moPPIt-v3: Motif-Specific Peptides Generated via Multi-Objective-Guided Discrete Flow Matching

Tong Chen,[1,*] **Zachary Quinn,**[2,*] **Yinuo Zhang,**[3]
Pranam Chatterjee[1,4,†]

[1]Department of Computer and Information Science, University of Pennsylvania
[2]Department of Biomedical Engineering, Duke University
[3]Centre for Computational Biology, Duke-NUS Medical School, Singapore
[4]Department of Bioengineering, University of Pennsylvania

[*]These authors contributed equally
[†]Corresponding author: pranam@seas.upenn.edu

Abstract

Precise targeting of therapeutic proteins to specific subsequence motifs within disease-related targets, such as conserved viral epitopes or mutant transcriptional domains, is critical for improving treatment efficacy and minimizing off-target interactions. Current computational binder design methods struggle to achieve this specificity, especially without reliable structural information. Here, we introduce **moPPIt-v3**, a generative, sequence-only model capable of the *de novo* design of high-affinity, motif-specific peptide binders. By coupling our **M**ulti-**O**bjective-**G**uided **D**iscrete **F**low **M**atching (**MOG-DFM**) framework with a residue-level interaction predictor, BindEvaluator, and a pretrained affinity predictor, we can guide peptide generation towards both sequence specificity and binding affinity. BindEvaluator is a transformer-based model, trained on over 510,000 annotated protein-protein interactions, 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. 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 integrating BindEvaluator, we demonstrate moPPIt-v3's *in silico* efficacy by designing high-quality binders to specific motifs within target sequences with and without known peptide binders, including both structured and disordered targets. Moreover, we validate the motif-specificity of moPPIt-generated peptides *in vitro* by showing sensitive and specific binding toward distinct domains of cancer receptor NCAM1. Altogether, moPPIt-v3 is a powerful tool for developing highly-specific peptide therapeutics without relying on target structure or known binding partner.

1 Introduction

Motif-specific targeting of protein–protein interactions (PPIs) enables highly selective biotherapeutics that modulate protein function while minimizing off-target effects, unlike traditional small molecules that rely on well-defined ligand binding sites [1]. Such specificity is critical in many therapeutic contexts, from restoring mutant p53 function in cancer [2], inhibiting BRAF kinase in melanoma, [3] or blocking tau aggregation in tauopathies [4]. Experimental binder discovery methods, such as animal immunization, phage, or yeast display, are costly and labor-intensive, motivating more efficient

Published at Frontiers in Probabilistic Inference: Sampling Meets Learning (NeurIPS 2025).

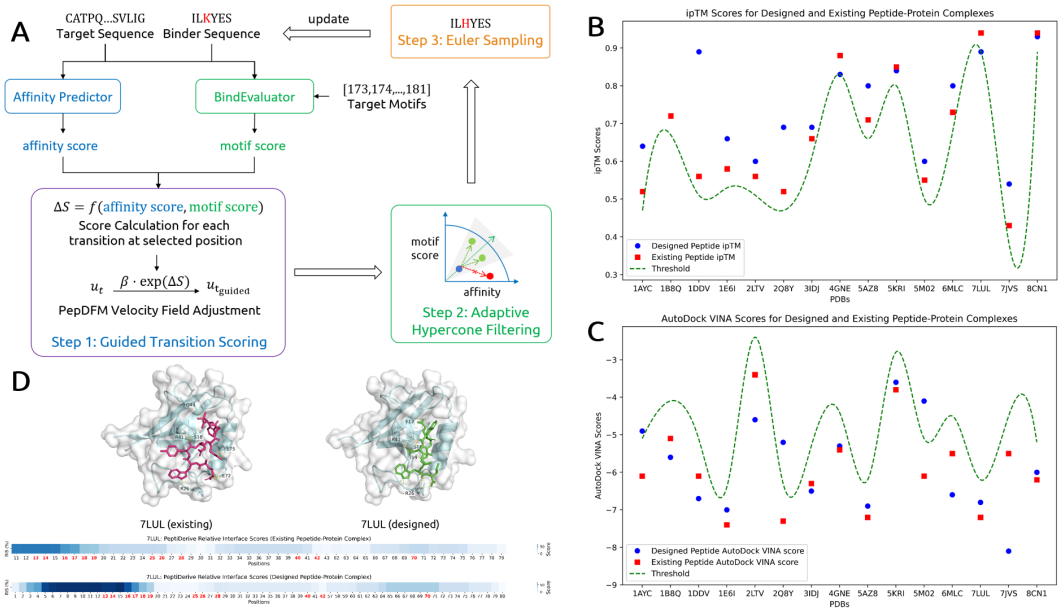


Figure 1: **moPPIt. (A)** Schematic of moPPIt. **(B), (C)** Hit rates on structured targets with known binders, evaluated by ipTM and AutoDock VINA. Scores of known peptides (red) from PDB structures were compared with moPPIt-designed peptides (blue). Hits were defined as ipTM within 0.05 or VINA within 1.0 of the known peptide score (green line). **(D)** AutoDock VINA docking visualization of protein (PDB ID: 7LUL) with existing and designed peptide binders, highlighting interacting residues.

computational approaches [5]. Structure-based models such as RFDiffusion and BindCraft show promise for motif-specific binder design [6, 7], but they are not suited for intrinsically disordered proteins (IDPs), which are generally underrepresented in their training sets. Prior **mo**tif-specific **PPI t**argeting (**moPPIt**) algorithms, moPPIt-v1 and v2 [8, 9], addressed this gap but lacked optimization for other key molecular properties like binding affinity. To overcome this, we extend the Multi-Objective-Guided Discrete Flow Matching (MOG-DFM) framework [10], which excels at guiding sequence-based peptide generation across multiple objectives, to enable motif-specific binder design. As such, our key contributions are as follows:

1. **moPPIt-v3**, a sequence-only algorithm capable of generating both high-affinity and motif-specific peptide binders of varying lengths across diverse protein targets.
2. **BindEvaluator**, a pre-trained transformer model that accurately predicts binding hotspots.
3. **Extension of MOG-DFM**, a multi-objective generation framework [10] that is extended to jointly optimize motif specificity and binding affinity for *de novo* peptide generation using BindEvalulator and a pre-trained binding affinity predictor.
4. ***In silico* and *in vitro* validation results.** Using AlphaFold3, AutoDock VINA, and PeptiDerive [11–13], we demonstrate moPPIt-v3's ability to design epitope-specific binders across diverse targets, including kinases, transcription factors, G-coupled protein receptors (GPCRs), and intrinsically disordered regions (IDRs). These *in silico* findings are corroborated by *in vitro* validation of domain-specific binders against both full-length and truncated NCAM1.

2 Methods and Results

moPPIt-v3 applies MOG-DFM for motif-specific peptide binder generation

To enable motif-specific peptide binder generation, we first developed BindEvaluator, a transformer model that predicts binding motifs directly from target and binder sequences. Trained on a dataset containing 510,000 known PPIs, BindEvaluator achieves excellent test performance in predicting interacting residues (AUC > 0.94), especially when finetuned on peptide-protein binding data (AUC > 0.96) (Section A). While accurate, BindEvaluator does not capture thermodynamic binding strength

Table 1: Ablation results for binder design targeting 5AZ8 (PDB ID), 7JVS (PDB ID), and MYC with different guidance settings. For each setting, 100 binders were designed by moPPIt-v3 with lengths of 11, 11, and 8, respectively. The average scores are displayed.

Guidance Settings		**5AZ8**		**7JVS**		**MYC**	
Motif	**Affinity**	**Motif Score**	**Affinity Score**	**Motif Score**	**Affinity Score**	**Motif Score**	**Affinity Score**
✓	✓	0.7048	7.3871	0.7970	7.8295	0.4950	7.2433
✓	×	0.6990	6.1803	0.8273	6.4606	0.5325	5.8708
×	✓	0.5430	8.2470	0.4775	8.4952	0.1789	8.1899
×	×	0.4876	5.6212	0.5442	6.0628	0.2014	6.0884

and often produces low-affinity peptides. To ameliorate this, we extend the MOG-DFM framework, which steers pretrained discrete flow matching (DFM) generators toward Pareto-efficient trade-offs across multiple objectives [10]. MOG-DFM has shown strong performance in peptide binder design using PepDFM, an unconditional DFM for diverse peptide generation. In moPPIt-v3, we guide PepDFM with BindEvaluator [8, 9] and a pretrained affinity predictor (Section B), ensuring binders with both high specificity and affinity (Figure 1A).

The inputs to moPPIt-v3 include a target protein sequence, target motifs, and a specified binder length. At initialization, a peptide sequence is randomly generated and weight vectors are initialized via the Das-Dennis simplex lattice [14], with one weight vector chosen to guide the optimization towards the Pareto front in the current run. Beginning at a random amino acid position, BindEvaluator and the pretrained affinity predictor compute the motif and affinity scores for the current peptide and for all single-site variants at that position. These scores update the PepDFM velocity field, favoring transitions that improve specificity and affinity. To ensure each candidate token replacement steers the sequence towards the desired trade-off direction, adaptive hypercone filtering is applied, restricting candidate transitions to a cone around the weight vector. Finally, Euler sampling is employed to select the best candidate transition. After a fixed number of iterations, the optimized peptide is expected to achieve both high motif specificity and binding affinity.

moPPIt-v3 generates epitope-specific binders to target proteins

To validate that the MOG-DFM framework can balance trade-offs between motif specificity and binding affinity, we performed binder generation experiments targeting three proteins: 5AZ8, 7JVS, and MYC (Table 1). Ablation experiments reveal that removing one or both objectives dramatically decreases the corresponding property score, while only modestly improving the other. By contrast, enabling both guidance signals produces the most balanced binder profiles across motif and affinity scores. Notably, 5AZ8 and 7JVS are structured proteins with known binders, whereas MYC is a disordered protein lacking pre-existing binders, thus illustrating moPPIt-v3's ability to balance potentially conflicting objectives while designing novel binders to diverse targets with both high motif specificity and strong binding affinity.

To comprehensively evaluate moPPIt-v3 in a controlled setting, we designed binders for 15 structured proteins and compared them to known peptide binders from the PDB using AlphaFold3 ipTM scores and AutoDock VINA docking scores [11, 13]. In all cases, moPPIt-v3-generated binders achieved comparable or superior performance to known binders, with none of the designed peptides falling below the ipTM stability threshold (set at 0.05 lower than the reference complex) and only 3 out of 15 scoring below the docking threshold (set at 1.0 lower), demonstrating comparable binding affinity (Table 2, Figure 1B, C). PeptiDerive analysis [12] further revealed that compared to known complexes, moPPIt-v3-designed complexes had similar or higher relative interface scores (RIS), which quantifies the energy contribution of specific residues to the overall free energy of the binder-target complex. Crucially, high-RIS residues were concentrated near specified binding motifs, confirming both specificity and binding potential (Figure 7, 8). Together, these results highlight moPPIt-v3's ability to generate structurally stable, high-affinity, and motif-specific binders of diverse lengths (Table 2).

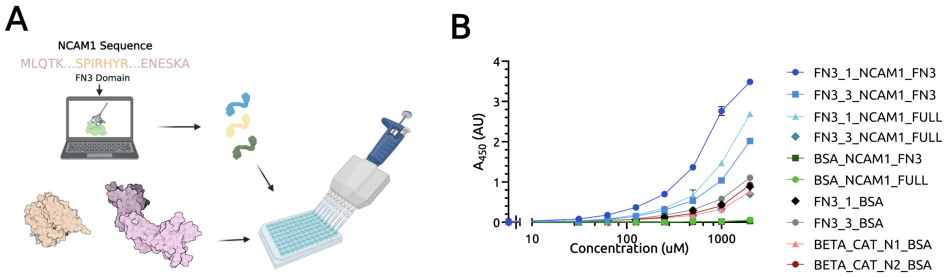


Figure 2: **Experimental validation of moPPIt-v3-designed peptide binders *in vitro*. (A)** Schematic of experimental pipeline. Peptides were designed to target the FN3 domain of the NCAM1 protein for downstream ELISA binding affinity measurements. **(B)** moPPIt-v3 peptides targeting the NCAM1 FN3 domain were expressed as C-terminal SUMO-tag fusions and screened for binding to NCAM1-FN3 and full-length NCAM1. ELISA was performed on the top two candidates, with NCAM1-FN3, full-length NCAM1, or bovine serum albumin (BSA) immobilized on 96-well plates, incubated with serial dilutions of biotinylated peptides or BSA, and detected via streptavidin–HRP. Absorbance at 450 nm (A_{450}) was measured, with data shown as mean ± s.e.m. (n=3 biological replicates).

To further assess moPPIt-v3'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-v3's versatility in designing binders for diverse structured proteins without pre-identified binding sites. Since these targets lack known binders, potential binding sites are identified using APBS electrostatic analysis [15]. We evaluated the epitope specificity of the designed binders to their respective targets (Figure 3, 4, 5, Table 3). Notably, residues at specified binding motifs aligned with high RIS from PeptiDerive, while AutoDock VINA scores and 3D docking visualizations further confirmed strong binding affinity and correct spatial placement of the designed peptides. To demonstrate moPPIt-v3's capacity to design binders against IDPs, we generated peptides toward two highly disordered proteins: MYC and EWS::FLI1. For both targets, PeptiDerive scores align with the specified binding motifs, showing high predicted RIS, while 3D structural models reveal the designed peptides positioned adjacent to these motifs (Figure 6). High ipTM scores and AutoDock VINA docking scores further suggest high binding affinities (Table 3). These results indicate that moPPIt-v3 can effectively design binders targeting both ordered and disordered regions of IDPs.

moPPIt-v3-generated binders show motif-specificity *in vitro*

To experimentally validate the motif-specific capabilities of moPPIt-v3, we generated peptide binders against neural cell adhesion molecule 1 (NCAM1), a key marker of acute myeloid leukemia [16]. NCAM1 is composed of multiple distinct folded domains, five consecutive immunoglobulin (IgG) domains followed immediately by a two fibronectin-type 3 (FN3) domains [17]. Binders targeting the FN3 domain will demonstrate interactions with both the full-length protein and the FN3 domain alone. To facilitate this characterization, the NCAM1-FN3 domain was expressed in *E. coli*, while four peptide binders against this FN3 domain were designed and expressed as C-terminal fusions to a SUMO-tag protein (Figure 2A). Two of these peptides (FN3_1 and FN3_3) showed specific interactions at mid-low nanomolar levels of NCAM1-FN3 compared to BSA and peptide controls (BETA_CAT_N1+2) as measured by ELISA (Figure 2B). Moreover, one of these peptides, FN3_1, shows potent, albeit slightly diminished, binding to the full-length NCAM1 protein (IgG+FN3).

3 Conclusion

Designing highly specific peptide binders for disease targets without well-defined structural pockets or with IDRs remains a major challenge in therapeutic development. Here we present **moPPIt-v3**, a sequence-only framework that leverages discrete flow matching with multi-objective optimization to design motif-specific peptide binders with high affinity, regardless of protein structure. moPPIt-v3 can generate peptides targeting user-defined epitopes across both structured and disordered protein

motifs, with forthcoming validation against structure-based methods like RFDiffusion [6, 18] through binding assays and ubiquibody-mediated degradation studies [5, 19, 20]. Its motif-specific capability enables applications for mutant-selective binding, targeting post-translational modification sites [21] and fusion oncoprotein breakpoints [22], and designing peptides against conserved viral epitopes less susceptible to escape mutations [23], highlighting its promise for both therapeutic and diagnostic interventions.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] Leo Tianlai Chen, Zachary Quinn, Madeleine Dumas, Christina Peng, Lauren Hong, Moises Lopez-Gonzalez, Alexander Mestre, Rio Watson, Sophia Vincoff, Lin Zhao, Jianli Wu, Audrey Stavrand, Mayumi Schaepers-Cheu, Tian Zi Wang, Divya Srijay, Connor Monticello, Pranay Vure, Rishab Pulugurta, Sarah Pertsemlidis, Kseniia Kholina, Shrey Goel, Matthew P. DeLisa, Jen-Tsan Ashley Chi, Ray Truant, Hector C. Aguilar, and Pranam Chatterjee. Target sequence-conditioned design of peptide binders using masked language modeling. *Nature Biotechnology*, August 2025. ISSN 1546-1696. doi: 10.1038/s41587-025-02761-2. URL http://dx.doi.org/10.1038/s41587-025-02761-2.

[6] 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.

[7] Martin Pacesa, Lennart Nickel, Christian Schellhaas, Joseph Schmidt, Ekaterina Pyatova, Lucas Kissling, Patrick Barendse, Jagrity Choudhury, Srajan Kapoor, Ana Alcaraz-Serna, Yehlin Cho, Kourosh H. Ghamary, Laura Vinué, Brahm J. Yachnin, Andrew M. Wollacott, Stephen Buckley, Adrie H. Westphal, Simon Lindhoud, Sandrine Georgeon, Casper A. Goverde, Georgios N. Hatzopoulos, Pierre Gönczy, Yannick D. Muller, Gerald Schwank, Daan C. Swarts, Alex J. Vecchio, Bernard L. Schneider, Sergey Ovchinnikov, and Bruno E. Correia. Bindcraft: one-shot design of functional protein binders. October 2024. doi: 10.1101/2024.09.30.615802. URL http://dx.doi.org/10.1101/2024.09.30.615802.

[8] Tong Chen, Yinuo Zhang, and Pranam Chatterjee. moppit: De novo generation of motif-specific binders with protein language models. *bioRxiv*, 2024.

[9] Tong Chen, Yinuo Zhang, Zachary Quinn, and Pranam Chatterjee. moppit: De novo generation of motif-specific peptide binders via conditional uniform discrete diffusion. In *Learning Meaningful Representations of Life (LMRL) Workshop at ICLR 2025*, 2025.

[10] Tong Chen, Yinuo Zhang, Sophia Tang, and Pranam Chatterjee. Multi-objective-guided discrete flow matching for controllable biological sequence design. *arXiv preprint arXiv:2505.07086*, 2025.

[11] 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*, 630(8016):493–500, 2024.

[12] 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.

[13] 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.

[14] Indraneel Das and John E Dennis. Normal-boundary intersection: A new method for generating the pareto surface in nonlinear multicriteria optimization problems. *SIAM journal on optimization*, 8(3):631–657, 1998.

[15] Nathan A Baker, David Sept, Simpson Joseph, Michael J Holst, and J Andrew McCammon. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proceedings of the National Academy of Sciences*, 98(18):10037–10041, 2001.

[16] Daniel Sasca, Jakub Szybinski, Andrea Schüler, Viral Shah, Jan Heidelberger, Patricia S. Haehnel, Anna Dolnik, Oliver Kriege, Eva-Marie Fehr, Wolf H. Gebhardt, George Reid, Claudia Scholl, Matthias Theobald, Lars Bullinger, Petra Beli, and Thomas Kindler. Ncam1 (cd56) promotes leukemogenesis and confers drug resistance in aml. *Blood*, 133(21):2305–2319, May 2019. ISSN 1528-0020. doi: 10.1182/blood-2018-12-889725. URL http://dx.doi.org/10.1182/blood-2018-12-889725.

[17] Federico Carafoli, Jane L. Saffell, and Erhard Hohenester. Structure of the tandem fibronectin type 3 domains of neural cell adhesion molecule. *Journal of Molecular Biology*, 377(2): 524–534, March 2008. ISSN 0022-2836. doi: 10.1016/j.jmb.2008.01.030. URL http://dx.doi.org/10.1016/j.jmb.2008.01.030.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] Sophia Vincoff, Shrey Goel, Kseniia Kholina, Rishab Pulugurta, Pranay Vure, and Pranam Chatterjee. Fuson-plm: a fusion oncoprotein-specific language model via adjusted rate masking. *Nature Communications*, 16(1), February 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-56745-6. URL http://dx.doi.org/10.1038/s41467-025-56745-6.

[23] 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.

[24] Anton Bushuiev, Roman Bushuiev, Petr Kouba, Anatolii Filkin, Marketa Gabrielova, Michal Gabriel, Jiri Sedlar, Tomas Pluskal, Jiri Damborsky, Stanislav Mazurenko, and Josef Sivic. Learning to design protein-protein interactions with enhanced generalization, 2023. URL https://arxiv.org/abs/2310.18515.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] Sangmin Seo, Jonghwan Choi, Seungyeon Choi, Jieun Lee, Chihyun Park, and Sanghyun Park. Pseq2sites: Enhancing protein sequence-based ligand binding-site prediction accuracy via the deep convolutional network and attention mechanism. *Engineering Applications of Artificial Intelligence*, 127:107257, 2024.

[30] Ruochi Zhang, Haoran Wu, Yuting Xiu, Kewei Li, Ningning Chen, Yu Wang, Yan Wang, Xin Gao, and Fengfeng Zhou. Pepland: a large-scale pre-trained peptide representation model for a comprehensive landscape of both canonical and non-canonical amino acids. *arXiv preprint arXiv:2311.04419*, 2023.

[31] Chakradhar Guntuboina, Adrita Das, Parisa Mollaei, Seongwon Kim, and Amir Barati Farimani. Peptidebert: A language model based on transformers for peptide property prediction. *The Journal of Physical Chemistry Letters*, 14(46):10427–10434, 2023.

[32] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *International Conference on Knowledge Discovery and Data Mining*, pages 2623–2631, 2019.

[33] Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky TQ Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. Discrete flow matching. *Advances in Neural Information Processing Systems*, 37:133345–133385, 2024.

[34] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

A BindEvaluator Methods and Results

A.1 Dataset Preparation

The training data for BindEvaluator was curated from the PPIRef dataset, a large and non-redundant database 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 a target sequence, a 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 were curated from the PepNN and BioLip2 databases [25, 26]. 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 an 80/10/10 ratio into a training set, validation set, and test set.

A.2 BindEvaluator accurately predicts target binding sites provided two interacting sequences

To enable motif-specific peptide binder generation, we developed a BindEvaluator model to predict peptide-protein binding sites (Figure 9A). 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 [27]. 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 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 [28] 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 (Figure 9B). 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 (Figure 9B). 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 (Figure 9B).

A.3 BindEvaluator Model Architecture

As shown in Figure 9A, BindEvaluator takes a binder sequence and a target sequence as inputs to predict the binding residues on the target protein. The design of this model draws inspiration from the architectures of PepNN and Pseq2Sites, which have demonstrated effectiveness in similar tasks [25, 29].

Both binder and target sequences are first passed into the pre-trained ESM-2-650M model to obtain their embeddings [27]. For the target sequence, a dilated CNN module captures the local features of adjacent residues. Specifically, the module is composed of three stacked CNN blocks with different dilation rates (1, 2, and 3) to extract hierarchical features. Each block consists of three convolutional layers with different kernel widths (3, 5, and 7) to cover different receptive field sizes, accommodating different binding site sizes. Padding is added to each convolutional layer to maintain consistent output and input sizes. Since the focus is on identifying binding residues for the target protein, the dilated

CNN module is applied only to the target sequence. Given that no binding motifs in the training set contain more than 23 continuous residues, the dilated CNN module is sufficient to capture the binding region features.

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

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

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

where d_k is the model dimension.

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

A.4 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:

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

Here, λ is a hyperparameter 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.

B Affinity Predictor

Dataset Preparation. We collected 1,781 binding affinity data for classifier training from the PepLand and PeptideBERT datasets [30, 31]. All sequences taken are wild-type L-amino acids and are tokenized and represented by the ESM-2 protein language model [27]. The dataset was split into a 0.8/0.2 ratio, maintaining similar affinity score distributions across splits.

Model Architecture. We developed an unpooled reciprocal attention transformer model to predict protein-peptide binding affinity, leveraging latent representations from the ESM-2 650M protein language model [27]. Instead of relying on pooled representations, the model retains unpooled token-level embeddings from ESM-2, which are passed through convolutional layers followed by cross-attention layers.

Training Details. We used OPTUNA [32] for hyperparameter optimization, tracing validation correlation scores. The final model was trained for 50 epochs with a learning rate of 3.84e-5, a dropout rate of 0.15, 3 initial CNN kernel layers (dimension 384), 4 cross-attention layers (dimension

2048), and a shared prediction head (dimension 1024) in the end. The classifier reached 0.64 Spearman's correlation score on validation data.

C Discrete Flow Matching

In the discrete setting, we consider data $x = (x_1, \ldots, x_d)$ taking values in a finite state space $S = \mathcal{T}^d$, where $\mathcal{T} = [K] = \{1, 2, \ldots, K\}$ is called the vocabulary. We model a continuous-time Markov chain (CTMC) $\{X_t\}_{t\in[0,1]}$ whose time-dependent transition rates $u_t(y, x)$ transport the probability mass from an initial distribution p_0 to a target distribution p_1 [33]. The marginal probability at time t is denoted $p_t(x)$, and its evolution is governed by the Kolmogorov forward equation

$$\frac{\mathrm{d}}{\mathrm{d}t}p_t(y) \;=\; \sum_{x\in S} u_t(y, x)\, p_t(x)\,. \tag{4}$$

The learnable velocity field $u_t(y, x)$ is defined as the sum of factorized velocities:

$$u_t(y, x) = \sum_i \delta(y^{\bar{i}}, x^{\bar{i}})u_t^i(y^i, x), \tag{5}$$

where $\bar{i} = (1, \ldots, i-1, i+1, \ldots, d)$ denotes all indices excluding i. The rate conditions for factorized velocities $u_t^i(y^i, x)$ are required per dimension $i \in [d]$:

$$u_t(y, x) \geq 0 \text{ for all } y^i \neq x^i, \text{ and } \sum_{y^i\in\mathcal{T}} u_t^i(y^i, x) = 0 \text{ for all } x \in S, \tag{6}$$

so that for small $h > 0$, the one-step kernel

$$p_{t+h|t}(y \mid x) = \delta(y, x) + h\, u_t(y, x) + o(h) \tag{7}$$

remains a proper probability mass function.

In practice, we can further parameterize the velocity field using a mixture path. Specifically, a mixture path is defined with scheduler $\kappa_t \in [0, 1]$ so that each coordinate X_t^i equals x_0^i or x_1^i with probabilities $1 - \kappa_t$ and κ_t, respectively. The mixture marginal velocity is then obtained by averaging the conditional rates over the posterior of (x_0, x_1) given $X_t = x$, yielding

$$u_t^i(y^i, x) = \sum_{x_1^i} \frac{\dot{\kappa}_t}{1 - \kappa_t}\left[\delta(y^i, x_1^i) - \delta(y^i, x^i)\right] p_{1|t}^i(x_1^i \mid x), \tag{8}$$

where $\dot{\kappa}_t$ denotes the time derivative of κ_t.

D PepDFM

Peptide **d**iscrete **f**low **m**atching model, **PepDFM**, is developed to generate biologically plausible peptide sequences unconditionally or with multi-objective guidance.

Model Architecture. The base model is a time-dependent architecture based on U-Net [34]. It uses two separate embedding layers for sequence and time, followed by five convolutional blocks with varying dilation rates to capture temporal dependencies, while incorporating time-conditioning through dense layers. The final output layer generates logits for each token. We used a polynomial convex schedule with a polynomial exponent of 2.0 for the mixture discrete probability path in the discrete flow matching.

Dataset Curation. The dataset for PepDFM training was curated from the PepNN, BioLip2, and PPIRef dataset [25, 26, 28]. All peptides from PepNN and BioLip2 were included, along with sequences from PPIRef ranging from 6 to 49 amino acids in length. The dataset was divided into training, validation, and test sets at an 80/10/10 ratio.

Training Strategy. The training is conducted on a 2xH100 NVIDIA NVL GPU system with 94 GB of VRAM for 200 epochs with batch size 512. The model checkpoint with the lowest evaluation loss was saved. The Adam optimizer was employed with a learning rate 1e-4. A learning rate scheduler with 20 warm-up epochs and cosine decay was used, with initial and minimum learning rates both

1e-5. The embedding dimension and hidden dimension were set to be 512 and 256 respectively for the base model.

Dynamic Batching. To enhance computational efficiency and manage variable-length token sequences, we implemented dynamic batching. Drawing inspiration from ESM-2's approach [27], input peptide sequences were sorted by length to optimize GPU memory utilization, with a maximum token size of 100 per GPU.

E moPPIt-v3 Formulation

moPPIt-v3 operates under the same setting as the discrete flow matching described in Section C. At its core, it leverages a pretrained discrete flow matching model, PepDFM, that defines a CTMC with a factorized velocity field $u_t^i(y^i, x)$, which transports probability mass from an initial distribution p_0 to the target distribution over plausible peptide sequences via mixture path parametrization. In addition, moPPIt-v3 uses two pre-trained score models, the affinity predictor s_1 and BindEvaluator s_2, that assign objective scores to any peptide sequence. The affinity predictor s_1 calculates the affinity score based on the peptide-protein pair, while s_2 predicts the motif score given a target protein sequence, a peptide binder sequence, and target motifs. Specifically, motif score is calculated as the average probability of each target motif residue participating in binding, as predicted by BindEvaluator:

$$s_2(x) = \frac{1}{n} \sum_{m_i \in M} \text{softmax(logits)}[m_i], \tag{9}$$

where M represents the target motifs.

We aim to generate novel sequences $x_1 \in \mathcal{S}$ whose objective vectors $\big(s_1(x_1),\ s_2(x_1)\big)$ lie near the Pareto front (not guaranteed to be Pareto optimal)

$$\text{PF} \ = \ \Big\{ x \in \mathcal{S} \ \Big| \ \nexists\, x' \in \mathcal{S} : \ s_n(x') \geq s_n(x)\ \forall n,\ \exists\, m : s_m(x') > s_m(x) \Big\}.$$

To achieve this, moPPIt-v3 applies the MOG-DFM framework that guides the CTMC sampling dynamics of PepDFM using multi-objective transition scores, steering the generative process toward Pareto-efficient regions of the state space (Figure 1A).

moPPIt-v3 begins by initializing the generative process at time $t = 0$ by sampling an initial sequence x_0 uniformly from the discrete state space $\mathcal{S} = [K]^d$. To steer the generation towards diverse Pareto-efficient solutions, we introduce a set of M weight vectors $\{\omega^k\}_{k=1}^M$, where $\omega \in \mathbb{R}^2$, that uniformly cover the 2-dimensional Pareto front. Intuitively, each ω encodes a particular trade-off among both objectives, so sampling different ω promotes exploration of distinct regions of the Pareto front. We construct these vectors via the Das–Dennis simplex lattice with H subdivisions, yielding components

$$\omega_i = \frac{k_i}{H}, \quad k_i \in \mathbb{Z}_{\geq 0}, \quad \sum_{i=1}^{N} k_i = H, \tag{10}$$

A single ω is then sampled randomly to define the optimization direction toward the Pareto front for the current run. Once initialized, moPPIt-v3 performs Step 1 (Guided Transition Scoring), Step 2 (Adaptive Hypercone Filtering), and Step 3 (Euler Sampling) over T iterations to generate a final sequence x_1 whose score vectors have been steered close to the Pareto front, with both objectives optimized. For detailed formulations of these steps, please refer to [10].

F Sampling Settings

The hyperparameters were set as follows during sampling: The number of divisions used in generating weight vectors, num_div, was set to 64, λ to 1.0, β to 1.0, α_r to 0.5, τ to 0.3, η to 1.0, Φ_{init} to 45°, Φ_{min} to 15°, Φ_{max} to 75°. The total sampling step T was 100. The importance vector was set to $[20, 1]$, corresponding to motif score and affinity score, respectively.

G Expression and purification of SUMO–peptide constructs

Peptides of interest were cloned into a pET-24a+ (Novagen) expression vector containing an N-terminal 6×-histidine–SUMO tag to facilitate downstream purification. Oligonucleotide primer pairs,

each encoding for one half of the peptide sequences, were designed using NEBaseChanger V2, then incorporated into the plasmid using Q5 site-directed mutagenesis, as per the manufacturer's instructions. Plasmid assembly was verified using Sanger sequencing (GENEWIZ) and then transformed into chemically competent Escherichia coli BL21(DE3) cells. Starter cultures (3ml of LB media, $50\mu g/ml$ kanamycin) were inoculated from freshly streaked agar plates or glycerol stocks and grown at 37°C with shaking at 225 r.p.m. overnight. Starter cultures were then diluted 1:500 in bulk cultures and grown to an optical density at 600 nm (OD600) of 0.6–0.8 and then induced at a concentration of 1 mM isopropyl β-d-thiogalactopyranoside (IPTG) overnight at 37 °C with shaking. Thirty minutes after induction, rifampicin was added to a final concentration of $150\mu g/ml$. Cells were then collected by centrifugation (4,500xg) at 4°C and washed twice with ice-cold 1× PBS. The resulting cell pellets were frozen at -20°C overnight, thawed to room temperature, and then lysed using BugBuster protein extraction reagent (Millipore Sigma, 70584-3) supplemented with recombinant lysozyme (Millipore Sigma, 71110-3) and benzonase endonuclease (Millipore Sigma, E1014-25KU) for 20 minutes at room temperature with gentle rocking. The corresponding lysate was diluted with lysis buffer (1× PBS, 20mM imidazole, 1× Halt protease inhibitor cocktail (Thermo Fisher Scientific, 78430)) and then centrifuged at 14,000xg for 30 minutes. The cleared supernatant was mixed end over end at 4°C for 30 minutes with HisPur Ni-NTA resin (Thermo Fisher Scientific, 88221) equilibrated with 20mM imidazole in 1× PBS. Resin was centrifuged at 700xg for 2 minutes and then washed three times with 50mM imidazole in 1× PBS. Protein was eluted with three consecutive washes with 500mM imidazole, concentrated (Millipore Sigma, 3K MWCO, UFC900308), and desalted using Zeba spin desalting columns (Thermo Fisher Scientific, 89892). Expression and purity of purified proteins in both the soluble and insoluble fractions, as well as purified fractions, were assessed using SDS-PAGE. Protein concentrations were quantified using a Qubit Protein Assay (Thermo Fisher Scientific, Q33211).

H Sandwich ELISA

Purified SUMO-tagged peptide constructs were coated onto 96-well plates (Corning, CLS9018) at a concentration of $5\mu g/ml$ in coating buffer (10mM phosphate, pH 7.4) at a volume of 50–100µl per well at 4°C overnight with gentle rocking. Plates were washed once with Tris-buffered saline supplemented with 0.05% Tween-20 (v/v) (TBS-T), then blocked with 300µl of SuperBlock in PBS (Thermo Fisher Scientific, 37516) per the manufacturer's instructions. Recombinant NCAM1 (Sino 10673-H08H) or NCAM1-FN3 were serially diluted in triplicate in SuperBlock with 0.05% Tween-20, after which 100µl of each solution was added to each well and incubated at room temperature with gentle rocking for 1 hour. Plates were then washed five times using 300µl of TBS-T per well and then incubated with 100µl of SA-HRP (Thermo Fisher Scientific, N100, diluted 1:10,000 SuperBlock with 0.05% Tween 20) for 1 hour at room temperature. Plates were again washed five times with 300µl of TBS-T and then incubated with 100µl per well of 3,3'-5,5'-tetramethylbenzidine substrate (1-Step Ultra TMB-ELISA; Thermo Fisher Scientific, 34029) for 30 minutes at room temperature with gentle rocking. Finally, the reaction was quenched with 100µl of 2N H2SO4, and absorbance at 450nm was immediately quantified using a Promega GloMax Discover plate reader.

Table 2: **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-v3 algorithm. The designed binders for each protein are presented.

PDB ID	ipTM score (existing binder)	ipTM score (designed binder)	VINA score (existing binder)	VINA score (designed binder)	Designed Binder
1AYC	0.52	0.64	-6.1	-4.9	YAYRYICYYCD
1B8Q	0.72	0.72	-5.1	-5.6	IVDWVCF
1DDV	0.56	0.89	-6.1	-6.7	RCVRWC
1E6I	0.58	0.66	-7.4	-7	GRWRC
2LTV	0.56	0.6	-3.4	-4.6	PTVEECSYWYHE
2Q8Y	0.52	0.69	-7.3	-5.2	WLSWCHVYC
3IDJ	0.66	0.69	-6.3	-6.5	IRRVRAP
4GNE	0.88	0.83	-5.4	-5.3	ARRVRWS
5AZ8	0.71	0.8	-7.2	-6.9	LRWEVYLVREV
5KRI	0.85	0.84	-3.8	-3.6	FAGMIVVNCIMR
5M02	0.55	0.6	-6.1	-4.1	PEVRWEVRD
6MLC	0.73	0.8	-5.5	-6.6	GRWYCW
7LUL	0.94	0.89	-7.2	-6.8	WEVTIWV
7JVS	0.43	0.54	-5.5	-8.1	CVGIICEIICP
8CN1	0.94	0.93	-6.2	-6	SAEV

Table 3: **pTM and ipTM scores and VINA docking scores for moPPIt-v3-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**	**VINA Score**	**Designed Binder**
Q16671	AMHR2	Kinases	0.73	-5.7	EFEYEEV
P49759	CLK1		0.5	-6.9	PEVAAKEEEVEC
P53041	PPP5	Phosphatases	0.71	-8.5	YFLVYNVC
Q9UNI6	DUSP12		0.52	-6.9	QTCRYVVEC
Q9Y5K5	UCHL5	DUB	0.58	-5.4	GDGMTQGV
O43613	OX1R-TM3	GPCRs	0.58	-8.8	GYYVKCVDDY
	OX1R-TM5		0.56	-8.2	MSYWCCCVGF
	OX1R-TM7		0.54	-7.7	ARYTYDWVYLFA
P01106	MYC	Disordered	0.55	-5.6	EVFYWTWW
B1PRL2	EWS::FLI1		0.67	-6	IDEVCRRW

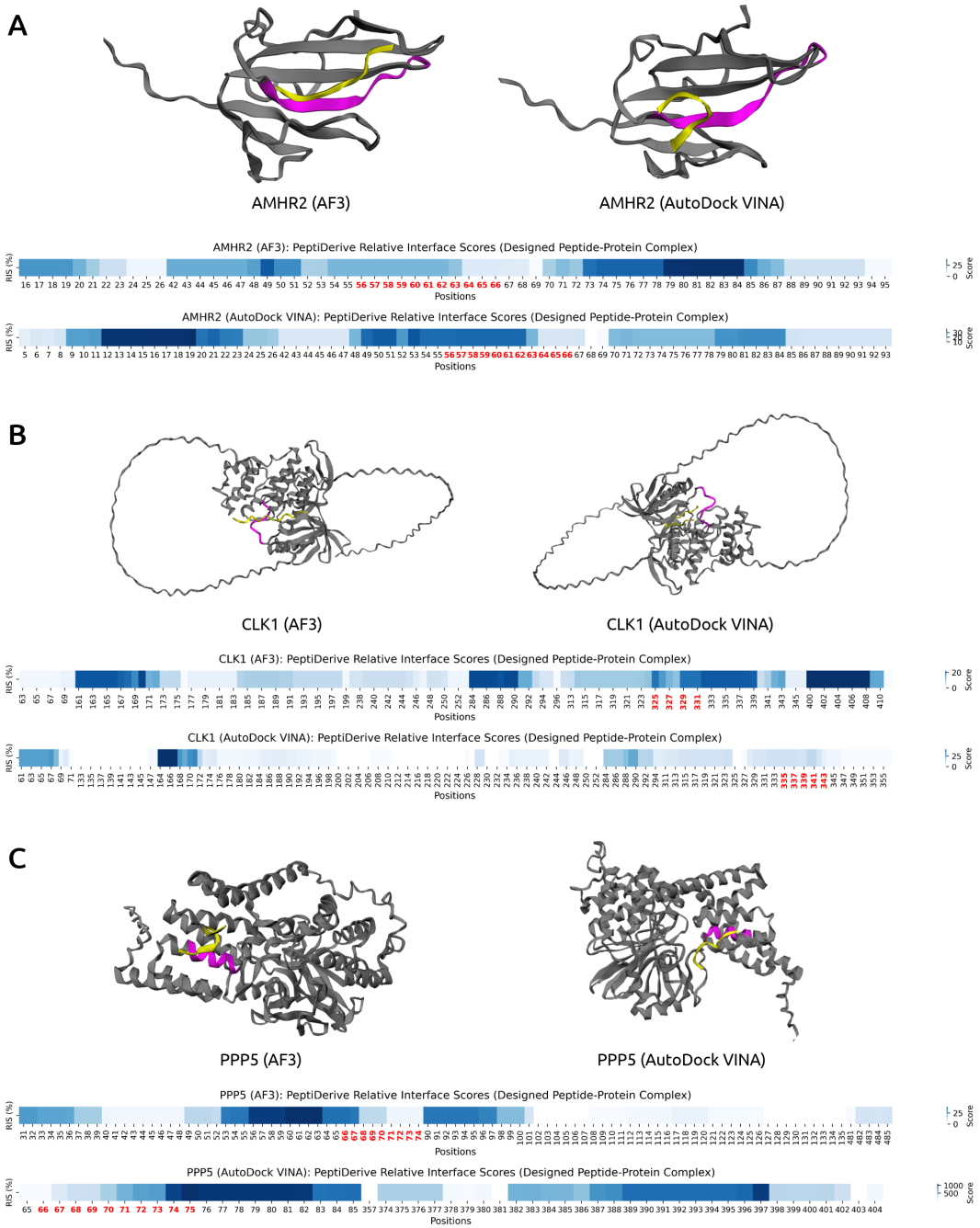


Figure 3: **Structural visualization and PeptiDerive relative interaction scores for designed peptides targeting structured motifs.** The peptide-protein complex structures are visualized for three proteins without known binders: **(A)** AMHR2, **(B)** CLK1, **(C)** PPP5 using AlphaFold3 and AutoDock VINA. The target proteins are depicted in grey, the designed peptide binders are shown in yellow, and the binding residues specified by the moPPIt-v3 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-v3 as the desired target amino acids.

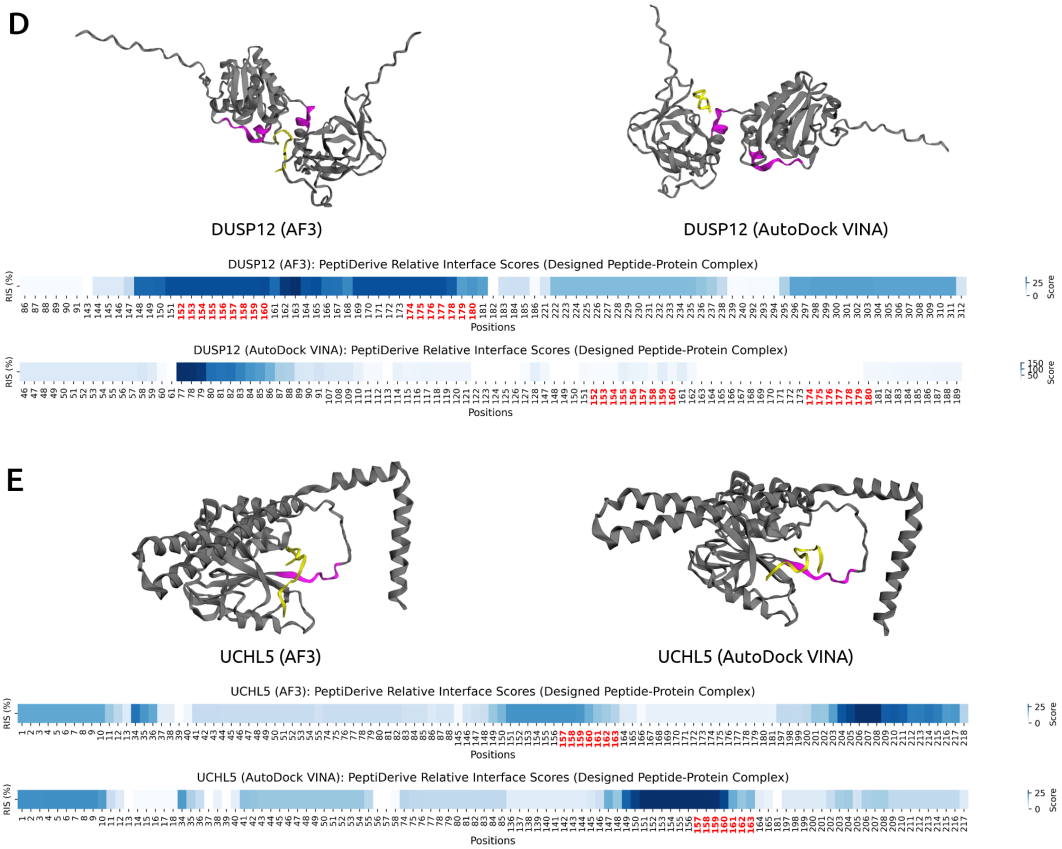


Figure 4: **Structural visualization and PeptiDerive relative interaction scores for designed peptides targeting structured motifs.** The peptide-protein complex structures are visualized for two proteins without known binders: **(D)** DUSP12, **(E)** UCHL5 using AlphaFold3 and AutoDock VINA. The target proteins are depicted in grey, the designed peptide binders are shown in yellow, and the binding residues specified by the moPPIt-v3 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-v3 as the desired target amino acids.

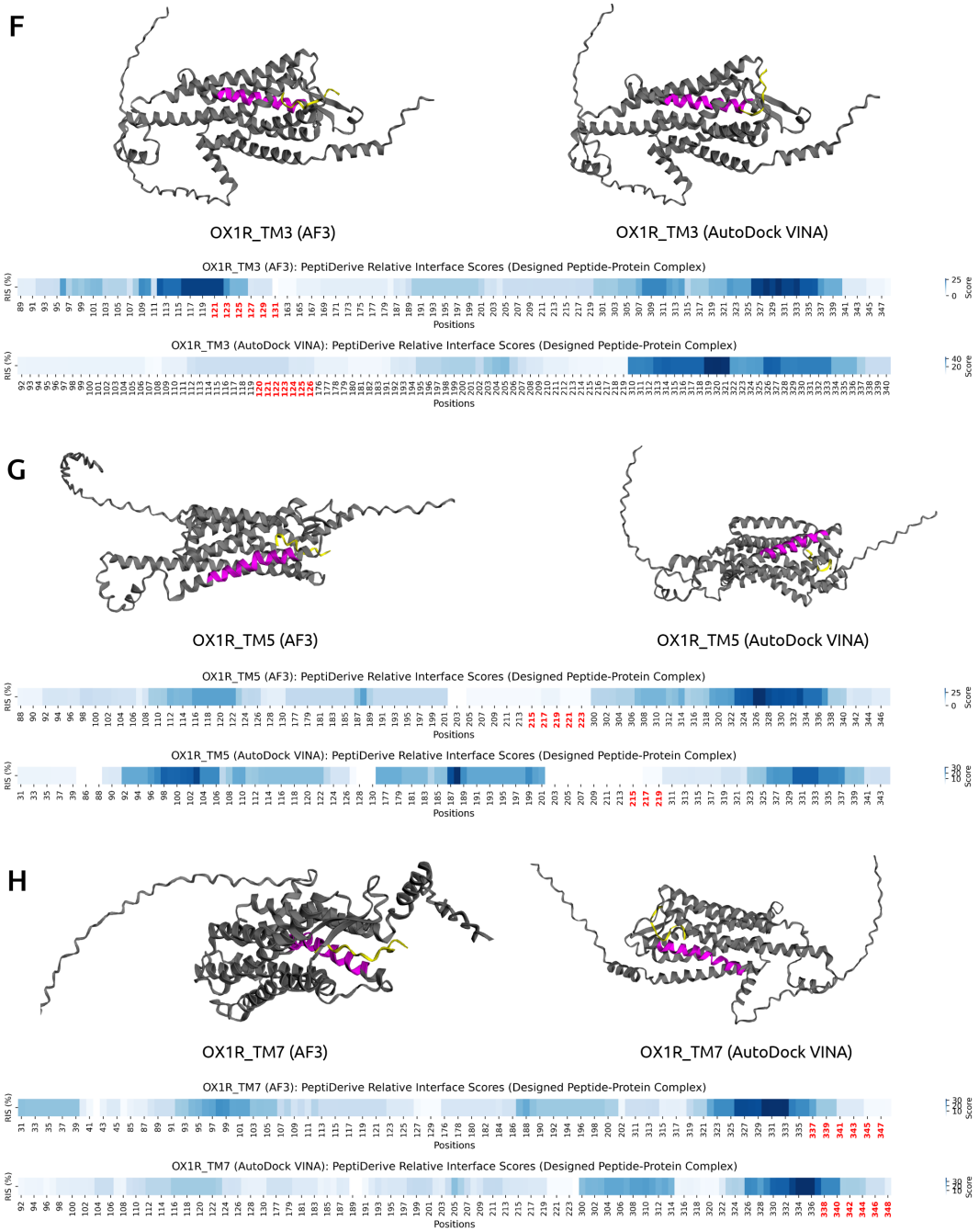


Figure 5: **Structural visualization and PeptiDerive relative interaction scores for designed peptides targeting structured motifs.** The peptide-complex structures are visualized for three different domains on OX1R: **(F)** Transmembrane (Name=3), **(G)** Transmembrane (Name=5), **(H)** Transmembrane (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 the moPPIt-v3 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-v3 as the desired target amino acids.

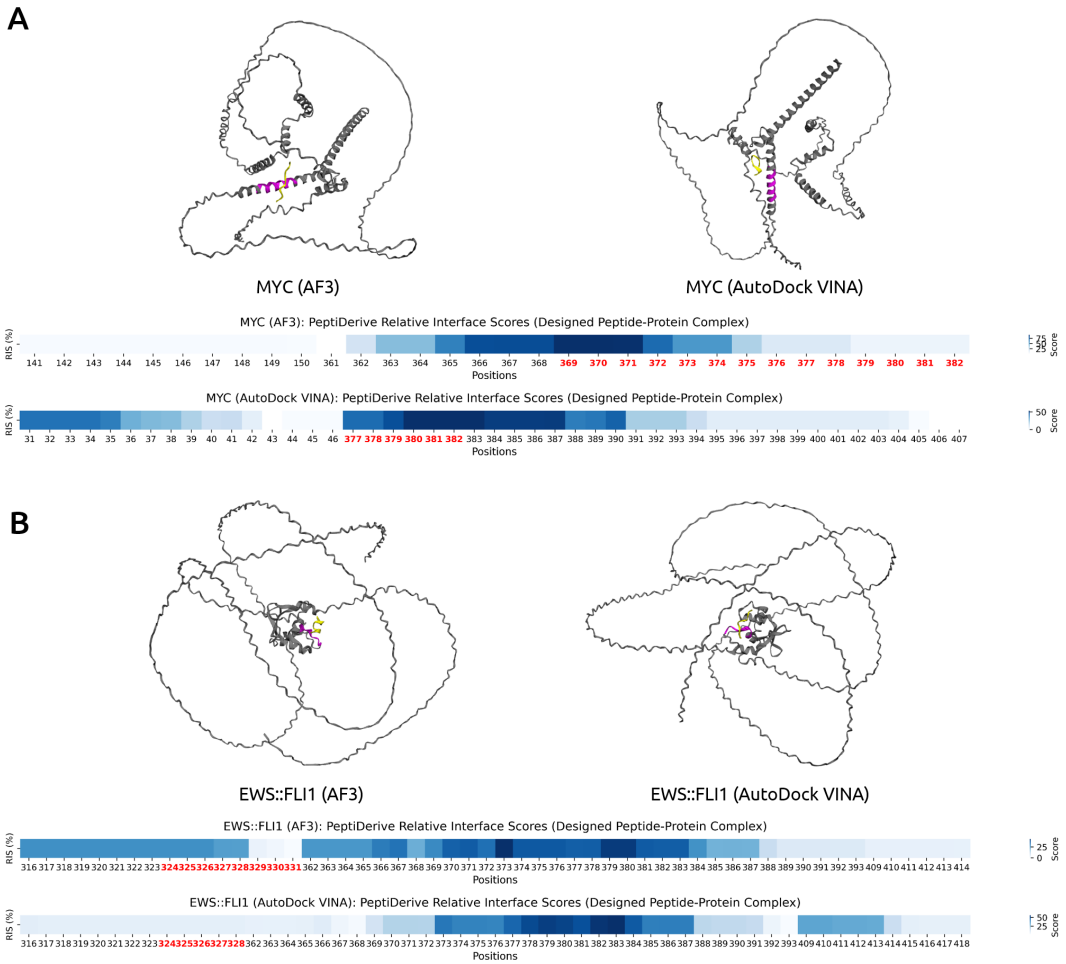


Figure 6: **Structural visualization and PeptiDerive relative interaction scores for designed peptides targeting intrinsically disordered proteins.** The peptide-complex structures are visualized for two intrinsically disordered proteins: **(A)** MYC, **(B)** EWS::FLI1 using AlphaFold3 and AutoDock VINA. The target proteins are depicted in grey, the designed peptide binders are shown in yellow, and the binding residues specified by the moPPIt-v3 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-v3 as the desired target amino acids.

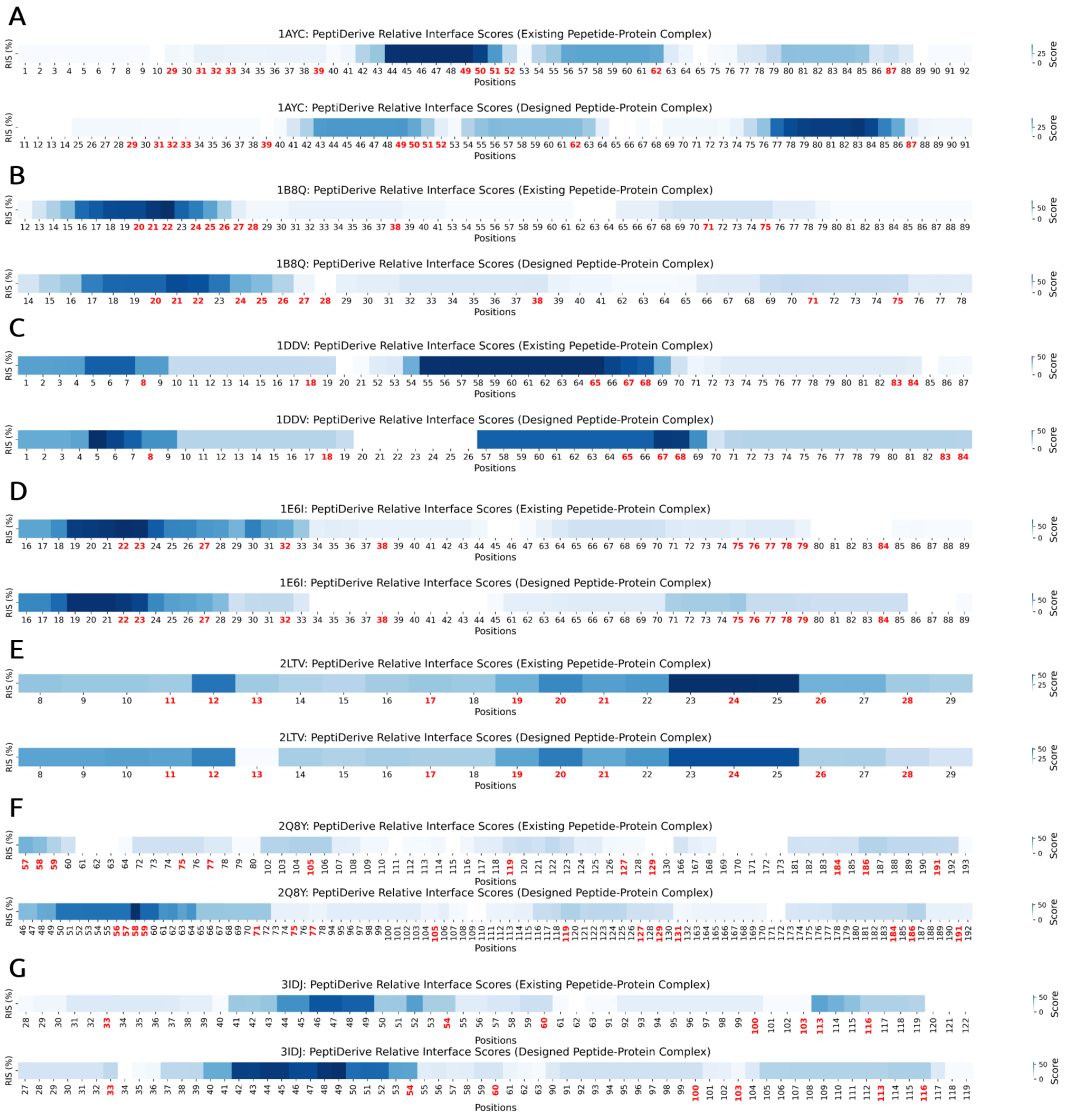


Figure 7: PeptiDerive relative interface scores for existing and designed peptide-protein complexes. Heatmaps of PeptiDerive relative interface scores (RIS) are shown for 7 peptide-protein complexes among 15 structured complexes with known binders that were tested: **(A)** 1AYC, **(B)** 1B8Q, **(C)** 1DDV, **(D)** 1E6I, **(E)** 2LTV, **(F)** 2Q8Y, **(G)** 3IDJ. The first heatmap for each protein shows the RIS of the existing peptide-protein complex, while the second heatmap shows the scores for the designed peptide-protein complex. For each heatmap, the x-axis indicates the residue positions, with highlighted positions in red representing the target binding amino acid positions that were input into moPPIt. High RIS at these positions indicate strong binding potential.

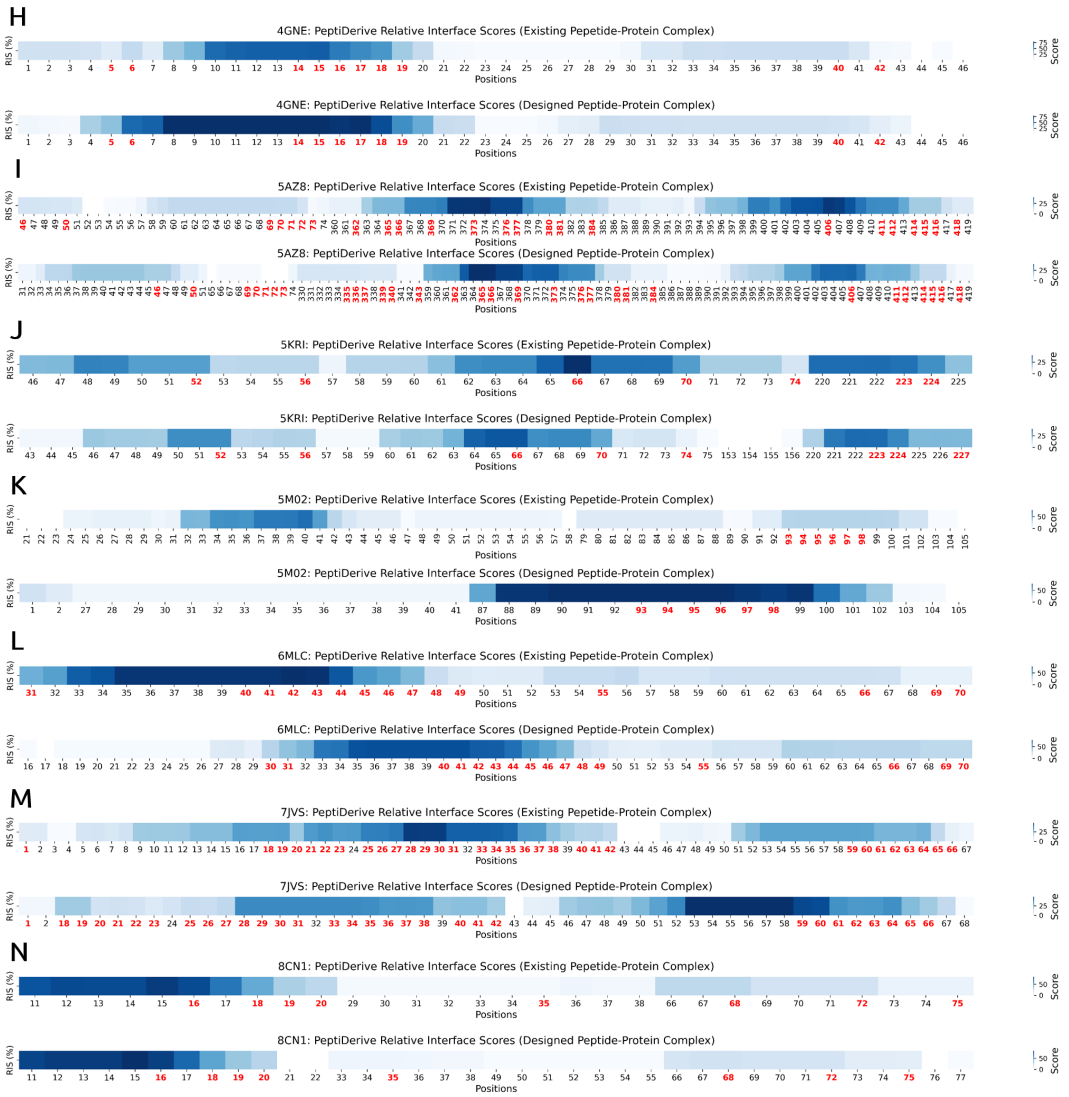
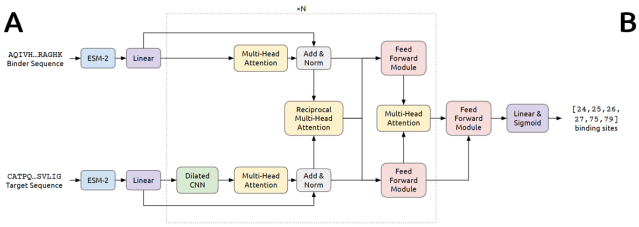


Figure 8: PeptiDerive relative interface scores for existing and designed peptide-protein complexes. Heatmaps of PeptiDerive relative interface scores (RIS) are shown for 7 peptide-protein complexes among 15 structured complexes with known binders that were tested: **(H)** 4GNE, **(I)** 5AZ8, **(J)** 5KRI, **(K)** 5M02, **(L)** 6MLC, **(M)** 7JVS, **(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.



B

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

Figure 9: **BindEvaluator. (A)** Overview of the architecture of BindEvaluator. BindEvaluator predicts the binding residues on the target protein given a target sequence and a binder sequence. The binder and target sequences are first processed using a pre-trained ESM-2-650M model to obtain their embeddings. The target sequence embeddings are further refined using a dilated CNN module to capture local features. Both embeddings are then passed through multi-head attention modules to capture global dependencies. Reciprocal multi-head attention modules integrate the representations of the target and binder sequences, allowing for the capture of binder-target interaction information. Feed-forward and linear layers subsequently process the refined embeddings to predict the binding sites. **(B)** Test performance metrics of BindEvaluator across different training configurations. Performance metrics were calculated for BindEvaluator across three configurations: trained without dilated CNN modules, trained with dilated CNN modules, and fine-tuned for peptide-protein binding site prediction. Metrics include overall loss, binary cross-entropy (BCE) loss, KL divergence loss, accuracy, area under the ROC curve (AUC), F1 score, and Matthews correlation coefficient (MCC).

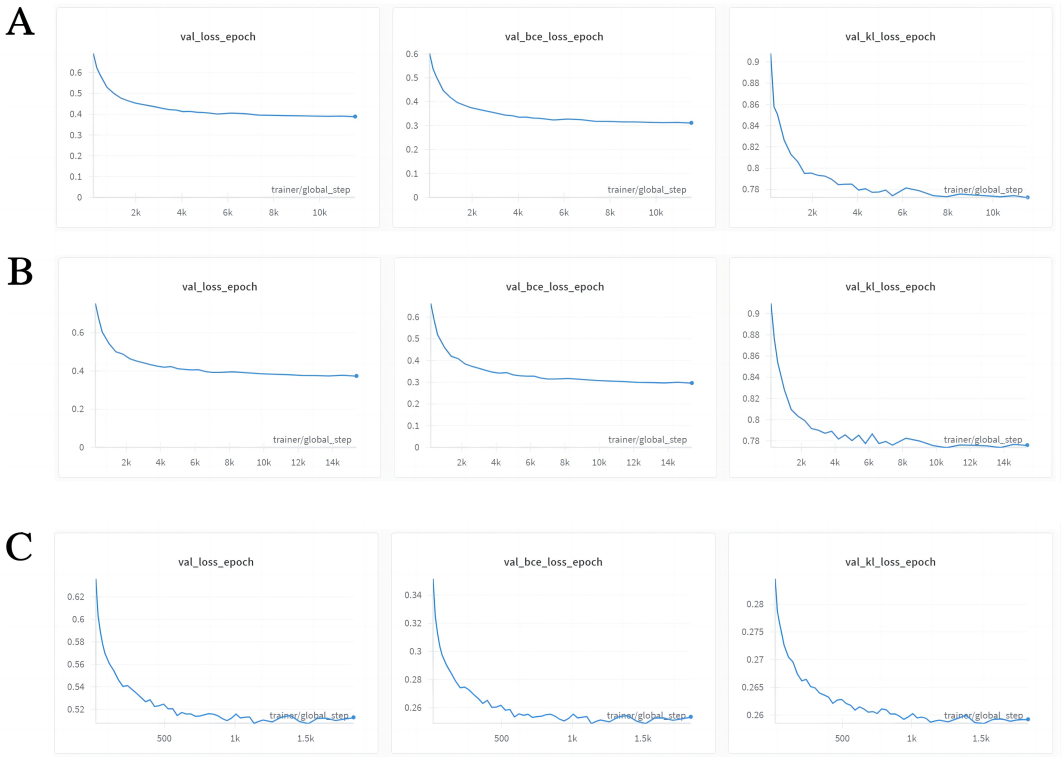


Figure 10: **Validation loss curves for BindEvaluator training and fine-tuning.** (A) Validation loss, binary cross-entropy (BCE) loss, and Kullback-Leibler (KL) divergence loss curves during training of BindEvaluator on the PPI dataset without dilated CNN modules. (B) Loss curves for training with dilated 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.