
Supplementary Materials for Linker-Tuning: Optimizing Continuous Prompts for Heterodimeric Protein Prediction

Anonymous Author(s)

Affiliation

Address

email

1 A Preliminary analysis of different backbones on the inter-chain contact 2 prediction task

3 We perform a preliminary experiment on the heterodimer test set on the inter-chain contact prediction
4 task to investigate how different backbones affect performance. We compare three backbones,
5 including ESM2-650M¹, ESM2-3B², and ESMFold-v1 which contains 3B+690M parameters.

6 For ESM2-650M and ESM2-3B, we train a distogram head on top of each model using single chains
7 and their distograms from the heterodimer training set. We leverage the same architecture for the
8 distogram head as in LM-design [1]. The distogram head contains only a linear layer with a softmax
9 activation function. Given a protein sequence, it takes the stacked attention map from ESM2 as input
10 and outputs a probability distribution on 18 distance bins for each residue pair. Specifically, the
11 stacked attention map for each protein is a matrix of shape (num_layers*num_heads, N , N), where
12 num_layers denotes the number of layers in ESM2, num_heads denotes the number of attention heads
13 in ESM2, and N denotes the number of residues in the protein. For ESM2-650M, num_layers is 33,
14 and num_heads is 20. For ESM2-3B, num_layers is 36, and num_heads is 40. We train the distogram
15 head with ESM2 frozen for 10 epochs using the distogram loss and select the best checkpoint based
16 on the validation distogram loss. For ESMFold, we use its pre-trained distogram head.

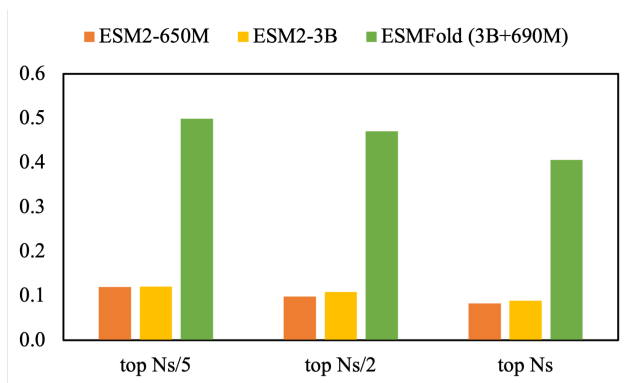


Figure 1: Inter-chain contact prediction results on the heterodimer test set using different backbones, including ESM2-650M, ESM2-3B, and ESMFold (3B+690M). Chains in the heterodimer are joined by the G25 linker before input into the backbone model.

¹https://dl.fbaipublicfiles.com/fair-esm/models/esm2_t33_650M_UR50D.pt

²https://dl.fbaipublicfiles.com/fair-esm/models/esm2_t36_3B_UR50D.pt

Figure 1 shows the inter-chain contact prediction precisions on the heterodimer test set using different backbones. ESM2-3B performs only slightly better than the ESM2-650M while containing >2B more parameters. In contrast, ESMFold significantly improves performance over ESM2-3B while containing only 690M more parameters. This result indicates that the Folding Module plays a critical role in extracting the structure information out of the sequence. Therefore, we suspect that ESM2 is not so sensitive to structure prediction and is harder to control. Although it is natural to place the prompt at ESM2 for NLP researchers, however, for protein structure prediction, Folding Module is a better choice for prompt tuning.

B Structure prediction results on a larger antibody VH-VL test set

In order to extend the applicability of our method beyond antibody heavy chain light chain Fv region docking, we conduct additional testing on an expanded test set consisting of complete heavy chains and complete light chains from various antibodies. The dataset contains 171 samples, so we call it VH-VL171 for short. VH-VL171 is extracted from antibodies released in PDB between 2022-01-16 and 2022-04-13. It filters out samples that contain more than 10 consecutive missing residues in a chain. It also filters out chains whose length falls outside the range of [30, 1,024]. For each sample in VH-VL171, the heavy chain and the light chain extracted are interacted pairs. The average sequence length is 337 and the maximum is 454.

Table 1 shows the structure prediction results of our methods and baseline methods on VH-VL171. Across all 171 test samples, ESMFold-Linker achieves an average DockQ score of 0.55, outperforming AlphaFold-Linker and HDock by large margins. In comparison, our method ESMFold-Linker* achieves an average DockQ score of 0.65, outperforming the ESMFold-Linker baseline by +18.18%. Meanwhile, it achieves an average TM-score of 0.90, surpassing ESMFold-Linker by +5.88%. It also reduces almost half of the RMSD compared to ESMFold-Linker. Furthermore, it successfully predicts 164 out of 171 interfaces, the same as AF-Multimer(v3 best). These results double validate that our method indeed generalizes well to antibody heavy chain light chain docking.

Table 1: Structure prediction results on the antibody **VH-VL171** test set.

	Structure metrics			Interface quality count			
	DockQ \uparrow	RMSD \downarrow	TM-score \uparrow	Incorrect	Acceptable	Medium	High
AlphaFold-Linker	0.35 ± 0.35	12.97 ± 11.61	0.71 ± 0.23	87	4	69	11
HDock	0.49 ± 0.30	6.91 ± 8.40	0.82 ± 0.17	38	35	72	26
ESMFold-Linker	0.55 ± 0.25	5.03 ± 5.84	0.85 ± 0.14	35	7	118	11
ESMFold-Gap	0.64 ± 0.18	3.25 ± 3.87	0.89 ± 0.10	10	10	135	16
ESMFold-Linker*(ours)	0.65 ± 0.16	2.95 ± 3.46	0.90 ± 0.09	7	9	136	19
ESMFold-Linker*-Gap(ours)	0.65 ± 0.17	3.00 ± 3.49	0.90 ± 0.10	7	8	131	25
AF-multimer (v3 best)	0.70 ± 0.18	2.75 ± 3.59	0.91 ± 0.11	7	6	107	51

C Time analysis

We evaluate the speed of the model by testing on the VH-VL test set on a single Nvidia A100 80G GPU. As shown in Table 2, ESMFold-Linker* makes a prediction on a protein with 231 residues in 3.5 seconds, $9\times$ faster than a single AF-Multimer model. In addition, the search process in CPU for constructing MSAs for a protein can take >10 min with the high-sensitivity protocols used by the published version of AF2, which further increases the time needed for AF-Multimer to predict protein structures. In contrast, ESMFold-Linker* does not require MSAs, which is convenient and fast.

Table 2: Time analysis on the VH-VL test set.

(Seconds per sample)	MSA search time	Model inference time	Total time
ESMFold-Gap	0	2.7	2.7
ESMFold-Linker	0	3.5	3.5
ESMFold-Linker*	0	3.5	3.5
AF-Multimer(1 model)	>600	32.0	>632.0

D Effect of linker length

We study how the linker length L affects the performance of our model. For each linker length L in $\{5, 10, 15, 25\}$, we train a corresponding model on the heterodimer training samples with sequence lengths less than or equal to $220 - L$. The number of training data for each model is less than 614. For all models, the number of recycles is set to 1 and `residue_index_offset` is set to 0 during both training and inference. Figure 2 shows the weighted distogram loss on the validation set (left) as well as the top Ns/5 inter-chain contact precision on the heterodimer test set (right) with different linker lengths. As shown in Figure 2, we can see that when $L = 5$, the inter-chain contact performance of the model after training 100 epochs is much worse than the performance of the other three models with $L \geq 10$. The performance improves as L increases, and they achieve comparable performance when training to convergence. However, for models with smaller L , it takes longer to converge during training. In general, L should not be too small (≤ 5). We can choose the value of L in the range of $L \geq 10$, depending on the training and inference speed trade-off. Empirically, the model with $L = 25$ enjoys a fast convergence in training, although the inference time increases slightly (+0.8 seconds as shown in Table 2) than not using the linker.

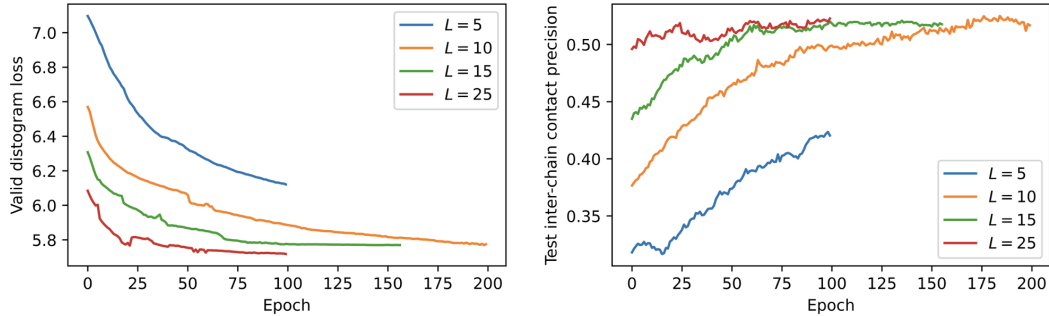


Figure 2: Effect of linker length L .

E Effect of weighted distogram loss

We study how the value of λ in the weighted distogram loss affects the performance of our model. For each λ in $\{2, 4, 6\}$, we train a corresponding model for 20K steps. For all models, the number of recycles is set to 1 and `residue_index_offset` is set to 0 during both training and inference. The top Ns/5 inter-chain contact precisions for models with $\lambda = 2, 4, 6$ on the heterodimer test set are 47.04%, 47.86%, and 47.57%, respectively. Although the model with $\lambda = 4$ achieves the best result, there is not much difference between the performances of the three models. It indicates that our method is not sensitive to the value of λ . Empirically, we set $\lambda = 4$.

Furthermore, for the model with $\lambda = 4$, we record the validation weighted distogram loss, validation inter-chain contact precision, and test inter-chain contact precision for every epoch during training. We find a Pearson correlation coefficient of -0.75 between the test inter-chain contact precision and the validation weighted distogram loss, which is larger in absolute value than the Pearson correlation coefficient of 0.47 between the test and validation inter-chain contact precision. This result suggests that the weighted distogram loss on the validation set correlates well with the quality of the predicted interface. Therefore, we recommend using the weighted distogram loss rather than the inter-chain contact precision as the model selection metric.

80 **F Data availability**

81 The training, validation, and test sets are available in the folder named data.

82 **G Code availability**

83 The codes for Linker-tuning are available in the folder named code. The model checkpoints are
84 available in the folder named checkpoint.

85 **References**

- 86 [1] Robert Verkuil, Ori Kabeli, Yilun Du, Basile IM Wicky, Lukas F Milles, Justas Dauparas, David Baker,
87 Sergey Ovchinnikov, Tom Sercu, and Alexander Rives. Language models generalize beyond natural proteins.
88 *bioRxiv*, pages 2022–12, 2022.