CO-FOLDING WITH A SOUP OF REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Co-folding models such as AlphaFold3, Protenix, ESMFold2, and OpenDDE have advanced rapidly, yet no single model consistently performs best across all biomolecular complexes. In this paper, we show that their pair representations encode complementary information that can be transferred across models to improve structure prediction. We introduce SOUPFOLD, which combines pair representations from multiple co-folding models in a common representation space and generates structures from the combined representation. Importantly, SOUPFOLD does not retrain the co-folding models and learns only simple mappings to transfer representations across models. We evaluate SOUPFOLD on antibody-antigen, protein-protein, protein-ligand, molecular glue, GPCR, and oligomeric complex prediction using AlphaFold3, Protenix, ESMFold2, and OpenDDE. By combining representations across models, SOUPFOLD improves over individual co-folding models across the considered benchmarks.

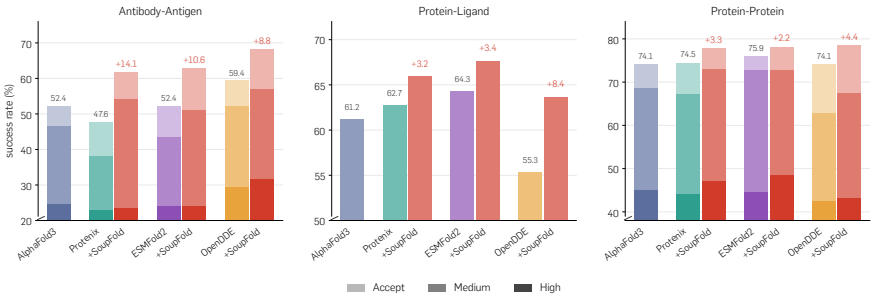


Figure 1: **Antibody-antigen, protein-ligand, and protein-protein predictions on FoldBench.** Accept, Medium, and High denote DockQ ≥ 0.23, ≥ 0.49, and ≥ 0.80. Protein-ligand success denotes LRMSD < 2 Å and lDDT-PLI > 0.8. SOUPFOLD improves performance over the baselines.

1 INTRODUCTION

Co-folding models have developed rapidly in recent years. Since AlphaFold3 (Abramson et al., 2024) extended structure prediction from single proteins to general biomolecular complexes, several models have followed, including Boltz (Wohlwend et al., 2025; Passaro et al., 2025), Chai (Chai Discovery team et al., 2024), Protenix (ByteDance AML AI4Science Team et al., 2025), OpenFold3 (The OpenFold3 Team, 2025), ESMFold2 (Candido et al., 2026), and OpenDDE (Aureka Research, 2026). These models share a similar pipeline: a Pairformer-based trunk constructs pair representations of a target complex, which are then used by a diffusion module to predict its structure. They mainly differ in their training recipes, including training data, objectives, and hyperparameters.

With several models now addressing the same prediction tasks, ensembling offers a way to improve accuracy without training a new model from scratch. This is especially promising because no single model consistently performs best at capturing structural patterns. For example, in an active-state GPCR system (Figure 2), where the (1) ligand pose and (2) coupled TM6 opening need to be captured jointly, AlphaFold3 predicts (1) the correct ligand pose but (2) misplaces the TM6 opening, whereas other models show the opposite behavior. Such differences suggest that the models capture complementary structural information.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

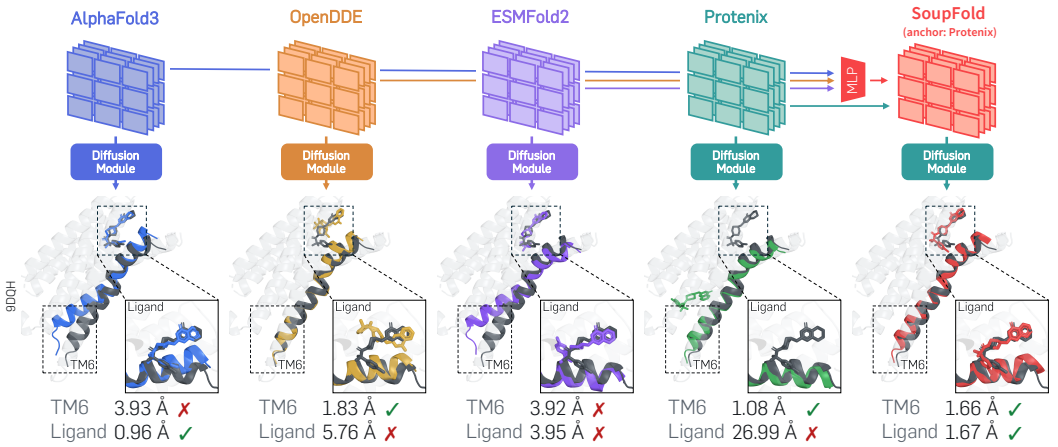


Figure 2: **Examples of SoupFold on high-order co-folding prediction.** Accurate co-folding structure prediction of active GPCRs requires capturing both the ligand pose that induces the active state and the corresponding active-state conformation of transmembrane helix 6 (TM6). Individual co-folding models often recover only one of these features, whereas SoupFold integrates their complementary structural information and accurately predicts both.

To ensemble complementary structural information across models, we note that these signals may already be encoded in the pair representations used to condition structure generation. We empirically compare representations across co-folding models and find that they are generally aligned, while their alignment weakens when the models predict different structures. These findings suggest a shared representational basis across models alongside complementary structural signals.

Motivated by these observations, we investigate whether transferring the representations across models can improve the standalone model's predictions. We thus ask:

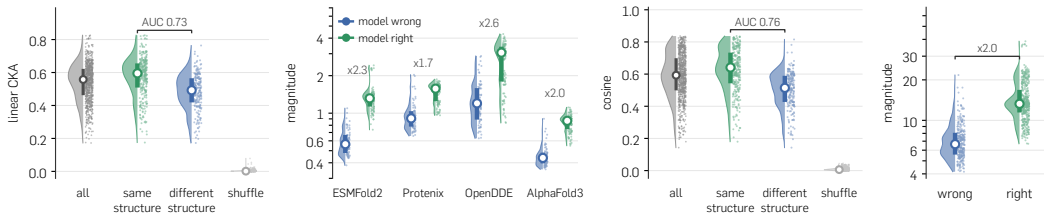
> *Can complementary information encoded in representations be transferred across co-folding models to improve structure prediction?*

To this end, we propose SoupFold, which ensembles representations from off-the-shelf co-folding models to generate high-quality structure predictions. Given an anchor model, lightweight transfer networks map representations from peer models into the anchor's representation space. The mapped representations are then combined with the anchor representation through a weighted average, augmenting its representation with structural signals captured by peers. The combined representation is then passed to the anchor model's diffusion module for structure generation. Importantly, SoupFold does not retrain the co-folding models and only trains the lightweight transfer networks.

We evaluate SoupFold using four cutting-edge co-folding models: AlphaFold3, Protenix-v1, ESMFold2, and OpenDDE. As benchmarks, we consider FoldBench (Xu et al., 2025) covering protein-protein, protein-ligand, and antibody-antigen complexes. We further consider some high-order benchmarks such as molecular glue ternary from MGBench (Liao et al., 2025), GPCR from Shen et al. (2025), and oligomer complexes from CASP16 (Zhang et al., 2026a). We also evaluate on two challenging benchmarks: the low-similarity subset of Runs N' Poses (Škrinjar et al., 2026) for protein-ligand prediction, and ARK-AB (Aureka Research, 2026) for antibody-antigen prediction.

Across these benchmarks, we show that

- SoupFold improves over standalone models and achieves the best or competitive performance across considered benchmarks, with larger relative gains on higher-order complexes that require multiple structural patterns to be captured jointly.
- SoupFold exploits complementary representations to improve prediction quality and can recover high-quality structures even when all standalone models fail.
- SoupFold outperforms sample-level ensembling across co-folding models and scaling standalone models with larger budgets.



(a) CKA and per-model magnitude.

(b) Cosine similarity and magnitude.

Figure 3: **Alignments and magnitudes of representations.** (a) Representational alignment, measured by Centered Kernel Alignment (CKA) (Kornblith et al., 2019), across model pairs, and interface representation magnitude measured by channel-wise standard deviation. Similar and dissimilar structures denote pairs with DockQ ≥ 0.8 and DockQ ≤ 0.23, respectively. Alignment weakens for dissimilar structures, while unsuccessful predictions have smaller representation magnitudes. (b) Trends after mapping representations into a shared space (ESMFold2 representation space).

2 SoupFold: Representation ensembling for co-folding

We introduce SoupFold, which improves structure prediction by ensembling pair representations across multiple co-folding models. We first describe the common structure of modern co-folding models (Section 2.1). We then present observations across these representations that support representation ensembling (Section 2.2) and describe the method (Section 2.3).

2.1 Co-folding representations

Modern co-folding models predict the joint structure of biomolecular complexes containing proteins, nucleic acids, and small molecules. Models such as AlphaFold3 (Abramson et al., 2024), Protenix (ByteDance AML AI4Science Team et al., 2025), ESMFold2 (Candido et al., 2026), and OpenDDE (Aureka Research, 2026) broadly decompose this task into two stages: encoding the complex into pair representations and generating structures conditioned on them.

A large trunk network first processes input features to construct pair representations, where each entry captures the structural features in biomolecules. Through recycling, the trunk iteratively refines the representations. A relatively lightweight diffusion module then generates atomic coordinates conditioned on them. At a high level, this process can be written as

$$H = \mathrm{Trunk}(\mathrm{Input}), \qquad \mathrm{Structure} \sim \mathrm{Diffusion}(\cdot \mid H),$$

where Input includes a biomolecular token sequence of length N, such as atoms or residues, and multiple sequence alignments (MSAs), while $H \in \mathbb{R}^{N \times N \times D}$ denotes the pair representation encoding relationships between pairs of tokens. Our method ensembles H from multiple co-folding models before generating the final structure using the $\mathrm{Diffusion}$ module. For convenience, we omit the single representations of co-folding models and do not ensemble them.

2.2 Observations

We characterize pair representations across different co-folding models. We find that their representations are generally well aligned, but alignment weakens when they predict different structures, suggesting that they share a representational basis while retaining potentially complementary structural signals. Meanwhile, unsuccessful representations tend to have smaller magnitudes, reducing the influence of potentially conflicting signals when ensembled.

Selectively aligned representations support ensembling. Pair representations from different co-folding models are generally well aligned. Their alignment is stronger when the models predict similar structures and weaker when their predictions differ (Figure 3a, left). This suggests that the representations share common structure while capturing complementary structural information when predictions diverge, providing a basis for representation ensembling.

Unsuccessful predictions correlate with small magnitude of representations. When grouped by prediction quality, representations associated with unsuccessful predictions tend to have smaller magnitudes, quantified by the channel-wise standard deviation at the interface (Figure 3a, right). These smaller magnitudes correspond to less differentiated patterns across channels, suggesting weaker structural signals rather than strong conflicting ones. Their weaker signals are therefore expected to have less influence than stronger signals from successful predictions when combined.

Representations can be mapped and aligned in a common space. We find that representations from different co-folding models can be meaningfully mapped from one model to another using a simple multi-layer perceptrons (MLPs). The mapped representations retain the patterns observed above: they are more aligned when the models predict similar structures, less aligned when their predictions differ, and smaller in magnitude for unsuccessful predictions (Figure 3b).

2.3 METHOD

From the observations in Section 2.2, we ensemble representations across multiple co-folding models. Given a set of co-folding models $\mathcal{M}$, we choose one model $m^\star \in \mathcal{M}$ as the anchor model, whose diffusion module generates the structure, and let the remaining models $\mathcal{M} \setminus m^\star$ be peers that provide additional representations. For each peer $m \in \mathcal{M} \setminus \{m^\star\}$, we train a transfer network $f_{m \to m^\star}$ that maps its representation into the anchor representation space for ensembling.

At inference time, we combine the representations by a weighted average in representation space of the anchor model. Each peer model's representation $H^{(m)}$ is mapped into the anchor model's representation space through a peer-to-anchor map $f_{m \to m^\star}$, and the peer representations are averaged with the anchor representation:

$$H = \frac{w_{m^\star} H^{(m^\star)} + \sum_{m \in \mathcal{M} \setminus \{m^\star\}} w_m f_{m \to m^\star}(H^{(m)})}{w_{m^\star} + \sum_{m \in \mathcal{M} \setminus \{m^\star\}} w_m}$$

where $w_{m^\star}$ and w_m are the coefficients for the anchor model and each peer, respectively. The updated representation H is then passed to the anchor model's diffusion module to predict the structure. In all experiments, we use $w_{m^\star} = 1$ and $w_m = 1$ for all peers by default.

Training transfer networks. We train a transfer network $f_{m \to m^\star}$ for each directed pair of co-folding models. Each map is a one-layer MLP of width 1024, applied independently to each pair entry. Given representations from two co-folding models for the same input, we train each transfer network with the following mean-squared error loss:

$$\mathcal{L}(f_{m \to m^\star}) = \mathbb{E}_{(H^{(m)}, H^{(m^\star)}) \in \mathcal{D}} \mathbb{E}_{(i,j)} \left[\left\| f_{m \to m^\star}(H_{ij}^{(m)}) - H_{ij}^{(m^\star)} \right\|_2^2 \right].$$

Here, $H_{ij}^{(m)}$ denotes the pair representation of model m for token pair (i, j). Note that we apply channel-wise normalization for training stability. The representation training set $\mathcal{D}$ is constructed from RCSB entries released before September 30, 2021, following the conventional training cutoff used by co-folding models and remaining temporally separated from most evaluation benchmarks. We provide additional details in Section A.

3 RESULTS

We evaluate whether SOUPFOLD can improve protein co-folding structure prediction by combining representations from multiple co-folding models. We consider AlphaFold3 (Abramson et al., 2024), Protenix-v1 (Zhang et al., 2026b), ESMFold2 (Candido et al., 2026), and OpenDDE (Aureka Research, 2026) on conventional antibody-antigen, protein-ligand, and protein-protein benchmarks, as well as higher-order complex benchmarks including molecular glue, GPCR, and oligomer complexes. We use Protenix-v1, ESMFold2, and OpenDDE as anchor models in turn, ensembling their representations with those of the other three peer models and generating structures using their diffusion modules. We do not use AlphaFold3 as an anchor model due to difficulties injecting representations into its JAX-compiled graph, but include it as a peer.

Table 1: **Co-folding prediction on FoldBench.** For antibody-antigen (170 interfaces) and protein-protein (274 interfaces) prediction, Accept, Medium, and High denote DockQ ≥ 0.23, ≥ 0.49, and ≥ 0.80. Protein-ligand success (526 interfaces) requires LRMSD < 2Å and lDDT-PLI > 0.8. The top line in each cell reports the best-confidence-of-25 result (5 samples per seed, 5 seeds) with gains of SoupFold over the anchor denoted with ↑. The bottom line reports the mean ± standard deviation over the 25 samples. SoupFold consistently improves the anchor model, and achieves the best performance. **Bold** indicates the better result within each anchor.

	AbAg			PL	PP		
Model	Accept	Medium	High	Success	Accept	Medium	High
AlphaFold3	52.35 43.25 ± 2.14	46.47 35.95 ± 1.73	24.71 18.07 ± 1.75	61.22 62.16 ± 0.68	74.09 74.09 ± 0.73	68.61 68.47 ± 0.20	44.89 44.45 ± 1.35
Protenix	47.65 42.73 ± 2.57	38.24 34.85 ± 1.55	22.94 18.54 ± 2.24	62.74 61.54 ± 0.77	74.45 72.34 ± 1.12	67.15 64.34 ± 1.57	44.16 42.18 ± 1.53
+SoupFold	**61.76** ↑14.11 55.53 ± 2.79	**54.12** ↑15.88 45.60 ± 2.35	**23.53** ↑0.59 20.42 ± 1.82	**65.97** ↑3.23 65.16 ± 1.00	**77.74** ↑3.29 75.69 ± 1.37	**72.99** ↑5.84 70.25 ± 1.26	**47.08** ↑2.92 44.35 ± 1.67
ESMFold2	52.35 46.66 ± 3.10	43.53 37.48 ± 2.68	**24.12** 18.59 ± 1.58	64.26 62.94 ± 0.66	75.91 73.14 ± 1.16	**72.63** 67.68 ± 1.12	44.53 38.73 ± 1.75
+SoupFold	**62.94** ↑10.59 52.96 ± 2.47	**51.18** ↑7.64 43.58 ± 1.71	**24.12** ↑0.00 20.12 ± 1.70	**67.68** ↑3.42 65.40 ± 0.92	**78.10** ↑2.19 75.33 ± 0.94	**72.63** ↑0.00 70.22 ± 0.95	**48.54** ↑4.01 45.07 ± 1.79
OpenDDE	59.41 56.16 ± 1.91	52.35 47.76 ± 1.57	29.41 25.48 ± 1.84	55.32 54.43 ± 0.71	74.09 72.74 ± 0.64	62.77 61.80 ± 0.58	42.34 39.56 ± 0.99
+SoupFold	**68.24** ↑8.83 61.51 ± 1.84	**57.06** ↑4.71 52.05 ± 1.26	**31.76** ↑2.35 28.14 ± 1.38	**63.69** ↑8.37 62.71 ± 0.81	**78.47** ↑4.38 77.37 ± 0.82	**67.52** ↑4.75 66.88 ± 0.66	**43.07** ↑0.73 41.58 ± 1.20

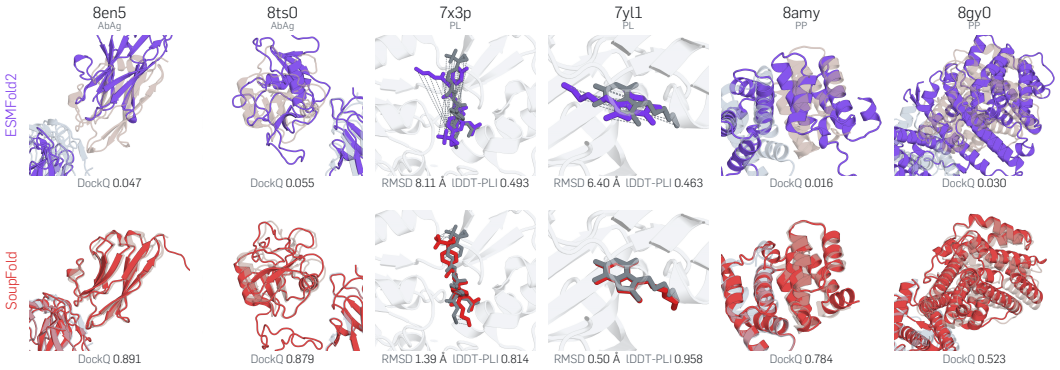


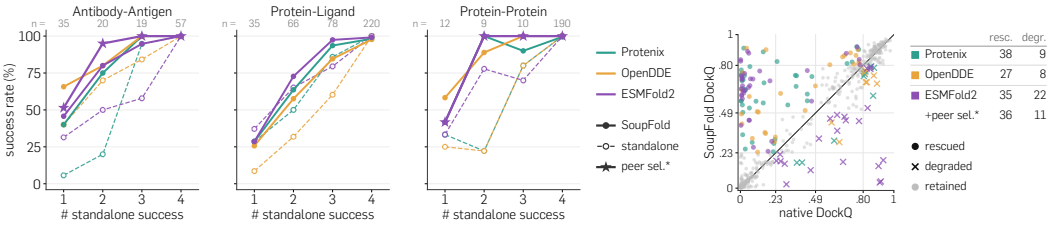
Figure 4: **SoupFold recovers failure predictions using peer representations.** Examples of antibody-antigen, protein-protein, and protein-ligand interfaces whose predictions are rescued by SoupFold using other peer representations.

Configuration. All models use unpaired and paired MSAs, 10 recycling steps, and their default diffusion steps. We use templates for AlphaFold3, Protenix, and OpenDDE, while ESMFold2 does not support the use of templates. Templates are prepared from PDBe structures released on or before September 30, 2021. The details are provided in Section B.

Benchmark. On FoldBench (Xu et al., 2025), we report DockQ-based success for antibody-antigen and protein-protein complexes, and LRMSD and lDDT-PLI-based success for protein-ligand complexes. We further evaluate higher-order complexes using MGBench molecular glue benchmark (Liao et al., 2025), GPCR benchmark (Shen et al., 2025), and CASP16 oligomer benchmark (Zhang et al., 2026a). MGBench evaluates protein-protein and protein-ligand success, while GPCR evaluates ligand-pose success and TM6 prediction. For CASP16 oligomers, we report TM-score, QSbest, DockQ, lDDT, ICS, and IPS. We additionally evaluate ARK-AB (Aureka Research, 2026) and the low-similarity subset of Runs N' Poses (Škrinjar et al., 2026) for challenging antibody-antigen and protein-ligand prediction, respectively. Detailed settings are provided in Section C.

Table 2: **Co-folding prediction on molecular-glue, GPCR, and oligomer benchmarks.** For molecular-glue (88 complexes) prediction, PLP (ternary success) requires PP success and PL success simultaneously. For GPCR (69 complexes) prediction, TM6-PL success requires the RMSD of the TM6 tip to be < 3.0 Å and PL success simultaneously. For oligomer (18 complexes) prediction, we report TM-score, QSbest, and DockQ. We report the best-confidence result (5 samples per seed, 5 seeds), and the mean ± standard deviation. See Sections D.1 to D.3 for full tables.

	Molecular Glue			GPCR			Oligomer		
Model	PLP	PP	PL	TM6-PL	TM6	PL	TM	QSbest	DockQ
AlphaFold3	31.82 33.9 ± 1.1	47.73 49.6 ± 2.7	32.95 34.1 ± 1.0	49.28 48.5 ± 2.6	1.24 1.22 ± .01	53.62 50.9 ± 2.3	0.838 0.802 ± .017	0.637 0.642 ± .016	0.511 0.478 ± .023
Protenix	28.41 31.2 ± 2.4	50.00 49.5 ± 1.3	28.41 31.5 ± 2.5	49.28 50.6 ± 2.7	1.32 1.34 ± .04	50.72 53.7 ± 2.5	0.826 0.760 ± .026	0.655 0.616 ± .024	0.488 0.443 ± .019
+SOUPFOLD	**36.36** 35.9 ± 1.2	**54.55** 51.9 ± 1.3	**36.36** 36.1 ± 1.2	**56.52** 51.0 ± 3.3	**1.19** 1.20 ± .02	**57.97** 53.3 ± 2.9	**0.832** 0.815 ± .026	**0.697** 0.657 ± .025	**0.551** 0.485 ± .026
ESMFold2	30.68 26.6 ± 2.2	**53.41** 48.6 ± 1.4	31.82 26.6 ± 2.2	43.48 42.4 ± 2.6	1.33 1.36 ± .03	47.83 45.9 ± 2.2	0.752 0.693 ± .033	0.605 0.570 ± .031	0.450 0.405 ± .027
+SOUPFOLD	**39.77** 34.1 ± 2.7	52.27 51.4 ± 1.7	**39.77** 34.2 ± 2.8	**50.72** 48.5 ± 2.1	**1.19** 1.19 ± .02	**53.62** 50.2 ± 2.4	**0.818** 0.757 ± .026	**0.659** 0.603 ± .031	**0.489** 0.446 ± .020
OpenDDE	**31.82** 31.1 ± 1.6	50.00 57.6 ± 3.1	**31.82** 31.1 ± 1.6	46.38 42.8 ± 2.8	1.39 1.35 ± .03	47.83 46.5 ± 2.2	0.651 0.658 ± .023	0.506 0.499 ± .013	0.454 0.447 ± .012
+SOUPFOLD	27.27 27.4 ± 1.9	**52.27** 55.2 ± 2.5	27.27 27.4 ± 1.9	**49.28** 49.5 ± 2.5	**1.18** 1.21 ± .03	**50.72** 51.3 ± 2.2	**0.721** 0.687 ± .018	**0.569** 0.529 ± .019	**0.530** 0.475 ± .020



(a) Success by the number of successful standalone models. (b) Sample-wise DockQ changes.

Figure 5: **Gains from SOUPFOLD.** (a) SOUPFOLD improves even when only two models succeed, consistent with failed representations having less influence (Section 2.2). (b) SOUPFOLD improves more antibody-antigen targets than it degrades. Peer sel.* denotes the variant in Table 4.

3.1 MAIN RESULTS

We evaluate whether representations from other models can improve the performance of an anchor model. In Table 1, SOUPFOLD consistently improves antibody-antigen, protein-ligand, and protein-protein prediction. For each anchor model, SOUPFOLD outperforms or is competitive with the strongest standalone model, with large gains on antibody-antigen prediction. These results suggest that independently trained models provide complementary representations that improve predictions.

Table 3: **Additional challenging benchmarks.** See full results in Sections D.4 and D.5.

Model	ARK	RnP*
AlphaFold3	47.66	28.89
Protenix	45.94	21.11
+SOUPFOLD	**56.08**	**25.56**
ESMFold2	48.68	27.78
+SOUPFOLD	**60.53**	**28.89**
OpenDDE	58.92	15.56
+SOUPFOLD	**64.35**	**24.44**

The performance gains are also notable on the higher-order complexes in Table 2. For molecular glue-induced ternary complexes, SOUPFOLD substantially improves joint success requiring both protein-protein and protein-ligand predictions to succeed. Similar gains are observed for GPCR complexes, where joint success requires both accurate ligand-pose and TM6-branch prediction. For oligomer complexes, Protenix+SOUPFOLD matches AlphaFold3 in TM-score while improving QSbest and DockQ, indicating complementary gains in global and interface quality. Additionally, SOUPFOLD consistently improves over the anchor models on ARK-AB and the low-similarity subset of Runs N' Poses (Table 3). On Runs N' Poses, SOUPFOLD matches AlphaFold3 in success rate but further improves RMSD (Section D.5).

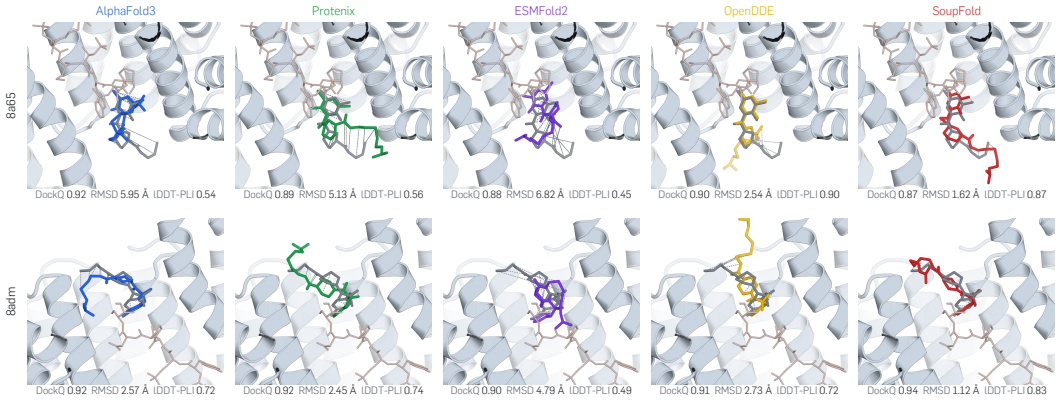


Figure 8: **SOUPFOLD succeeds even when all standalone co-folding models strictly fail on molecular glue ternary complexes.** Molecular glue examples where SOUPFOLD succeeds with its highest-confidence prediction, while all standalone models fail. For 8a65, SOUPFOLD succeeds with ESMFold2 and OpenDDE, and for 8adm, with all three anchors. Notably, the standalone models still fail even under oracle selection over one hundred generated samples.

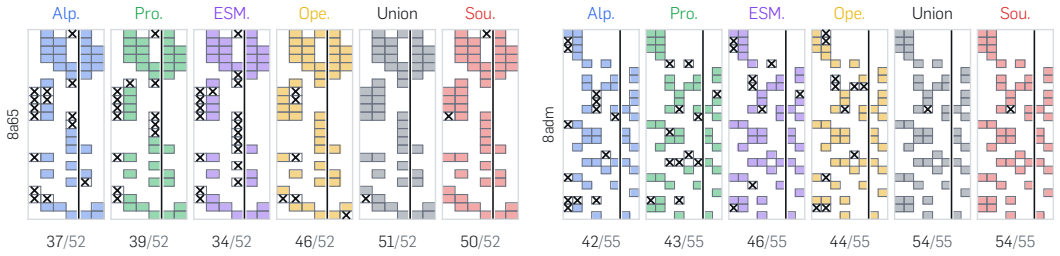


Figure 9: **SOUPFOLD recovers complete interfaces (Figure 8) by combining complementary information scattered across representations.** We illustrate residue-atom contacts for examples in Figure 8. A contact is filled when it is recovered in at least 5 of 25 samples, while ☒ marks a missed contact. Each standalone model captures only a subset of the contacts, while their union recovers nearly all ground-truth contacts. This suggests that SOUPFOLD combines complementary contacts across model representations to recover the high-quality structure.

3.2 PERFORMANCE GAINS ARE ASSOCIATED WITH COMPLEMENTARY REPRESENTATIONS

We further examine when SOUPFOLD improves prediction quality. As shown in Figure 5a, SOUPFOLD improves over the standalone models even when only two of four models succeed, consistent with unsuccessful representations having less influence on the ensemble (Section 2.2). Figure 5b further shows that, although most predictions retain their original interface quality, improvements are more frequent than degradations. However, the benefit varies substantially across targets.

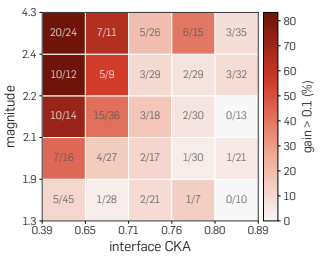


Figure 6: **CKA and magnitude vs. DockQ gain.** Dissimilar high-magnitude representation benefits.

We hypothesize that larger gains stem from stronger complementary signals, giving SOUPFOLD more useful information to exploit. We compare DockQ gains with representation similarity between anchor and peers, shown in Figure 6. Here, the similarity is calculated via CKA. As hypothesized, we find gains to be larger when peer representations are less aligned with the anchor and have larger magnitudes, suggesting that stronger complementary representations can provide greater benefit.

Table 4: **Teacher selection by representation complementarity.** ESMFold2 is the anchor model. Peers are ranked by increasing CKA to the anchor representation, measured after mapping. Top-1 and Top-2 use the first one and two teachers, respectively. Equal uses equal weights. Weighted scales each by its interface magnitude. Accept, medium, and high rates follow the DockQ thresholds in Table 1. Gains over the anchor are denoted by ↑.

		Antibody-Antigen			Protein-Protein		
Criterion	Selection	Accept	Medium	High	Accept	Medium	High
ESMFold2	(No Peer)	52.35	43.53	24.12	75.91	72.63	44.53
Equal	All	62.94 ↑10.6	51.18 ↑7.6	24.12 ↑0.0	78.10 ↑2.2	72.63 ↑0.0	48.54 ↑4.0
	Top-1	65.29 ↑12.9	55.88 ↑12.3	25.88 ↑1.8	78.83 ↑2.9	73.36 ↑0.7	45.26 ↑0.7
	Top-2	62.94 ↑10.6	52.94 ↑9.4	26.47 ↑2.3	78.83 ↑2.9	72.63 ↑0.00	47.08 ↑2.5
Weighted	All	64.12 ↑11.8	54.12 ↑10.6	25.88 ↑1.8	77.74 ↑1.8	72.63 ↑0.0	48.91 ↑4.4
	Top-1	66.47 ↑14.1	55.88 ↑12.3	26.47 ↑2.3	79.20 ↑3.3	74.09 ↑1.5	43.80 ↓0.7
	Top-2	62.94 ↑10.6	52.94 ↑9.4	26.47 ↑2.3	77.74 ↑1.8	72.26 ↓0.4	47.08 ↑2.5

Table 5: **SoupFold vs. best-confidence selection from four pooled standalone models.** We consider FoldBench, MGBench ternary complexes, CASP16 oligomers, ARK, and the low-similarity subset of Runs N' Poses. SoupFold remains competitive with or outperforms.

Method	AbAg	PL	PP	MG	GPCR	Oligo.	AbAg_{ARK}	PL_{RNP}
Pooled (100 samples)	60.59	63.49	78.10	31.81	46.38	0.797	59.67	24.44
Protenix + SoupFold	61.76	65.97	77.74	36.36	56.52	0.826	56.08	25.56
ESMFold2 + SoupFold	62.94	67.78	78.10	39.77	50.72	0.818	60.53	28.89
OpenDDE + SoupFold	68.24	63.69	78.47	27.27	49.28	0.720	64.35	24.44

SoupFold combines low-quality standalone representations into high-quality predictions. SoupFold is effective for higher-order complexes (Table 2), which require multiple structural patterns to be captured jointly and can benefit from combining partial information across standalone models. For example, in the GPCR example in Figure 2, standalone models correctly capture either the ligand pose or the TM6 opening, but not both, while SoupFold recovers both by integrating complementary information.

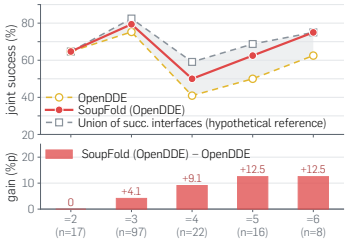


Figure 7: **Entry-level success for multi-interface entries.** Gains increase with the number of chains.

We further examine this benefit by extending the standard interface-level evaluation of ARK-AB to the entry level, where success requires all interfaces in an entry to have DockQ > 0.23. As entries contain more chains and interfaces, individual models may capture only subsets of them, while SoupFold can combine complementary information across models to recover the full entry. Accordingly, in Figure 7, the gain from SoupFold increases with the number of chains, while its performance approaches a hypothetical upper bound given by the union of successful interfaces from all standalone models.

More interestingly, complementarity also appears at a finer interface level. As shown in Figures 8 and 9, each standalone model recovers only a subset of native contacts, while their union covers nearly all contacts required for the correct structure. Thus, even when no individual model produces a successful prediction, their representations contain complementary interface signals that SoupFold integrates to recover a high-quality structure.

Exploiting representation complementarity for peer selection. Based on the representation complementarity observed above, we design peer selection to favor stronger complementary signals. Rather than using all peers with equal weights, we consider weighting peers by their interface representation magnitude and selecting peers with the lowest CKA to the anchor representation. In Table 4, both strategies improve overall performance. However, they can also reduce performance on high-quality predictions for protein-protein complexes, as excluding peers may remove useful information and magnitude weighting may amplify incorrect signals. This suggests a trade-off in peer selection that depends on the task.

3.3 SAMPLE-LEVEL AND SEED-LEVEL SCALING VS. SOUPFOLD

We compare SOUPFOLD with sample-level ensembling across co-folding models, which pools predictions across multiple standalone models and selects the highest-confidence structure. As shown in Table 5, despite comparing only 25 SOUPFOLD samples against 100 pooled standalone samples, SOUPFOLD is competitive or substantially better. This also reflects the miscalibration of confidence scores across models to select best-confidence samples. In contrast, SOUPFOLD combines information before structure generation and selects samples within a single anchor model.

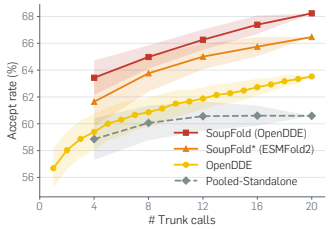


Figure 10: **Scaling budget.** Our method beats scaling OpenDDE.

We next compare SOUPFOLD with scaling a standalone model under a larger sampling budget. This is particularly important for antibody-antigen prediction, where prior work reports improved performance with more seeds (Candido et al., 2026; Abramson et al., 2024). Each SOUPFOLD seed evaluates all four model trunks once, so we allocate the standalone model four times as many seeds, matching the number of trunk evaluations while giving it a larger diffusion sampling budget. We use OpenDDE, the strongest standalone model for antibody-antigen prediction, despite its substantially larger and more expensive trunk. As shown in Figure 10, SOUPFOLD still outperforms scaling OpenDDE with the larger sampling budget.

4 RELATED WORKS

Protein co-folding models. Modern co-folding models predict the structure of biomolecular complexes containing proteins, nucleic acids, and small molecules. AlphaFold3 (Abramson et al., 2024) introduce a co-folding framework for biomolecular structure prediction, followed by several models including, Boltz (Wohlwend et al., 2025; Passaro et al., 2025), Chai (Chai Discovery team et al., 2024), Protenix (ByteDance AML AI4Science Team et al., 2025), OpenFold3 (The OpenFold3 Team, 2025), ESMFold2 (Candido et al., 2026), and OpenDDE (Aureka Research, 2026). While Protenix closely follows the AlphaFold3 training recipe, ESMFold2 builds on representations from the pretrained 6B ESMC protein language model, and OpenDDE uses a multi-stage curriculum with emphasis on challenging interactions such as antibody-antigen complexes.

Representations in co-folding models. Several studies have examined the internal representations of co-folding models. AlphaLink and AlphaLink2 (Stahl et al., 2023; 2024) inject crosslink-derived distance restraints into pair representations to guide monomer and complex structure prediction toward specific conformations. More recently, Jang et al. (2026) use representations from Boltz as atom-level molecular features and show that they transfer to downstream tasks. Suzuki and Amagasa (2026) show that scaling the pair representations of AlphaFold3 alters its conformational sampling. In contrast, we transfer representations across multiple co-folding models.

Representation distillation and ensembling. Representation distillation transfers information between models by aligning intermediate representations of pretrained teacher and student models (Ahn et al., 2019). Recent work has extended this idea to generative models for images (Yu et al., 2025) and protein structures (Kim et al., 2026). Beyond distillation, independently trained models can be combined by averaging their parameters, as in Model Soups (Wortsman et al., 2022), or by learning mappings between their representation spaces (Bansal et al., 2021).

5 CONCLUSION

We study whether co-folding models encode complementary information in their internal representations and whether it can be transferred across models to improve structure prediction. We introduce SOUPFOLD, which learns lightweight mappings between pair-representation spaces and combines representations before structure generation. Without retraining the underlying models, SOUPFOLD improves prediction performance across standard, challenging, and higher-order complex benchmarks. Future work could explore adaptive peer weighting using representation-level interface uncertainty, or mixture-of-experts-style co-folding models specialized in complementary structural patterns.

AI USE STATEMENT

We used Claude Code to assist with author-specified data processing, implementation of author-specified methods, debugging, and resolving package conflicts or kernel issues when running co-folding models. We also used it to assist with visualizations and to polish figures and tables. It was not used to determine the methodology or experimental settings, which were determined by the authors. All AI-assisted outputs were reviewed and verified by the authors, and all results were interpreted by the authors. ChatGPT was used for writing and language polishing. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

This work could be misused to support the design of harmful compounds. All co-folding models used in this study were used for non-commercial academic research in accordance with their respective licenses or terms of use. This work uses publicly available biomolecular structure data and benchmarks and does not involve human participants, private data, or animal experiments.

REPRODUCIBILITY STATEMENT

We provide full details of the method, training, experimental settings, and evaluation procedures in the Appendix. The Supplementary Material includes the code and trained transfer-network weights.

REFERENCES

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.

Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9155–9163. IEEE, 2019.

Aureka Research. Folding, reasoning, and scaling with open-source drug discovery engine, 2026. URL https://github.com/aurekaresearch/OpenDDE.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. *Advances in neural information processing systems*, 34:225–236, 2021.

ByteDance AML AI4Science Team, Xinshi Chen, Yuxuan Zhang, Chan Lu, Wenzhi Ma, Jiaqi Guan, Chengyue Gong, Jincai Yang, Hanyu Zhang, Ke Zhang, et al. Protenix-advancing structure prediction through a comprehensive alphafold3 reproduction. *BioRxiv*, pages 2025–01, 2025.

Salvatore Candido, Thomas Hayes, Alexander Derry, Roshan Rao, Zeming Lin, Robert Verkuil, Bryan Z Wu, Jin Sub Lee, Elise S Bruguera, Jehan A Keval, et al. Language modeling materializes a world model of protein biology. *bioRxiv*, pages 2026–06, 2026.

Chai Discovery team, Jacques Boitreaud, Jack Dent, Matthew McPartlon, Joshua Meier, Vinicius Reis, Alex Rogozhonikov, and Kevin Wu. Chai-1: Decoding the molecular interactions of life. *BioRxiv*, pages 2024–10, 2024.

Hyosoon Jang, Hyunjin Seo, Honghui Kim, Seonghyun Park, Taewon Kim, Yunhui Jang, and Sungsoo Ahn. A systematic evaluation of co-folding model representations for small-molecule learning. *arXiv preprint arXiv:2602.13249*, 2026.

Taewon Kim, Hyosoon Jang, Hyunjin Seo, Seonghwan Seo, Hyeongwoo Kim, Wonho Zhung, Mingyeong Shin, Wooyoun Kim, and Sungsoo Ahn. Atom-level protein representation learning improves protein structure prediction. *arXiv preprint arXiv:2605.22133*, 2026.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMlR, 2019.

Yiyan Liao, Jintao Zhu, Juan Xie, Luhua Lai, and Jianfeng Pei. Benchmarking cofolding methods for molecular glue ternary structure prediction. *Journal of Chemical Information and Modeling*, 65 (20):11136–11148, 2025.

Claudio Mirabello and Björn Wallner. Dockq v2: improved automatic quality measure for protein multimers, nucleic acids, and small molecules. *Bioinformatics*, 40(10):btae586, 2024.

Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, et al. Boltz-2: Towards accurate and efficient binding affinity prediction. *BioRxiv*, 2025.

Shi-yi Shen, Jun-rui Li, Yu-song Wang, Shao-ning Li, H Eric Xu, and Xin-heng He. An update for alphafold3 versus experimental structures: assessing the precision of small molecule binding in gpcrs. *Acta Pharmacologica Sinica*, 46(12):3355–3364, 2025.

Peter Škrinjar, Jérôme Eberhardt, Gabriel Studer, Gerardo Tauriello, Torsten Schwede, and Janani Durairaj. Evaluating generalization in protein–ligand cofolding methods. *Nature Structural & Molecular Biology*, pages 1–13, 2026.

Kolja Stahl, Andrea Graziadei, Therese Dau, Oliver Brock, and Juri Rappsilber. Protein structure prediction with in-cell photo-crosslinking mass spectrometry and deep learning. *Nature biotechnology*, 41(12):1810–1819, 2023.

Kolja Stahl, Robert Warneke, Lorenz Demann, Rica Bremenkamp, Björn Hormes, Oliver Brock, Jörg Stülke, and Juri Rappsilber. Modelling protein complexes with crosslinking mass spectrometry and deep learning. *Nature communications*, 15(1):7866, 2024.

Shosuke Suzuki and Toshiyuki Amagasa. Biasing conformational sampling in alphafold 3 and boltz-2 via pair representation scaling. *Journal of Chemical Information and Modeling*, 66(17): 10862–10876, 2026.

The OpenFold3 Team. Openfold3-preview, 2025. URL https://github.com/aqlaboratory/openfold-3.

Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Noah Getz, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, Liam Atkinson, Tally Portnoi, Itamar Chinn, et al. Boltz-1 democratizing biomolecular interaction modeling. *BioRxiv*, pages 2024–11, 2025.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pages 23965–23998. Pmlr, 2022.

Sheng Xu, Qiantai Feng, Lifeng Qiao, Hao Wu, Tao Shen, Yu Cheng, Shuangjia Zheng, and Siqi Sun. Benchmarking all-atom biomolecular structure prediction with foldbench. *Nature Communications*, 17(1):442, 2025.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations*, 2025.

Jing Zhang, Rongqing Yuan, Andriy Kryshtafovych, Jimin Pei, Rachael C Kretsch, R Dustin Schaeffer, Jian Zhou, Rhiju Das, Nick V Grishin, and Qian Cong. Assessment of protein complex predictions in casp16: Are we making progress? *Proteins: Structure, Function, and Bioinformatics*, 94(1): 106–130, 2026a.

Yuxuan Zhang, Chengyue Gong, Hanyu Zhang, Wenzhi Ma, Zhenyu Liu, Xinshi Chen, Jiaqi Guan, Lan Wang, Yanping Yang, Yu Xia, and Wenzhi Xiao. Protenix-v1: Toward high-accuracy open-source biomolecular structure prediction. *bioRxiv*, 2026b. doi: 10.64898/2026.02.05.703733. URL https://www.biorxiv.org/content/early/2026/02/22/2026.02.05.703733.1.

A DETAILS OF SOUPFOLD

Training transfer networks. We train a transfer network $f_{m\to m^\star}$ for each directed pair of ESMFold2, AlphaFold3, Protenix, and OpenDDE, where m denotes the peer model and $m^\star$ denotes the anchor model. Each map is a one-layer MLP of width 1024, applied independently to each pair entry. The corresponding pair representation dimensions are 256, 128, 128, and 384, respectively.

Pair representations are obtained after ten recycling steps and standardized per channel using the channel-wise mean and standard deviation computed on the held-out training set. For model m, $H_{ij}^{(m)} \in \mathbb{R}^{C_m}$ denotes the pair representation for token pair (i, j). Each transfer network is trained with

$$\mathcal{L}(f_{m\to m^\star}) = \mathbb{E}_{(H^{(m)},H^{(m^\star)})\in\mathcal{D}}\left[\mathbb{E}_{(i,j)}\left\|f_{m\to m^\star}(\hat{H}_{ij}^{(m)}) - \hat{H}_{ij}^{(m^\star)}\right\|_2^2\right].$$

Here, $\hat{H}^{(m)}$ denotes the channel-wise standardized representation. Because the models use different tokenizations, we construct token maps across models and define the transfer loss only on corresponding pair entries. We use random crops of 512 tokens with AdamW, a learning rate of 3×10^{-4}, no weight decay, and gradient clipping at 1.0.

Held-out training data for transfer networks. We train the transfer maps on a held-out set collected from the RCSB PDB with deposition dates before September 30, 2021. We restrict the set to protein-only complexes with at least two chains. To reduce redundancy, we cluster distinct protein sequences using MMseqs2 at 40% sequence identity and 80% coverage. We then group complexes according to the sequence clusters of their constituent chains and retain one representative per complex family. From the resulting candidates, we select 3,000 structurally diverse complexes.

For each complex, we prepare MSAs and template features using the same preprocessing pipeline across models. Templates are restricted to structures released before the corresponding cutoff. We extract pair representations from all four co-folding models after ten recycling iterations. The remaining systems are discarded so that every transfer map is trained on the same set of complexes. We train the transfer networks for five epochs.

Token maps. The four models may occasionally tokenize and order the same complex differently. Thus, we construct token maps between models using normalized token fingerprints based on chain, residue index, and representative atom. For each anchor-peer pair, the token map matches corresponding tokens and leaves unmatched or ambiguous tokens undefined. A pair entry (i, j) is matched only when both tokens have valid correspondences.

Inference-time representation ensembling. At inference, we choose one model $m^\star$ as the anchor and use the remaining models as peers. Each model independently produces its pair representation $H^{(m)}$ after ten recycling iterations. Peer representations are aligned to the anchor token order using the token maps, standardized using their channel statistics, and mapped into the anchor representation space using the corresponding transfer networks. We then compute

$$\hat{H} = \frac{w_{m^\star}\hat{H}^{(m^\star)} + \sum_{m\neq m^\star} w_m f_{m\to m^\star}\left(\tilde{\hat{H}}^{(m)}\right)}{w_{m^\star} + \sum_{m\neq m^\star} w_m}, \qquad H' = \sigma^{(m^\star)} \odot \hat{H} + \mu^{(m^\star)},$$

where $\mu^{(m^\star)}$ and $\sigma^{(m^\star)}$ are the anchor channel statistics. Unmatched peers are excluded from both the numerator and denominator for the corresponding pair entry. We use uniform weights $w_{m^\star} = w_m = 1$ throughout. The resulting H' replaces the anchor pair representation, while its single representation and all other inputs remain unchanged. The anchor diffusion module then generates the final structure conditioned on H'.

B Model Configuration

B.1 Input preparation

Sequences. We construct all inputs from the official FoldBench ground-truth mmCIF. We first construct one canonical record. From this record, we generate the model-specific input for each of the four models. Thus, all four models receive the same chains, sequences, and ligands. Polymer sequences are obtained from SEQRES. Ligands, ions, and cofactors are included with their deposited copy numbers. Non-standard polymer residues are represented as modifications, except residues canonicalized to their parent amino acid by the reference pipelines (MSE, SEC, and PYL).

MSAs. We generate multiple sequence alignments once using the ColabFold MMseqs2 server with UniRef30 and environmental databases and share them across all models. For complexes containing at least two protein chains, including homo-oligomers, we additionally generate a paired A3M. Following the Protenix convention, paired alignments are also appended to the unpaired MSA block using `msa_pair_as_unpair = True`.

Templates. Template search follows the AlphaFold 3 pipeline. We run `hmmbuild` and `hmmsearch` with the AF3 settings against the PDB SeqRes database, with one search per unique sequence. We retain only PDB entries released on or before September 30, 2021. The top template hits are retrieved as mmCIF files from PDBe and stored locally. Each model then applies its own template filtering and uses at most four templates per chain.

B.2 Models, checkpoints, and source modifications

Model	Checkpoint	C_z
AlphaFold 3	alphafold3 3.0.1, released weights	128
OpenDDE	`opendde_v1`	384
Protenix	`protenix_base_default_v1.0.0`	128
ESMFold2	esm 3.4.0, ESMFold2 + ESMC-6B language model	256

Table 6: Packages, checkpoints, and pair-representation width C_z.

Source modifications for consistent tokenization. For SoupFold exclusively, we make minor modifications to three packages so that their pair representations are better aligned token-by-token.

- **AF3** (`atom_layout.py`, 2 lines): we retain the glycan O1 atom, which AF3 otherwise removes and which changes the token count for glycan-containing systems.
- **Protenix** and **OpenDDE** (`data/core/parser.py`, 8 and 4 lines): entities specified as ligands or ions are tokenized per atom even when their CCD type is "protein". This handles free amino-acid ligands such as ILE, VAL, and TRP consistently with AF3 and ESMFold2.
- **ESMFold2**: no source modifications.

B.3 Inference Configuration

By default, we use 10 recycling steps. We use the default number of diffusion steps for each co-folding model. We generate 25 samples per target. We select the highest-confidence prediction using each model's default confidence score. Each system produces five diffusion samples per seed.

B.4 Representation extraction

For each system and model, we run the full network once with 10 recycling iterations and store the final trunk pair representation $z \in \mathbb{R}^{N \times N \times C_z}$. All representations are stored in fp32.

C BENCHMARK

FoldBench. We use the official FoldBench protocol (Xu et al., 2025). We evaluate the protein-protein, antibody-antigen, and protein-ligand subsets. Ligands are specified by their CCD codes. We report results on the systems for which all four co-folding models produce a prediction: 274 interfaces for protein-protein, 170 interfaces for antibody-antigen, and 526 poses for protein-ligand complexes. Following official metrics, for protein-protein and antibody-antigen complexes, we compute DockQ with `DockQ v2.1.3` (Mirabello and Wallner, 2024) and report the success rate at DockQ ≥ 0.23, 0.49, and 0.80. For protein-ligand complexes, we compute lDDT-PLI and LRMSD (BiSyRMSD) with OpenStructure and count a ligand as successful when LRMSD < 2 Å and lDDT-PLI > 0.8.

MGBench. We evaluate molecular glue-induced ternary complexes using MGBench (Liao et al., 2025). We use the published ternary protein chains and molecular-glue annotations and evaluate the 88 complexes successfully processed by all four co-folding models. Each system contains the two annotated protein chains and the molecular glue specified by its CCD code, and the deposited mmCIF structure is used as ground truth. Following official metrics, we evaluate the protein-protein interface using DockQ and define protein-protein success as DockQ ≥ 0.23. For the molecular-glue pose, we compute LRMSD and lDDT-PLI with OpenStructure, and define protein-ligand success as LRMSD < 2 Å and lDDT-PLI > 0.8. We additionally consider ternary success that requires both protein-protein and protein-ligand success for the same predicted structure.

GPCR. We evaluate GPCR complexes using the benchmark from the recent study (Shen et al., 2025). We consider the 69 systems successfully processed by all four co-folding models. Following benchmark settings, each input contains the receptor, the transducer when present, and the bound small molecules specified by their CCD codes. Following Shen et al. (2025), crystallization aids and receptor fusion proteins are removed, with fusion regions replaced by the native receptor sequence. Following the official measures, we use TM6 RMSD and ligand LRMSD. We align predictions to the ground truth using the Cα atoms of TM1-5 and TM7 and compute the TM6 Cα RMSD. We additionally define P-L success as LRMSD < 2 Å for every ligand in the complex. We additionally define TM6 success as a Cα RMSD < 3 Å. We additionally consider TM6-P-L success that requires both criteria to be satisfied by the same prediction.

CASP16 oligomer. We use the CASP16 (Zhang et al., 2026a) oligomer targets with their published stoichiometry and the official target structures as ground truth. We evaluate the 18 targets for which all four co-folding models produce a prediction. Following official metrics, we score each prediction with OpenStructure `compare-structures` and report the CASP16 assembly metrics: interface DockQ weighted by the number of native contacts, ICS, IPS, QS-best, lDDT, and TM-score.

ARK-AB. We use the ARK-AB antibody-antigen benchmark released with OpenDDE (Aureka Research, 2026), with its interface definitions. Each input is the first biological assembly of the entry. We evaluate the 162 entries for which all four co-folding models produce a prediction, comprising 398 interfaces in 155 interface clusters. Interface clusters are defined by the benchmark as unordered pairs of MMseqs2 entity clusters (40 % sequence identity, 80 % coverage). We compute DockQ for each interface and report the success rate at DockQ ≥ 0.23, 0.49, and 0.80. Following the official metrics, the success rate is normalized over interface clusters: we compute the fraction of successful interfaces within each cluster and average this fraction over clusters.

Runs N' Poses. We use the challenging subset, 0-20 % similarity bin, of Runs N' Poses benchmark (Škrinjar et al., 2026). Here, ligands are specified by their CCD codes. We evaluate the 83 systems (90 ligands) for which all four co-folding models produce a prediction. Following official metrics, we compute lDDT-PLI and BiSyRMSD with OpenStructure and count a ligand as successful when BiSyRMSD < 2 Å and lDDT-PLI > 0.8.

D ADDITIONAL RESULTS

D.1 MGBENCH

Table 7: **Performance on MGBench molecular-glue ternary complexes.** We evaluate the 88 molecular glue ternary complexes for which all four co-folding models produce a prediction. PLP counts a system as successful only when PP (DockQ ≥ 0.23) and PL (LRMSD < 2 Å and lDDT-PLI > 0.8) hold for the same structure. The first line in each cell reports the highest-confidence prediction among 25 samples (5 samples per seed, 5 seeds). The second line reports the mean $\pm$ standard deviation over the 25 samples.

Model	PLP (%)	PP (%)	PL (%)	DockQ ↑	LRMSD ↓	lDDT-PLI ↑
AlphaFold3	31.82 33.91 ± 1.05	47.73 49.64 ± 2.63	32.95 34.05 ± 0.99	0.439 0.452 ± 0.011	4.77 5.10 ± 0.69	0.637 0.645 ± 0.010
ESMFold2	30.68 26.55 ± 2.13	**53.41** 48.59 ± 1.34	31.82 26.59 ± 2.11	0.481 0.453 ± 0.009	4.94 5.71 ± 0.70	0.669 0.659 ± 0.009
+SOUPFOLD	**39.77** 34.14 ± 2.66	52.27 51.36 ± 1.64	**39.77** 34.18 ± 2.73	**0.487** 0.466 ± 0.010	**4.38** 4.98 ± 0.82	**0.702** 0.664 ± 0.016
Protenix	28.41 31.18 ± 2.34	50.00 49.50 ± 1.25	28.41 31.45 ± 2.40	0.459 0.454 ± 0.008	**5.76** 5.48 ± 0.23	0.656 0.663 ± 0.017
+SOUPFOLD	**36.36** 35.86 ± 1.16	**54.55** 51.91 ± 1.27	**36.36** 36.05 ± 1.14	**0.479** 0.475 ± 0.007	5.93 4.96 ± 0.35	**0.672** 0.683 ± 0.007
OpenDDE	**31.82** 31.09 ± 1.54	50.00 57.55 ± 3.08	**31.82** 31.09 ± 1.54	0.471 0.491 ± 0.013	5.45 4.96 ± 0.27	**0.685** 0.694 ± 0.009
+SOUPFOLD	27.27 27.41 ± 1.86	**52.27** 55.23 ± 2.41	27.27 27.41 ± 1.86	**0.484** 0.489 ± 0.010	**4.93** 4.90 ± 0.18	0.683 0.688 ± 0.005

D.2 GPCR

Table 8: **Performance on the GPCR benchmark.** The benchmark comprises 69 complexes for which all four co-folding models produce a prediction. TM6 and TM6 tip are the mean Cα RMSD (Å) of the TM6 helix and of its cytoplasmic end, after superposition on TM1-5 and TM7. Following official metrics, P-L success requires LRMSD < 2 Å for every ligand in the complex. TM6-P-L additionally requires TM6-tip RMSD < 3.0 Å in the same prediction. We report the best-confidence result (5 samples per seed, 5 seeds), and the mean $\pm$ standard deviation over the 25 samples.

Model	TM6-P-L↑	TM6 (Å)↓	TM6 tip (Å)↓	P-L↑	LRMSD (Å)↓
AlphaFold3	49.28 48.46 ± 2.60	1.24 1.22 ± .01	1.39 1.38 ± .02	53.62 50.90 ± 2.29	5.47 5.06 ± .53
Protenix	49.28 50.55 ± 2.73	1.32 1.34 ± .04	1.58 1.62 ± .07	50.72 53.68 ± 2.54	6.13 5.95 ± .61
+SOUPFOLD	**56.52** ↑7.25 51.01 ± 3.28	**1.19** ↑.13 1.20 ± .02	**1.42** ↑.16 1.44 ± .03	**57.97** ↑7.25 53.28 ± 2.91	**4.49** ↑1.64 5.04 ± .71
ESMFold2	43.48 42.38 ± 2.57	1.33 1.36 ± .03	1.72 1.73 ± .07	47.83 45.86 ± 2.24	5.33 5.45 ± .33
+SOUPFOLD	**50.72** ↑7.25 48.46 ± 2.13	**1.19** ↑.15 1.19 ± .02	**1.40** ↑.32 1.40 ± .04	**53.62** ↑5.80 50.20 ± 2.39	**4.15** ↑1.19 4.83 ± .39
OpenDDE	46.38 42.84 ± 2.78	1.39 1.35 ± .03	1.59 1.56 ± .05	47.83 46.49 ± 2.20	5.75 5.35 ± .32
+SOUPFOLD	**49.28** ↑2.90 49.45 ± 2.50	**1.18** ↑.20 1.21 ± .03	**1.33** ↑.25 1.41 ± .05	**50.72** ↑2.90 51.30 ± 2.21	**5.15** ↑.59 5.26 ± .58

D.3 CASP16 OLIGOMER BENCHMARKS

Table 9: **Performance on CASP16 oligomer complexes.** The benchmark comprises 18 oligomer complexes for which all four co-folding models produce a prediction. The first line in each cell reports the highest-confidence prediction among 25 samples (5 samples per seed, 5 seeds). The second line reports the mean ± standard deviation over the 25 samples. Following official metrics, DockQ aggregates interface-level DockQ across each oligomer, weighting interface i by $\log_{10}(\text{number of interacting residues}_i/2)$.

Model	DockQ	lDDT	TM-score	QSbest	ICS	IPS
AlphaFold3	0.5106 0.4783 ± 0.0226	0.8326 0.8306 ± 0.0027	0.8379 0.8016 ± 0.0171	0.6375 0.6424 ± 0.0157	0.5116 0.5126 ± 0.0097	0.5983 0.5882 ± 0.0073
ESMFold2	0.4500 0.4052 ± 0.0275	0.7830 0.7644 ± 0.0073	0.7521 0.6925 ± 0.0329	0.6051 0.5703 ± 0.0314	0.5024 0.4751 ± 0.0222	0.5606 0.5489 ± 0.0188
+SOUPFOLD	**0.4890** 0.4457 ± 0.0201	**0.8218** 0.7935 ± 0.0127	**0.8182** 0.7566 ± 0.0255	**0.6592** 0.6028 ± 0.0308	**0.5117** 0.4754 ± 0.0192	**0.5891** 0.5648 ± 0.0114
Protenix	0.4883 0.4426 ± 0.0195	0.8155 0.8066 ± 0.0035	0.8261 0.7600 ± 0.0263	0.6553 0.6160 ± 0.0243	0.5151 0.4887 ± 0.0150	0.5681 0.5523 ± 0.0120
+SOUPFOLD	**0.5512** 0.4847 ± 0.0258	**0.8353** 0.8266 ± 0.0040	**0.8324** 0.8148 ± 0.0257	**0.6971** 0.6568 ± 0.0253	**0.5497** 0.5201 ± 0.0170	**0.6089** 0.5970 ± 0.0144
OpenDDE	0.4537 0.4468 ± 0.0120	0.6426 0.6438 ± 0.0024	0.6507 0.6581 ± 0.0228	0.5062 0.4992 ± 0.0134	0.4109 0.4071 ± 0.0095	0.5114 0.5102 ± 0.0109
+SOUPFOLD	**0.5304** 0.4750 ± 0.0202	**0.6673** 0.6585 ± 0.0024	**0.7207** 0.6873 ± 0.0184	**0.5688** 0.5288 ± 0.0187	**0.4647** 0.4316 ± 0.0169	**0.5362** 0.5252 ± 0.0106

D.4 ARK ANTIBODY-ANTIGEN BENCHMARKS

Table 10: **Performance on ARK-AB benchmark.** The benchmark comprises 162 recent, low-homology PDB complexes, decomposed into 398 antibody-antigen interfaces grouped into 155 interface clusters for which all four co-folding models produce a prediction. We report success at three DockQ quality tiers: acceptable (≥ 0.23), medium (≥ 0.49), and high (≥ 0.80). *Cluster success (default)* averages the per-cluster success rate over the clusters, while *Interface success* reports the raw success rate over all interfaces. For each metric, we report the best-confidence result (5 samples per seed, 5 seeds) and the mean $\pm$ standard deviation over the 25 samples.

	Cluster Success		Interface Success	
Model	Best-conf.	Mean $\pm$ Std	Best-conf.	Mean $\pm$ Std
DockQ > 0.23 (acceptable)				
AlphaFold3	47.66	42.27 $\pm$ 2.00	58.04	51.80 $\pm$ 2.05
Protenix	45.94	42.08 $\pm$ 1.66	57.54	53.29 $\pm$ 1.23
+SOUPFOLD	**56.08**	54.02 $\pm$ 3.37	**67.34**	64.17 $\pm$ 2.86
ESMFold2	48.68	39.91 $\pm$ 2.53	59.30	47.70 $\pm$ 2.17
+SOUPFOLD	**60.53**	52.38 $\pm$ 2.08	**69.85**	61.31 $\pm$ 1.58
OpenDDE	58.92	57.51 $\pm$ 1.60	68.59	64.91 $\pm$ 1.56
+SOUPFOLD	**64.35**	61.37 $\pm$ 1.75	**73.12**	70.24 $\pm$ 1.60
DockQ > 0.49 (medium)				
AlphaFold3	41.62	36.29 $\pm$ 1.72	53.77	47.67 $\pm$ 1.64
Protenix	37.85	34.03 $\pm$ 1.40	50.75	46.61 $\pm$ 1.18
+SOUPFOLD	**47.48**	46.60 $\pm$ 2.38	**60.80**	58.31 $\pm$ 2.18
ESMFold2	42.88	34.66 $\pm$ 2.27	54.52	43.24 $\pm$ 2.12
+SOUPFOLD	**53.54**	44.75 $\pm$ 1.91	**64.07**	55.25 $\pm$ 1.46
OpenDDE	55.73	53.84 $\pm$ 1.73	65.33	61.81 $\pm$ 1.42
+SOUPFOLD	**57.74**	56.34 $\pm$ 2.09	**67.84**	65.23 $\pm$ 1.62
DockQ > 0.80 (high)				
AlphaFold3	21.60	19.54 $\pm$ 1.49	34.92	30.58 $\pm$ 1.30
Protenix	19.97	16.42 $\pm$ 1.33	31.91	27.69 $\pm$ 1.20
+SOUPFOLD	**20.42**	19.89 $\pm$ 1.38	**34.92**	32.89 $\pm$ 1.08
ESMFold2	19.40	13.42 $\pm$ 1.81	32.66	23.65 $\pm$ 1.90
+SOUPFOLD	**20.40**	18.92 $\pm$ 1.64	**34.67**	32.14 $\pm$ 1.34
OpenDDE	**32.18**	30.15 $\pm$ 1.33	**43.47**	40.86 $\pm$ 1.09
+SOUPFOLD	30.41	29.28 $\pm$ 1.14	43.22	41.53 $\pm$ 0.77

D.5 LOW-SIMILARITY SUBSET OF RUNS N' POSES

Table 11: **Performance on low-similarity subset of Runs N' Poses.** We evaluate the $[0, 20)$ low-similarity subset of Runs N' Poses benchmarks. We use 90 ligands from 83 systems for which all four co-folding models produce a prediction. The first line in each cell reports the highest-confidence prediction among 25 samples (5 samples per seed, 5 seeds). The second line reports the mean $\pm$ standard deviation over the 25 samples.

Model	Success (%)	BiSyRMSD ↓	lDDT-PLI ↑
AlphaFold3	28.89 25.64 ± 1.63	11.27 11.14 ± 0.54	0.40 0.38 ± 0.01
ESMFold2	27.78 25.33 ± 1.47	12.20 12.31 ± 0.37	0.37 0.35 ± 0.01
+SOUPFOLD	**28.89** 25.06 ± 1.97	**10.91** 11.29 ± 0.48	**0.41** 0.39 ± 0.01
Protenix	21.11 21.78 ± 0.89	11.67 11.41 ± 0.31	0.38 0.39 ± 0.01
+SOUPFOLD	**25.56** 24.93 ± 1.71	**11.25** 11.62 ± 0.53	**0.40** 0.39 ± 0.01
OpenDDE	15.56 18.02 ± 2.22	10.42 10.17 ± 0.48	0.35 0.36 ± 0.02
+SOUPFOLD	**24.44** 24.58 ± 1.59	**9.86** **10.13** ± 0.41	**0.39** **0.39** ± 0.01

D.6 TRANSFER TRAINING

Table 12: **Training representation transfer maps.** Ablation on FoldBench antibody-antigen and protein-protein complexes using ESMFold2 as the anchor model. The default maps are trained once on a held-out set. We compare them with maps trained directly on representations from each FoldBench benchmark. Accept, Medium, and High correspond to DockQ ≥ 0.23, ≥ 0.49, and ≥ 0.80, respectively. The first line in each cell reports the best-confidence-of-25 result (5 samples per seed, 5 seeds), and the second line reports the mean $\pm$ standard deviation over the 25 samples.

		DockQ success rate (%)		
Benchmark	Training data for f	Accept	Medium	High
Ab-Ag	Held-out (default)	**62.94** 52.96 ± 2.47	**51.18** 43.58 ± 1.71	**24.12** 20.12 ± 1.70
	FoldBench Antibody-Antigen	61.18 51.76 ± 2.54	48.82 41.91 ± 1.68	23.53 19.08 ± 1.97
P-P	Held-out (default)	**78.10** 75.33 ± 0.94	**72.63** 70.22 ± 0.95	48.54 45.07 ± 1.79
	FoldBench Protein-Protein	**78.10** 75.28 ± 1.00	**72.63** 69.71 ± 0.88	**49.27** 45.20 ± 1.47