Low-N OpenFold fine-tuning improves peptide design without additional structures

Theodore Sternlieb*
Dyno Therapeutics
Watertown, MA 02472
theodore.sternlieb@dynotx.com

Jakub Otwinowski*
Dyno Therapeutics
Watertown, MA 02472
jakub.otwinowski@dynotx.com

Sam Sinai
Dyno Therapeutics
Watertown, MA 02472
sam.sinai@dynotx.com

Jeffrey Chan
Dyno Therapeutics
Watertown, MA 02472
jeffrey.chan@dynotx.com

Abstract

Discovery of high-affinity peptide binders is broadly useful for designing novel therapeutics. Machine learning-based *in silico* screening is a promising approach for increasing the success rate of therapeutic peptide design. Structure-based prediction models, such as AlphaFold-multimer, have shown promising though insufficient zero-shot performance for *in silico* screens of diverse peptides. Incorporating interaction data produced during peptide discovery campaigns, we develop a low-N OpenFold fine-tuning procedure on the peptide recognition modules database (PRM-db). With a relatively small dataset, we find 13-60x fold increase in design hit rate with the fine-tuned model making a powerful model for improving peptide design success rates. Unexpectedly, we also find that interaction data also improves structure complex predictions critical for targeting binding sites during design campaigns. The framework introduced here demonstrates a data-efficient recipe for dramatically improving peptide-protein prediction and ultimately the success rate of peptide binder design, without the need for additional experimentally determined peptide-protein complexes.

1 Introduction

Peptides are a critical class of biological molecules with diverse roles in cellular processes, making them valuable therapeutic candidates (such as GLP-1 and insulin) for the treatment of a wide range of diseases [11]. Therapeutic peptides function by binding to a specific target binding site to induce the desired therapeutic effect. Furthermore, larger protein-protein interactions are often mediated by peptide-like short linear motifs (SLiMs), such that information gleaned from peptides may serve as building blocks which generalize to broader classes of interaction.

Therapeutic peptide candidates often require multiple desirable properties stability, solubility, and a specific binding site in addition to high target binding affinity. High-throughput experimental screens of random peptide libraries can generate some peptide binders against a target, however subsequent libraries are often needed to achieve a large enough pool for further filtering against these other biological properties. *In silico* binding predictors are a promising approach to design new libraries with a high rate of binding peptides. In addition, accurate structural complex prediction would enable the direct design of binders that bind to a specific target binding site often necessary for activating or blocking a mechanistic process of therapeutic interest.

Machine Learning for Structural Biology Workshop, NeurIPS 2024.

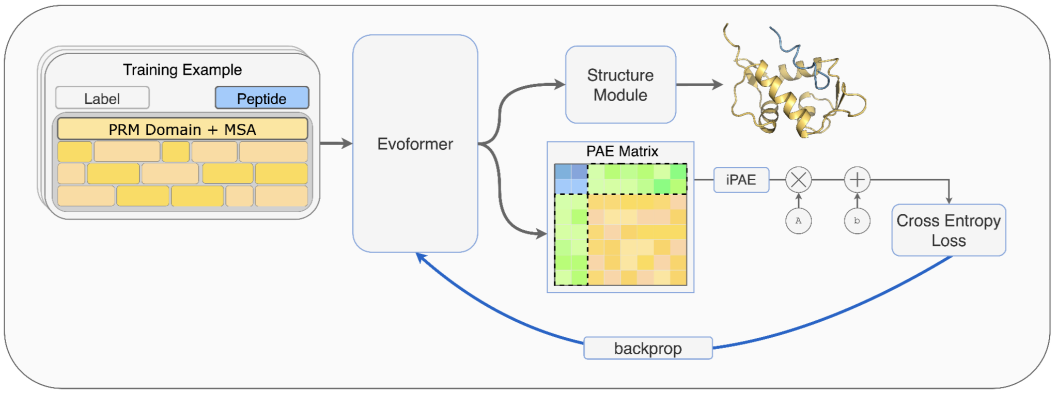

Figure 1: OpenFold binding predictor architecture. Peptides and PRM domains along with MSAs and are passed into the Evoformer layer. The mean of the PAE over the interaction region is calculated, adjusted by a scale and bias parameter, and fed in a cross entropy loss.

Critical to any such *in silico* screening approach is accurate prediction. AlphaFold-based methods have demonstrated success in the peptide binding prediction task [10, 2, 8]. However, predicting peptide binding is difficult, and there remains significant room for improvement. Peptides lack co-evolutionary information that typically encodes interaction information. Peptides may also be more flexible due their short length and have more structural disorder, limiting the availability of experimental structures for training models.

To address the need for improvement in binding prediction, developing an AlphaFold-based model that incorporates high-throughput experimental binding databases could dramatically increase the number of peptide binder candidates identified. However, high-throughput binding data is limited to only a moderate number (10s-100s) of positive binding examples per domain raising questions around whether typical dataset sizes can lead to meaningful improvements in modeling performance.

Recent work investigated peptide-MHC binding prediction by adding a prediction head to AlphaFold and fine-tuning on around 5,000 binding pairs [7]. However, this approach relied heavily on structural templates of the peptide-MHC complex, and it is unknown if this approach generalizes to other proteins with less structurally-characterized binding pockets and fewer binding examples per domain. In addition, this work suggested that fine-tuning mainly helped by lowering the rate of false positives. However, for designing peptide binders, reducing the number of false negatives is critical for *in silico* screens. It remains unclear whether binding data without structural information is sufficient for identifying the correct structural pose for misclassified peptides necessary for reducing false negatives.

To investigate the viability of utilizing experimental binding data to improve *in silico* screening for therapeutic peptide design, we demonstrate via fine-tuning that with relatively few positive examples per domain, we achieve a 13- to 60-fold increase in hit rate. Surprisingly, we find that despite not training on structure, structure prediction was often improved which enables usage of predicted binding site as a selection criteria. We believe this framework is promising as we incorporate increasingly larger interaction datasets.

2 Results

We added a binding prediction head to OpenFold, an alternative implementation of AlphaFold, and fine-tuned the parameters on a dataset of diverse peptides-domain interactions [9] without using templates. This data consisted of 163 small domains called peptide recognition modules (PRMs), and novel peptides discovered by high-throughput phage display. We curated the PRM data and augmented it with synthetic negatives (non-binding) data by shuffling peptide-domain pairs (see Methods and schematic in Fig. 1). We added a classification head to OpenFold's predicted aligned error confidence metric at the peptide-domain interface (iPAE), and fine-tuned the Evoformer and binding head parameters (see Methods). Notably, the Structure Module remains fixed during fine-tuning, since the structure loss is not optimized.

We evaluate how our model predicts binding of novel peptides to domains on two tasks:

- **Shared Domain Task:** Against domains present in the training set. Primarily useful for increasing the number of high-affinity binders via incorporation of ML designs in subsequent rounds of screening.
- **Novel Domain Task:** Against domains not in the training set. Primarily useful for replacing a large random libraries with a designed library or for design against targets difficult to perform high-throughput assays on.

We asses the performance of fine-tuned Openfold using positive likelihood ratio (PLR), a standard metric for diagnostic screens, because it do not depend on the prevalence of positives in test sets (see Appendix for further explanation). If binders are very rare among random peptides, PLR approximates the enrichment of binders over a random selection. On the shared domain task we see a large improvement of PLR after finetuning, jumping from 2.9 to 13-fold enrichment, in the top 20% of candidates ranked by prediction score and reaching as high as 20- to 60-fold enrichment for more stringent selection (Fig. 2A). Fine-tuning also improves the negative likelihood ratio, used when negatives are selected for follow-up testing, and other standard classifier statistics (Fig. S1). In particular, we see many false negatives turn into true positives after fine-tuning, a desirable property for peptide design (Fig. S4). We also find that two sequence-only baseline models are not competitive with fine-tuned openfold, suggesting structural information is valuable in this setting (Fig. 2A).

The Novel Domain Task is a more difficult test for generalization. We find that structure-based models remain better than sequence-only methods, though by a smaller margin. While fine-tuning is expected to degrade OpenFold's performance in out-of-distribution regimes, we find that overall the performance is largely the same (Fig. 2).

In peptide design, a particular binding site may be targeted, so it is important to know if structure prediction changes during fine-tuning. Despite not including a structure loss in our model, it is remarkable that domain structure predictions changed so little, as we see small root mean square distances (RMSD) between predictions before and after fine-tuning (Fig. S2). Peptides have larger RMSD, as they are more flexible and can change binding position. We see that prediction of binders that are 'rescued', that is initially predicted to not bind, and then are predicted to bind after fine-tuning, have their physical interfaces restructured during fine-tuning, as quantified by low similarity scores (DockQ [6]) between base and fine-tuned structure predictions (Fig. S3).

We further characterized the impact of fine-tuning on structure predictions by examining changes in predicted binding poses for a set of experimentally resolved peptide-domain complexes with similar domains and dissimilar peptides to those in the training data (see Methods). For a set of 12 peptide-domain complexes, we predict a number of structures with both fine-tuned and base OpenFold, take the top 10% by iPAE ranking, and calculate DockQ scores against the native structure. Despite not providing any additional structural information, we see similar or improved DockQ for most structures (Fig. 3).

3 Discussion

Fine-tuning a structure prediction model on binding data improves the ability to positively or negatively screen candidates. For protein engineering, we can estimate how an *in silico* screen can be used to design a peptide library. In the PRM experiment, roughly 10^3 binding peptides were discovered from a pool of approximately $4 \cdot 10^{10}$, resulting in a hit rate of approximately $3 \cdot 10^{-8}$. While a screen with a 4x improved hit rate for the Novel Domain Task is insufficient to replace the initial random peptide library, the 60x improvement rate from the Shared Domain Task can augment additional rounds of screening increasing the number of high-affinity binders to a particular binding site. Additionally, our fine-tuned model can be used as a screen in pre-existing peptide design workflows such as exploration around known hits or screening in conjunction with a generative model, such as RFDiffusion [12] to propose sequences. Aggregating larger diverse datasets [5, 3] to further fine-tune on may lead to even greater performance gains on both the Novel Domain and Shared Domain Task.

Notably, the PLR trends upwards as the selection is stricter, a useful property for lower-throughput studies, and may be even higher than we can detect at high thresholds due to low numbers. Data leakage, a common concern in protein machine learning, is less of a concern in our Shared Domain

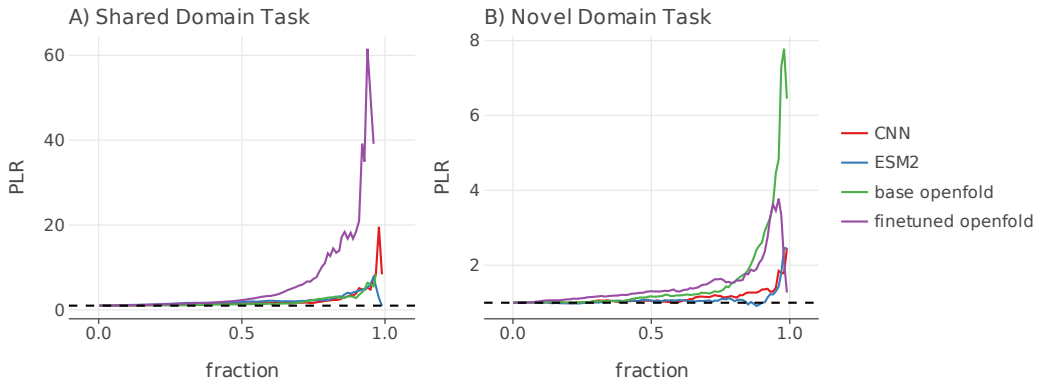

Figure 2: Fine-tuned OpenFold is better at screening peptide-domain binding than base OpenFold, CNN, and ESM2 baselines for the A) Shared Domain Task, and B) Novel Domain Task. A higher positive likelihood ratio (PLR) indicates a higher chance that selected peptides will be true binders. PLR depends on the prediction score threshold used to select peptides, shown here as the fraction of scores below a value (the empirical cumulative distribution). Dashed line indicates PLR = 1.

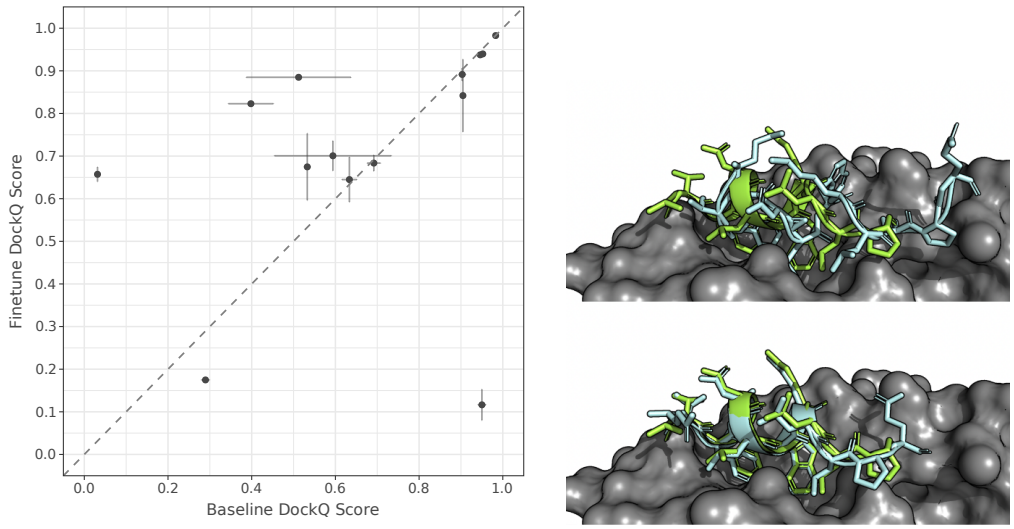

Figure 3: Of 12 peptide-domain complexes, most structure predictions are maintained or improved. Left: Mean DockQ scores for the top 10% of predictions per PDB using the baseline and fine-tuned OpenFold models with one recycle. Error bars capture standard deviation over 8 predictions. Right: Both images show OpenFold peptide predictions (cyan) overlaid on the native peptide (green). The top image shows the baseline model, while the bottom shows the fine-tuned model. After fine-tuning, OpenFold accurately predicts relevant inter-chain aromatic binding.

dataset. The peptides were generated in an stochastic experimental process, so a random selection of peptides is aligned with a realistic screening scenario, even if binders may share some sequence motifs [9]. For our Novel Domain dataset which contains different domains, the performance may depend on domain similarity.

Fine-tuning structure models may be most effective in regimes where the number of binders per domain is low to moderate. When this quantity is large, sequence-only models may catch up in performance, but this setting is difficult to achieve experimentally and obviates the need of machine learning-designed libraries. Furthermore, since peptides are merely small proteins, we believe jointly training on protein-protein and protein-peptide binding data can further improve performance.

We've also observed signs that fine-tuning on binding data can further improve structure prediction. We believe that this is a promising path forwards for the structure prediction field, since scaling up

data collection of structural binding complexes and co-evolutionary information is challenging for peptide-protein complexes while binding data is comparatively cheap to acquire.

Methods

Data preparation We downloaded dataset EV3 from [9], and extracted peptide and domain sequences. We predicted structures for domain sequences with base OpenFold, and excluded domains with an average pLDDT < 90. We also excluded domains with ≥ 400 amino acids to avoid out-of-memory errors, resulting in 115 unique domain sequences. We defined a test set (the Novel Domain Task) as domains with < 10 binding peptides. We randomly split the remaining data 80/20 into train and validation sets, grouping by domain such that peptides per domain are sampled with a 80/20 split. We use the validation set as the Shared Domain Task. We generated synthetic negative data by randomly assigning peptides to different domains per split, removing any pairs that appeared in the positive set by chance. The ratio of positive to negative data was approximately 1:3. See table S1 in appendix for data statistics. Since in the experiment the 16mer peptides had an unspecified number of glycines attached to both ends, we added 3 G's on both ends resulting in all peptides having length 22. In addition, we found that base OpenFold had better performance when trained with glycine flanked sequences compared to just the 16mer.

To examine changes in structure prediction quality induced by fine-tuning, we used dataset EV5 from [9], which contains experimentally determined peptide-domain structures with high sequence-similarity to domains from EV3. We first identifed each unique domain in our training data with the peptide-domain structure from EV5 sharing the greatest domain level sequence identity. We then select from those peptide-domain structures those with > 85% sequence similarity to a domain from the training data. This process produced 12 relevant experimental structures. Additionally, peptides from the training dataset had >10 Hamming distance from peptides in the identified experimental structures.

Fine-tuning OpenFold We added a binding prediction head to OpenFold, that links the predicted aligned error (PAE), averaged over the peptide-domain interface (iPAE), to a binary cross entropy loss. This head has two free (scale and bias) parameters, that were initialized by calculating iPAE from base OpenFold, and linearly regressing on the training set labels. We only used OpenFold model *model_1_multimer_v3*. We fine-tuned OpenFold, freezing all parameters except the Evoformer module and the binding head, with 0 recycles, batch size 64 by gradient accumulation, learning rate $3 \cdot 10^{-5}$, and gradients clipped at 0.1. We trained on 4x A100 GPUs with bf16 precision for 50 epochs with early stopping based on validation loss and patience setting of 10, and using the model with minimum validation loss at epoch 25. Inference for the Novel Domain Task and Shared Domain Task was run with 0 recycles.

Baseline models We used base OpenFold for evaluation with model *model_1_multimer_v3* and 2 recycles. We used a convolutional neural network classifier (CNN) with 2 convolutional layers with filter width 5, followed by a linear layer with output size 64 to generate embeddings. The classification logits are then generated by taking the cosine similarity between peptide and domain embeddings and the model is trained with a binary cross-entropy loss. We also experimented with parameterizing the embeddings using ESM2 [4], similar to PepPRClip [1], a model which has shown good performance on similar tasks. Our model generates protein and peptide embeddings using two copies of ESM2 with mean pooling. The embeddings are then fed into two separate dense heads, each with two layers, before generating logits using cosine similarity. The model is trained with a binary cross-entropy loss. We experimented with using the 8m, 35m and 150m parameter checkpoints for the protein embeddings and found best performance using the 8m model. We used the 8m checkpoint for peptide embeddings in all experiments.

References

[1] Suhaas Bhat et al. *De Novo Design of Peptide Binders to Conformationally Diverse Targets with Contrastive Language Modeling*. Pages: 2023.06.26.546591 Section: New Results. July 22, 2024. DOI: 10.1101/2023.06.26.546591. URL: https://www.biorxiv.org/content/10.1101/2023.06.26.546591v2 (visited on 09/16/2024).

[2] Liwei Chang and Alberto Perez. "Ranking Peptide Binders by Affinity with AlphaFold". In: *Angewandte Chemie International Edition* 62.7 (2023). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.202213362, e202213362. ISSN: 1521-3773. DOI: 10.1002/anie.202213362. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.202213362 (visited on 09/06/2024).

[3] Bifang He et al. "Biopanning data bank 2018: hugging next generation phage display". In: *Database* 2018 (Jan. 1, 2018), bay032. ISSN: 1758-0463. DOI: 10.1093/database/bay032. URL: https://doi.org/10.1093/database/bay032 (visited on 09/20/2024).

[4] Zeming Lin et al. "Evolutionary-scale prediction of atomic-level protein structure with a language model". In: *Science* 379.6637 (Mar. 17, 2023). Publisher: American Association for the Advancement of Science, pp. 1123–1130. DOI: 10.1126/science.ade2574. URL: https://www.science.org/doi/abs/10.1126/science.ade2574 (visited on 09/16/2024).

[5] Filip Mihalič et al. "Large-scale phage-based screening reveals extensive pan-viral mimicry of host short linear motifs". In: *Nature Communications* 14.1 (Apr. 26, 2023). Publisher: Nature Publishing Group, p. 2409. ISSN: 2041-1723. DOI: 10.1038/s41467-023-38015-5. URL: https://www.nature.com/articles/s41467-023-38015-5 (visited on 08/30/2024).

[6] Claudio Mirabello and Björn Wallner. *DockQ v2: Improved automatic quality measure for protein multimers, nucleic acids, and small molecules*. Pages: 2024.05.28.596225 Section: New Results. June 2, 2024. DOI: 10.1101/2024.05.28.596225. URL: https://www.biorxiv.org/content/10.1101/2024.05.28.596225v1 (visited on 09/04/2024).

[7] Amir Motmaen et al. "Peptide-binding specificity prediction using fine-tuned protein structure prediction networks". In: *Proceedings of the National Academy of Sciences* 120.9 (Feb. 28, 2023). Publisher: Proceedings of the National Academy of Sciences, e2216697120. DOI: 10.1073/pnas.2216697120. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2216697120 (visited on 09/06/2024).

[8] Felix Teufel et al. "Deorphanizing Peptides Using Structure Prediction". In: *Journal of Chemical Information and Modeling* 63.9 (May 8, 2023). Publisher: American Chemical Society, pp. 2651–2655. ISSN: 1549-9596. DOI: 10.1021/acs.jcim.3c00378. URL: https://doi.org/10.1021/acs.jcim.3c00378 (visited on 09/06/2024).

[9] Joan Teyra et al. "Large-scale survey and database of high affinity ligands for peptide recognition modules". In: *Molecular Systems Biology* 16.12 (Dec. 2020). Publisher: John Wiley & Sons, Ltd, e9310. ISSN: 1744-4292. DOI: 10.15252/msb.20199310. URL: https://www.embopress.org/doi/full/10.15252/msb.20199310 (visited on 08/29/2024).

[10] Tomer Tsaban et al. "Harnessing protein folding neural networks for peptide–protein docking". In: *Nature Communications* 13.1 (Jan. 10, 2022). Publisher: Nature Publishing Group, p. 176. ISSN: 2041-1723. DOI: 10.1038/s41467-021-27838-9. URL: https://www.nature.com/articles/s41467-021-27838-9 (visited on 09/06/2024).

[11] Lei Wang et al. "Therapeutic peptides: current applications and future directions". In: *Signal Transduction and Targeted Therapy* 7.1 (Feb. 14, 2022). Publisher: Nature Publishing Group, pp. 1–27. ISSN: 2059-3635. DOI: 10.1038/s41392-022-00904-4. URL: https://www.nature.com/articles/s41392-022-00904-4 (visited on 09/06/2024).

[12] Joseph L. Watson et al. "De novo design of protein structure and function with RFdiffusion". In: *Nature* 620.7976 (Aug. 2023). Publisher: Nature Publishing Group, pp. 1089–1100. ISSN: 1476-4687. DOI: 10.1038/s41586-023-06415-8. URL: https://www.nature.com/articles/s41586-023-06415-8 (visited on 09/06/2024).

A Appendix / supplemental material

A.1 A primer on diagnostic testing

Assume a fraction of a population is positive $p(x_+)$ and the rest negative $p(x_-) = 1 - p(x_+)$. We can interpret these fractions as Bayesian priors. Our test yields a number of proposed positives $\hat{x}_+$, and the Bayesian posterior is $p(x_+|\hat{x}_+)$, which is known as the precision or positive predictive value, and can be calculated directly from the confusion matrix. With Bayes rule we find

$$p(x_+|\hat{x}_+) = \frac{p(\hat{x}_+|x_+)p(x_+)}{p(\hat{x}_+|x_+)p(x_+) + p(\hat{x}_+|x_-)p(x_-)}.$$

The precision depends explicitly on the prevalence $p(x_+)$. Therefore, precision is not a useful metric to quantify a diagnostic screen when the population on which we are evaluating our test has a prevalence that is very different from the natural population that would be used in a practical application.

The posterior odds compares the two hypotheses, but is also dependent on the prevalence:

$$\frac{p(x_+|\hat{x}_+)}{p(x_-|\hat{x}_+)} = \frac{p(\hat{x}_+|x_+)p(x_+)}{p(\hat{x}_+|x_-)p(x_-)}.$$

However, the positive likelihood ratio (PLR) $p(\hat{x}_+|x_+)/p(\hat{x}_+|x_-)$, is independent of prevalence, and can be directly calculated from the confusion matrix. PLR and the negative likelihood ratio, $p(\hat{x}_-|x_+)/p(\hat{x}_-|x_-)$, are commonly used to quantify diagnostic tests. If in the natural population the prevalence is very rare, $p(x_+) << 1$, the PLR predicts fold change in the number of expected hits compared to a random sample.

Split	Binder	Peptide-Domain Pairs	Unique Domains	Unique Peptides
Train	Negative	1153	30	403
	Positive	403	30	403
Validation (Shared Domain)	Negative	288	30	99
	Positive	99	30	99
Test (Novel Domain)	Negative	969	85	318
	Positive	330	85	318

Table 1: Data split, label distribution, and number of unique sequences.

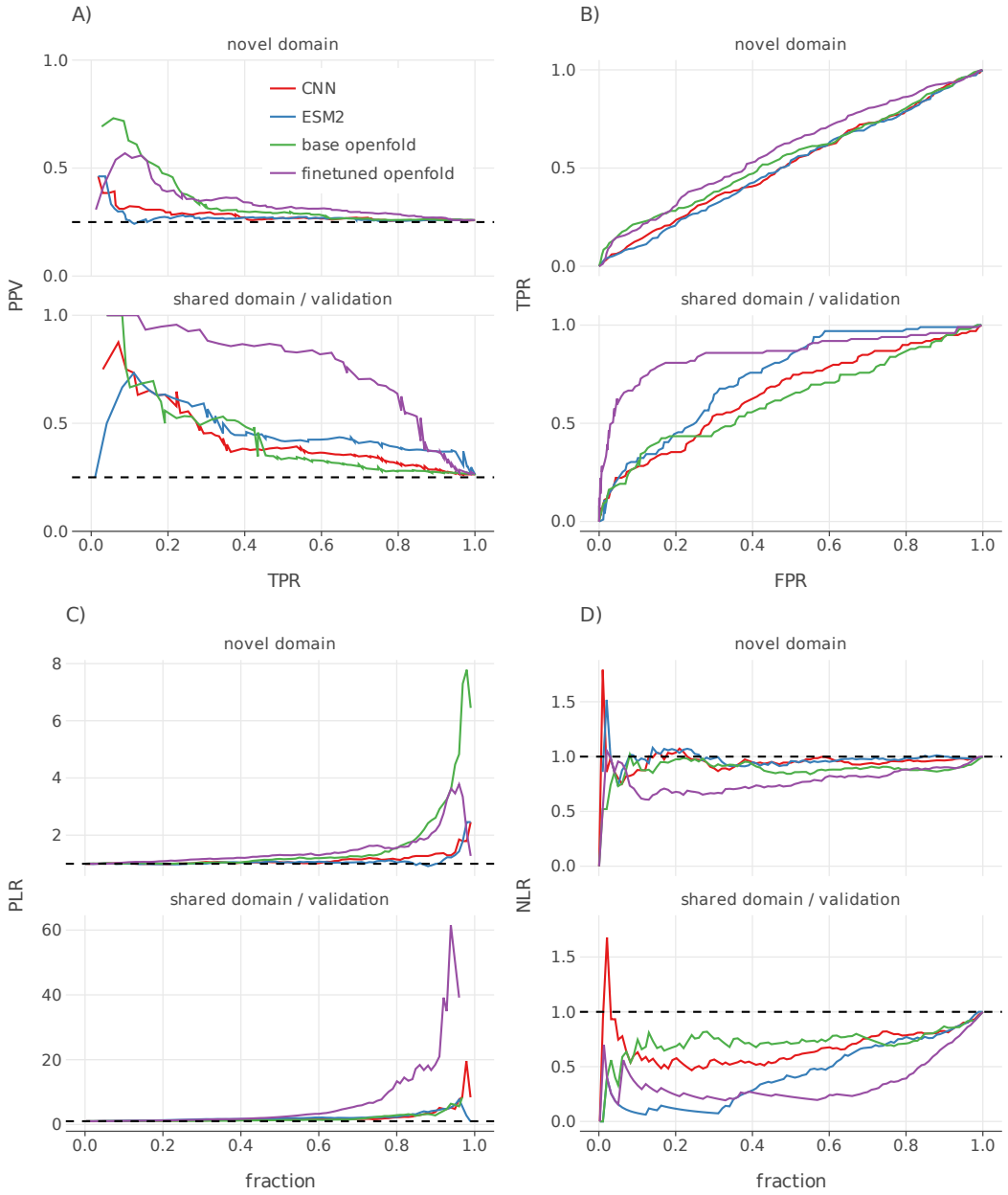

Figure S1: Fine-tuned OpenFold is better at classifying peptide-domain binding than base OpenFold, CNN, and ESM2 baselines. (A) positive predictive value (PPV) vs true positive rate (TPR). (B) false positive rate (FPR). (C) Positive likelihood ratio (PLR) and (D) negative negative likelihood ratio (NLR) vs fraction of binding scores below a threshold value.

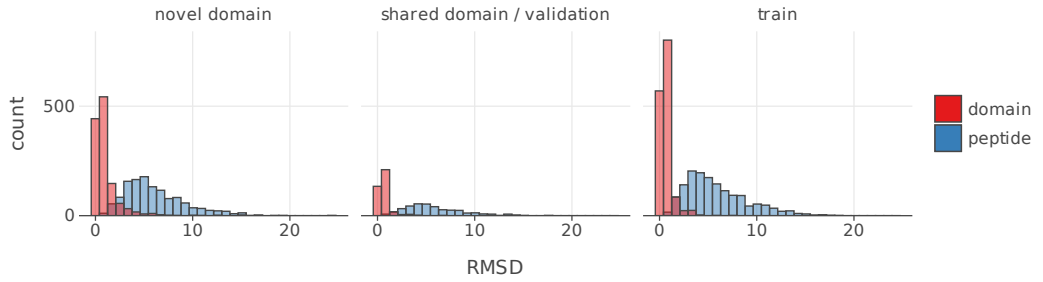

Figure S2: Structure prediction of domains are nearly unchanged by fine-tuning. The root mean square distances (RMSD) between predictions from the fine-tuned and base models is low for domains (red), indicating that the fine-tuning has not deteriorated the structure prediction quality. Peptide RMSD (blue) is much larger, as fine-tuning restructures the peptides and their binding interface. Three panels contain the data splits as labeled.

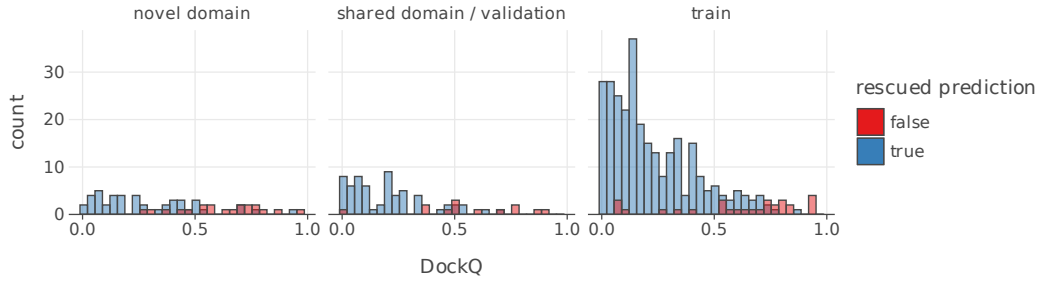

Figure S3: Binding predictions rescued by fine-tuning have remodeled interfaces. Rescued predictions (blue), i.e. peptide-domain binders (not synthetic negatives), that were predicted to not bind in base OpenFold but predicted to bind after fine-tuning, have a different physical interface after fine-tuning, with mostly low DockQ scores between fine-tuned model and base model structure predictions. Non-rescued binder predictions (red), which are actual binders that are predicted to bind before and after fine-tuning have relatively unchanged interfaces (DockQ > 0.23). Three panels contain the data splits as labeled.

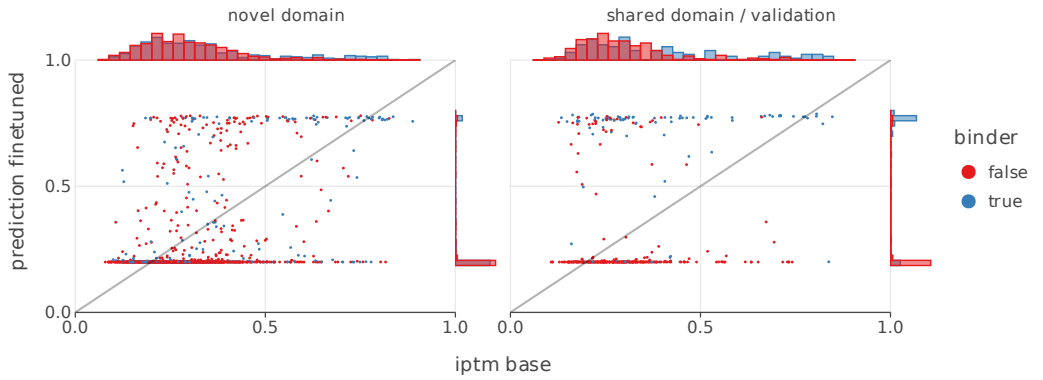

Figure S4: Fine-tuned classifier predictions vs base OpenFold iPTM values. Many base model false negatives (binding peptides (blue) with low iptm) have high scores after fine-tuning.

Top: Baseline, Bottom: Fine-tuned

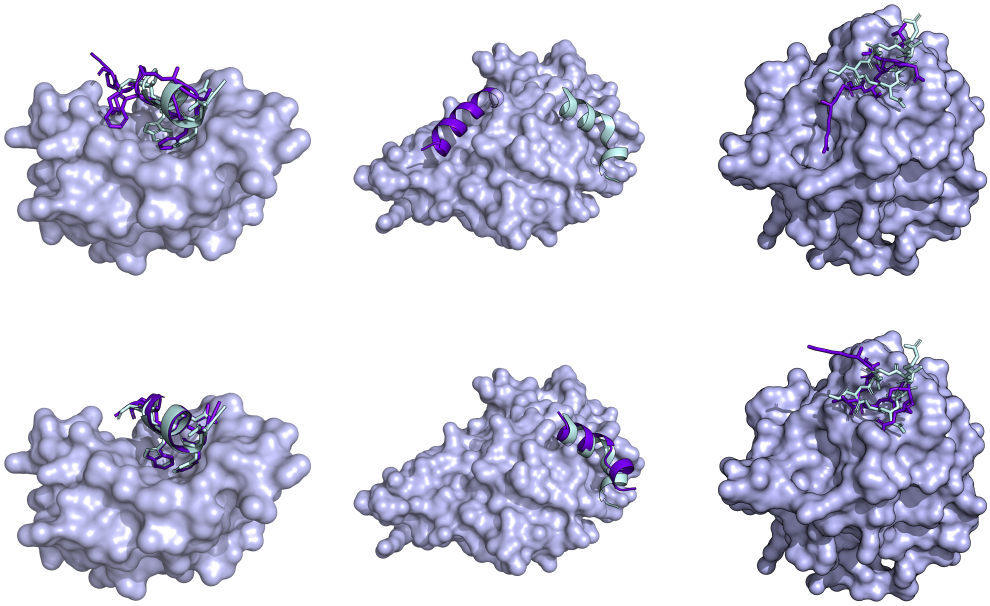

Figure S5: Sample structures with improved DockQ scores from fine-tuning. Top images are base OpenFold predictions and bottom images are fine-tuned OpenFold predictions. All images show the true peptide structure (cyan) and the predicted structure (purple). PDB IDs from left to right: 3dab, 3ipq, 5xn3. For 3dab and 5xn3 fine-tuning appears to improve the peptide orientation where as in 3ipq, fine-tuned OpenFold predicts an entirely different binding pocket.

Top: Baseline, Bottom: Fine-tuned

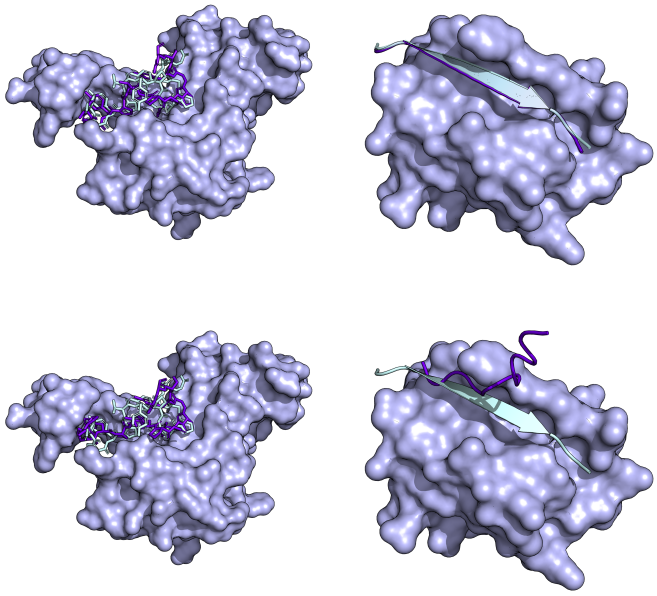

Figure S6: Sample structures with diminished DockQ scores from fine-tuning. Top images are base OpenFold predictions and bottom images are fine-tuned OpenFold predictions. All images show the true peptide structure (cyan) and the predicted structure (purple). PDB IDs from left to right: 2zne, 3zke.