, **(F)** UCHL5 using AlphaFold3 and AutoDock VINA. The target proteins are depicted in grey, the designed peptide binders are shown in yellow, and the binding residues specified by 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.

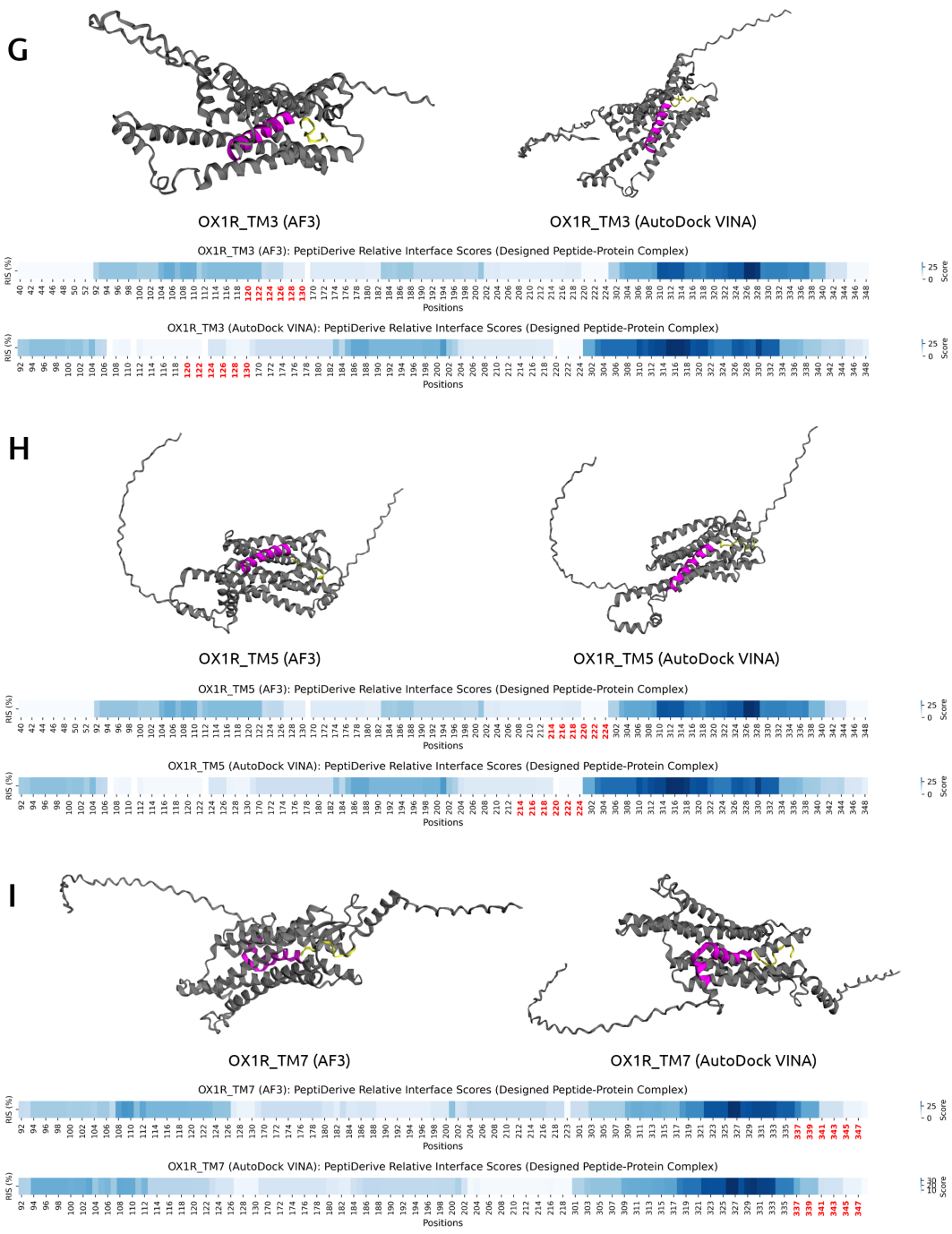

Figure 8: **Structural visualization and PeptiDerive relative interaction scores for designed peptides targeting structured motifs.** The peptide-complex structures are visualized for three different domains on OX1R: **(G)** Transmembrane (Name=3), **(H)** Transmembrane (Name=5), **(I)** Transmembrane (Name=7) using AlphaFold3 and AutoDock VINA. The target proteins are depicted in grey, the designed peptide binders are shown in yellow, and the binding residues specified by 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.

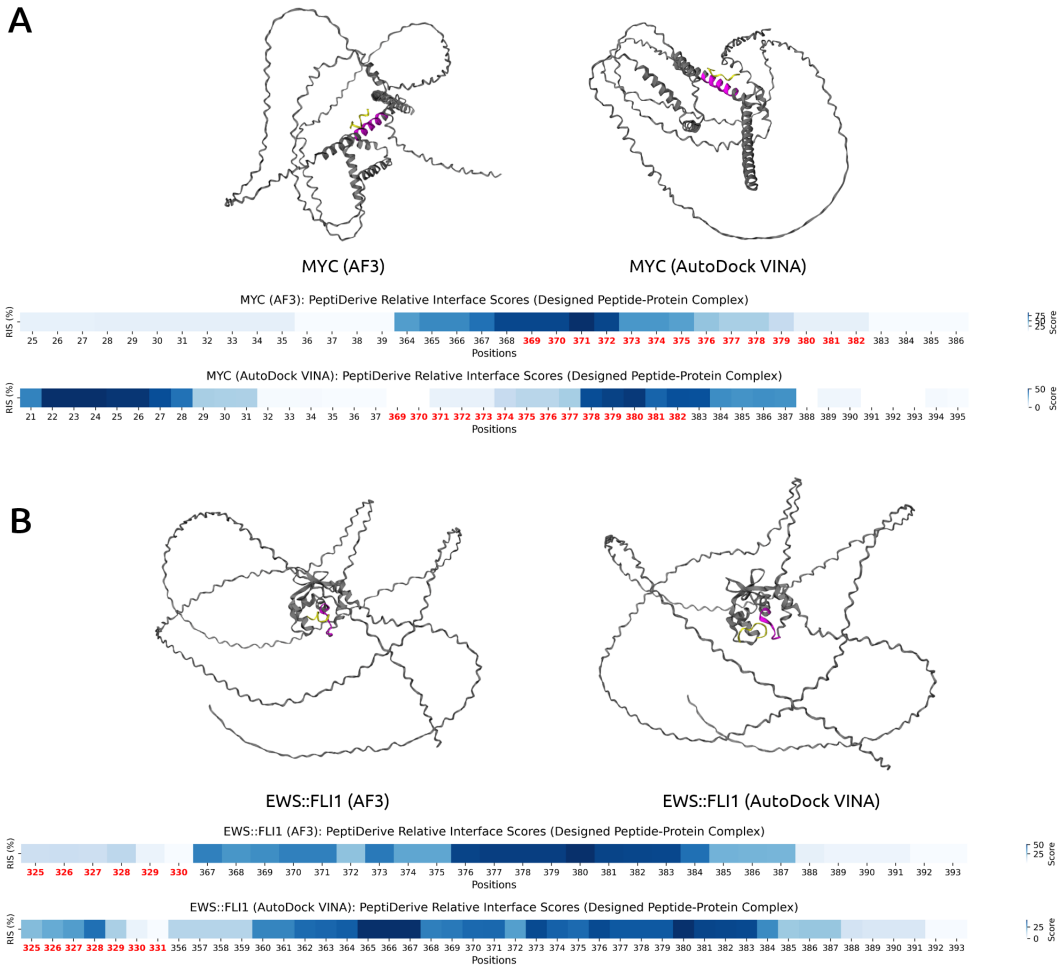

Figure 9: **Structural visualization and PeptiDerive relative interaction scores for designed peptides targeting intrinsically disordered proteins.** The peptide-complex structures are visualized for two intrinsically disordered proteins: **(A)** MYC, **(B)** EWS::FLI1 using AlphaFold3 and AutoDock VINA. The target proteins are depicted in grey, the designed peptide binders are shown in yellow, and the binding residues specified by 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.

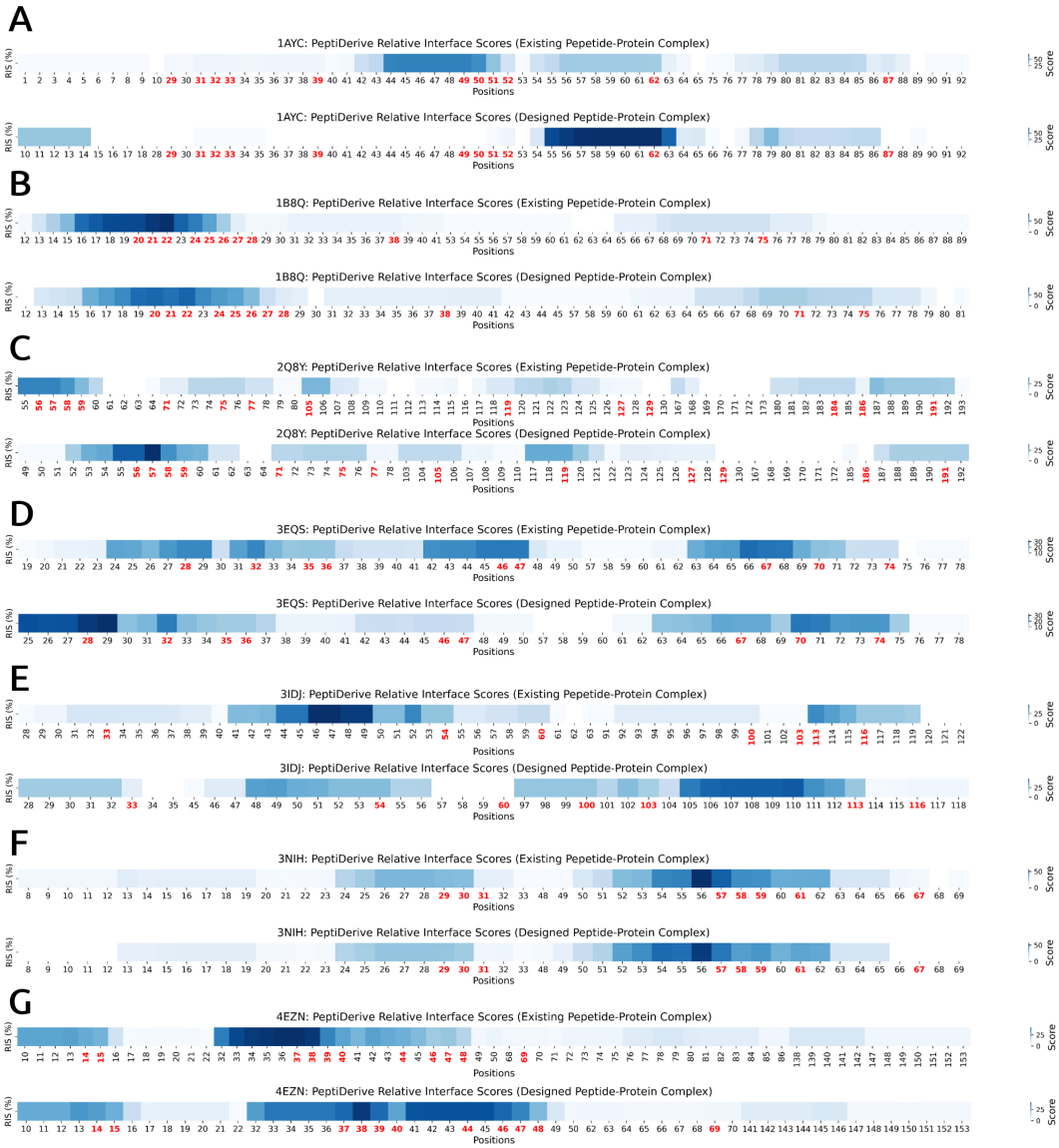

Figure 10: PeptiDerive relative interface scores for existing and designed peptide-protein complexes. Heatmaps of PeptiDerive relative interface scores (RIS) are shown for 7 peptide-protein complexes among 15 structured complexes with known binders that were tested: **(A)** 1AYC, **(B)** 1B8Q, **(C)** 2Q8Y, (**D**) 3EQS, **(E)** 3IDJ, (**F**) 3NIH, **(G)** 4EZN. 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.

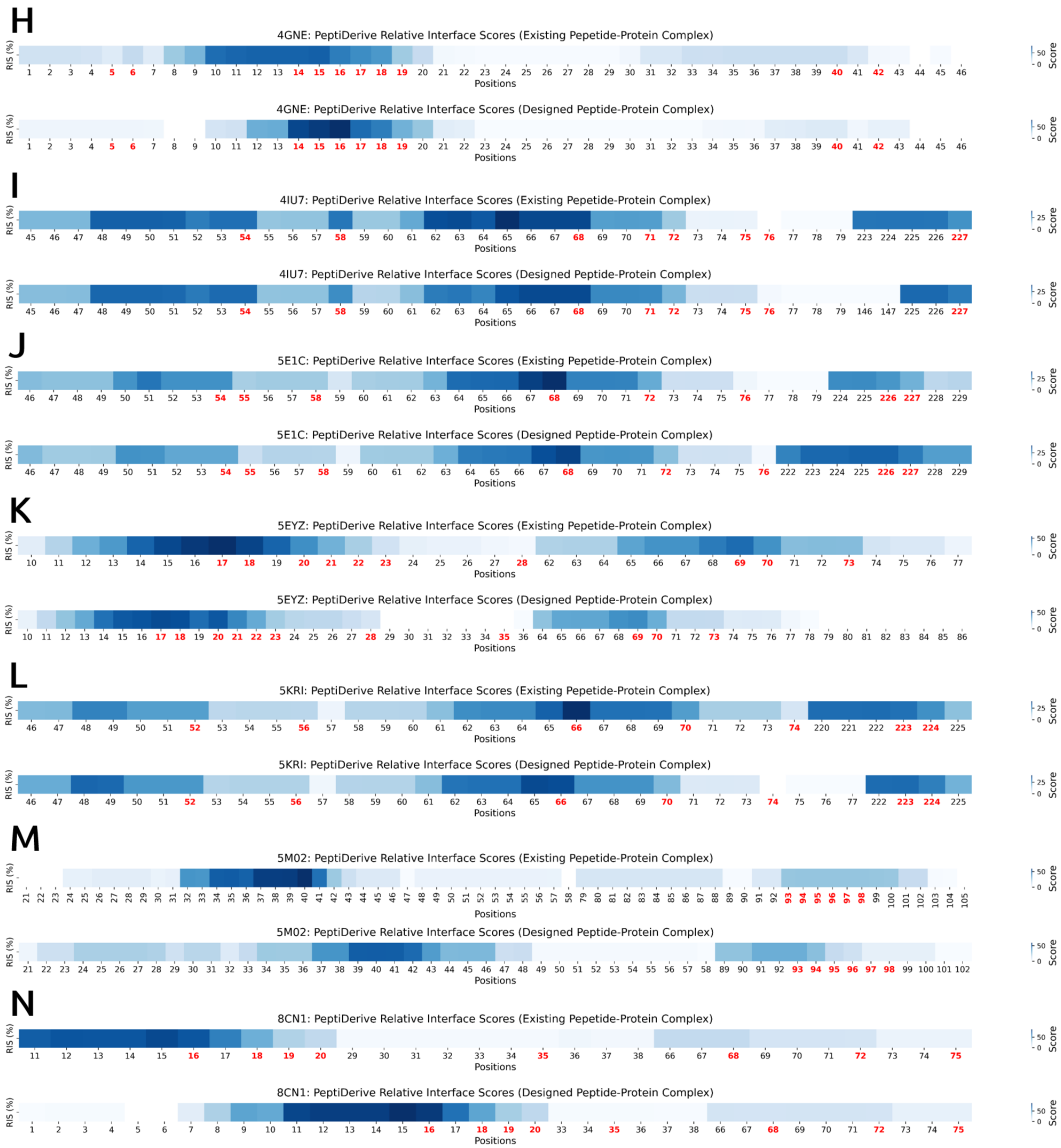

Figure 11: PeptiDerive relative interface scores for existing and designed peptide-protein complexes. Heatmaps of PeptiDerive relative interface scores (RIS) are shown for 7 peptide-protein complexes among 15 structured complexes with known binders that were tested: **(H)** 4GNE, **(I)** 4IU7, **(J)** 5E1C, **(K)** 5EYZ, **(L)** 5KRI, **(M)** 5M02, **(N)** 8CN1. 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.