COMPOSING CO-FOLDING MODELS WITH FEYNMAN–KAC CORRECTORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Feynman–Kac correctors (FKCs) compose pretrained diffusion models by simulating the sum of their scores and correcting the samples with importance weights and sequential Monte Carlo resampling. Diffusion-based all-atom structure predictors are natural candidates for such composition: no single predictor is reliably best, and pooling their outputs helps sampling more than selection because confidence scores are not comparable across models. We introduce FKCFold, which adapts FKCs to independently trained co-folding models that differ in architecture and tokenization but share the same all-atom denoiser parameterization. After each model encodes the complex with its own trunk, a single sampler evolves a population of candidate structures: at every step, all models denoise the same noisy structure, the sampler follows their normalized product of experts, and Feynman–Kac weights penalize disagreement between models. A common all-atom coordinate space, per-step rigid alignment of each prediction, native noise ranges for each model, and a correction adapted to the shared AlphaFold 3-style sampler make this composition well posed. FKCFold requires no training. Across multiple benchmarks, FKCFold combining Protenix-v2, OpenDDE and ESMFold2 consistently matches or outperforms each of the models it combines, with the largest gains on antibody–antigen complexes.

1 INTRODUCTION

Diffusion-based co-folding models predict the all-atom structures of biomolecular complexes. Following AlphaFold 3 (Abramson et al., 2024), open models including Protenix-v2, OpenDDE, and ESMFold2 have adopted this approach (Zhang et al., 2026; project, 2026; Candido et al., 2026). These models exhibit complementary strengths: a complex predicted inaccurately by one model may be resolved by another. Combining their predictions therefore offers a way to improve structure prediction without retraining the underlying models.

One approach is to run the models independently and pool their predictions. Pooling improves the coverage of accurate candidates, but native confidence scores are not calibrated across models, limiting the gains realized by candidate selection (Bae et al., 2026). The gap is widest for antibody–antigen complexes (Fromm et al., 2026; Solanki et al., 2026). Because pooling combines models only after generation, it leaves their complementary predictions unused during denoising. This motivates allowing several models to refine a shared population of candidate structures throughout generation.

Feynman–Kac correctors (FKCs) (Skreta et al., 2025a) provide a framework for such composition. They combine diffusion scores to define a product of experts and use importance weights and resampling to correct its sampling dynamics. Applying this framework to co-folding requires reconciling different atom representations, arbitrary rigid coordinate frames, and model-specific noise ranges. Moreover, the continuous-time derivation must be adapted to the discrete sampler used by AlphaFold 3-style models.

We introduce FKCFold, a training-free method for coupling heterogeneous all-atom denoisers using the FKC product-of-experts formulation. Each model first computes its own conditioning features. During sampling, the active models denoise the same candidate structures in a common atom space, and their rigidly aligned predictions are averaged to determine the next update. The FKC potential

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.

can then be expressed as the weighted disagreement among these denoised structures, requiring no additional network evaluations.

Our contributions are threefold. **(i)** We develop a shared sampling procedure that reconciles atom layouts, coordinate frames, and noise ranges while retaining each model's native conditioning and confidence modules. **(ii)** We express the FKC potential in denoiser space and distinguish its exact continuous-time interpretation from the approximations required in co-folding. **(iii)** We evaluate the composition of Protenix-v2, OpenDDE, and ESMFold2 on FoldBench, 2026ARK-AB, and PXM-2025 using a common DockQ implementation, matched candidate counts, and analyses that separate candidate generation from selection.

2 Background and related work

Diffusion in coordinate space. Let $x \in \mathbb{R}^{N\times 3}$ denote the coordinates of N atoms, and let σ denote the standard deviation of Gaussian noise added to a clean structure x_0. The resulting density, q_σ, defines a variance-exploding diffusion process (Song et al., 2020; Karras et al., 2022). A denoiser $D(x, \sigma)$ estimates the clean structure from the noisy coordinates. Its associated score is

$$s(x, \sigma) = \frac{D(x, \sigma) - x}{\sigma^2}. \tag{1}$$

For an exact conditional-mean denoiser, $s = \nabla_x \log q_\sigma$; thus the score points toward regions of higher probability at the current noise level. Sampling progressively reduces σ, using the denoiser to update the coordinates.

All-atom co-folding. AlphaFold 3 (Abramson et al., 2024) separates structure prediction into a conditioning network, or *trunk*, and an all-atom diffusion module. The trunk encodes sequence, multiple-sequence alignment (MSA), and template features into representations that remain fixed during a sampling trajectory. The diffusion module repeatedly denoises atomic coordinates using these representations, and a confidence module scores the final structures. Protenix, Boltz, Chai, and subsequent models follow this general design with different tokenizers, architectures, and training data (ByteDance AML AI4Science Team et al., 2025; Wohlwend et al., 2025; Passaro et al., 2025; Chai Discovery team et al., 2024; Zhang et al., 2026; project, 2026; Candido et al., 2026). For composition, each model can therefore be represented by a denoiser $D_e(x, \sigma; c_e)$, where e indexes the model and c_e denotes its fixed conditioning. This interface allows models to share coordinates while retaining their internal representations.

Composing diffusion models. For E models with noised densities $q_{e,\sigma}$, a normalized product of experts is

$$p_\sigma(x) = \frac{1}{Z_\sigma} \prod_{e=1}^{E} q_{e,\sigma}(x)^{\beta_e}. \tag{2}$$

Here Z_σ normalizes the density, and the nonnegative weights β_e sum to one and control each model's contribution. The product favors structures supported jointly by the models, and its score is the weighted mean of their scores. Score composition has been used to combine concepts in diffusion models (Liu et al., 2022). However, p_σ is generally not the result of adding Gaussian noise to p_0, so following the combined score alone does not recover the desired product distribution (Du et al., 2023).

MCMC correction (Du et al., 2023), density-based superposition (Skreta et al., 2025b), and Feynman–Kac correctors (Skreta et al., 2025a) address this mismatch. FKC augments the combined-score dynamics with importance weights and sequential Monte Carlo resampling (Doucet et al., 2009); under exact scores and continuous-time dynamics and the corresponding initialization, the ideal weighted law follows the product distribution. Related methods use sequential Monte Carlo to condition protein diffusion models (Trippe et al., 2022; Wu et al., 2023). Our method applies this framework to heterogeneous co-folding models. It also complements inference-time scaling approaches (Singhal et al., 2025; Ma et al., 2025a): drawing more candidates can improve oracle coverage, while confidence-based selection captures only part of that improvement (project, 2026; Bae et al., 2026).

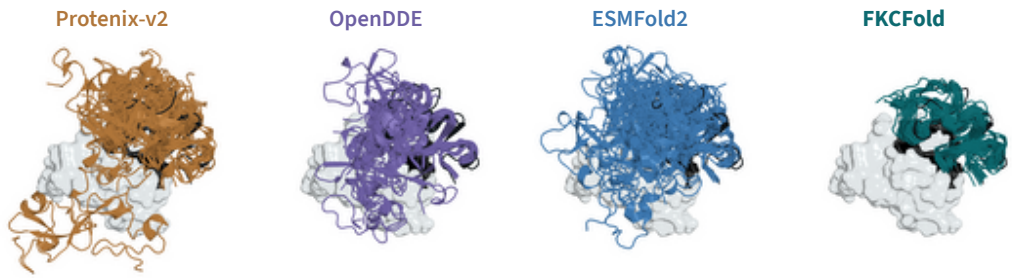

Figure 1: Prediction ensembles showing all candidates per method aligned on the reference receptor(grey); the native docked chain is black.

Algorithm 1 FKCFold

Require: denoisers $\{D_e, c_e, \sigma_e^{\max}\}_{e=1}^{E}$, weights β, schedule $\{\sigma_i, \gamma_i\}_{i=0}^{T}$, κ, α, particles K, FK window $[\sigma_{\min}^{\mathrm{FK}}, \sigma_{\max}^{\mathrm{FK}}]$

$x^k \sim \mathcal{N}(0, \sigma_0^2 I)$, $\ell^k \leftarrow 0$ ▷ $k = 1, \dots, K$
for $i = 0, \dots, T-1$ **do**
  $x^k \leftarrow R^k(x^k - \mathrm{mean}(x^k)) + t^k$ ▷ $R^k \sim \mathcal{U}(SO(3))$, $t^k \sim \mathcal{N}(0, I)$
  $\hat{x}^k \leftarrow x^k + \kappa\sqrt{\hat{\sigma}_i^2 - \sigma_i^2}\,\epsilon^k$ ▷ $\epsilon^k \sim \mathcal{N}(0, I)$, equation 6
  $\widetilde{D}_e^k \leftarrow \mathrm{Align}\big(D_e(\hat{x}_e^k, \hat{\sigma}_i; c_e), \hat{x}_e^k\big)$ for e with $\sigma_i \le \sigma_e^{\max}$ ▷ equation 3
  $\bar{D}_n^k \leftarrow \sum_e b_{e,n}(\sigma_i)\,\widetilde{D}_{e,n}^k$ ▷ equation 5
  $x^k \leftarrow \hat{x}^k + \eta_i(\bar{D}^k - \hat{x}^k)$ ▷ equation 6
  if $\sigma_{\min}^{\mathrm{FK}} \le \sigma_i \le \sigma_{\max}^{\mathrm{FK}}$ **then** ▷ optional FK correction
    $\ell^k \leftarrow \ell^k + \Delta \log w_{k,i}$ ▷ equation 8
    if $\mathrm{ESS}(e^{\ell}) < K/2$ **then**
      resample $\{x^k\}$ with probabilities $\propto e^{\ell^k}$; $\ell^k \leftarrow 0$
    end if
  end if
end for
return $\{x^k\}$, scored by each model's confidence module

3 METHODS

FKCFold couples E frozen co-folding models through shared-denoiser updates motivated by the product of experts in equation 2, using K particles, each a candidate structure (Algorithm 1). Each model e computes its conditioning c_e once; only the diffusion modules are queried during sampling.

3.1 ATOM MAPPING AND RIGID ALIGNMENT

Atoms are matched across models by entity, residue position, and atom name; their union gives a layout of N atoms. Let $m_{e,n} \in \{0, 1\}$ indicate that model e represents atom n, and x_e the coordinates of those atoms. Denoisers may return any rigid motion of the same structure, so each prediction is aligned to its input (Kabsch, 1976):

$$\widetilde{D}_e = R^\star D_e + t^\star, \qquad (R^\star, t^\star) = \underset{R \in SO(3),\, t}{\arg\min} \; \| R\, D_e(x_e, \sigma; c_e) + t - x_e \|_F^2. \tag{3}$$

3.2 THE SHARED DENOISING UPDATE

Writing each score as in equation 1, $s_e = (D_e - x)/\sigma^2$, with $\sum_e \beta_e = 1$, the product score is the score of one denoiser, the product-of-experts (PoE) denoiser $\bar{D}$:

$$\sum_e \beta_e s_e = \frac{\bar{D} - x}{\sigma^2}, \qquad \bar{D} = \sum_e \beta_e D_e. \tag{4}$$

In practice we combine aligned predictions atom by atom,

$$\bar{D}_n = \sum_e b_{e,n}\, \widetilde{D}_{e,n}, \qquad b_{e,n} = \frac{\beta_e m_{e,n} a_e(\sigma)}{\sum_j \beta_j m_{j,n} a_j(\sigma)}, \qquad a_e(\sigma) = \mathbf{1}[\sigma \le \sigma_e^{\max}], \tag{5}$$

renormalizing β_e over the active models that represent atom n. Activation is evaluated at σ_i, while denoisers are queried at the raised noise level $\hat{\sigma}_i$ defined below. With full layouts, all models active, and no alignment, equation 5 reduces to equation 4. Atoms with no active model keep $\bar{D}_n = x_n$ (zero drift).

We use the shared denoiser $\bar{D}$ in an AlphaFold 3-style sampler. At step i, the sampler raises the noise level to $\hat{\sigma}_i = (1+\gamma_i)\sigma_i$, then takes a scaled Euler step toward $\bar{D}$:

$$\hat{x}_i = x_i + \kappa\sqrt{\hat{\sigma}_i^2 - \sigma_i^2}\,\epsilon_i, \qquad x_{i+1} = \hat{x}_i + \eta_i\big(\bar{D}(\hat{x}_i, \hat{\sigma}_i) - \hat{x}_i\big), \qquad \eta_i = \alpha\Big(1 - \frac{\sigma_{i+1}}{\hat{\sigma}_i}\Big), \tag{6}$$

with $\epsilon_i \sim \mathcal{N}(0, I)$. In the ideal setting of equation 4, the denoising update with $\alpha = 1$ is an Euler step of $dx/d\sigma = -\sigma\bar{s}$, $\bar{s} = \sum_e \beta_e s_e$, from $\hat{\sigma}_i$ to σ_{i+1}. Each step costs one diffusion-module call per active model. Settings are in Appendix B.

3.3 Feynman–Kac correction in denoiser space

Following the PoE score alone does not generally sample the product in equation 2. In the ideal continuous-time construction, FKC corrects this mismatch with importance weights (Skreta et al., 2025a, Proposition D.5). For a variance-exploding process, the weight depends only on the spread of the denoised predictions (Appendix A):

$$d\log w = -\frac{V}{2\sigma^4}\, d\nu, \qquad V = \sum_e \beta_e \|D_e - \bar{D}\|_F^2, \tag{7}$$

where $d\nu > 0$ is the decrease in noise variance $\nu = \sigma^2$. Particles on which the models disagree lose weight, and V reuses the denoiser outputs. Per step, with the atomwise weights of equation 5,

$$\Delta \log w_{k,i} = -\frac{\sigma_i^2 - \sigma_{i+1}^2}{2\hat{\sigma}_i^4}\, V_{k,i}, \qquad V_{k,i} = \sum_{e,n} b_{e,n} \|\widetilde{D}_{e,n,k} - \bar{D}_{n,k}\|^2. \tag{8}$$

FKCFold combines the shared-denoiser update in equation 6 with an optional Feynman–Kac correction in equation 8. When enabled, the correction resamples particles when $\mathrm{ESS} = (\sum_k w_k)^2 / \sum_k w_k^2 < K/2$ and resets their weights. Sampler settings are reported in Appendix B.

Scope. FKC assumes exact scores, fixed β, and continuous-time reverse SDE dynamics. Alignment, atomwise renormalization, activation gates, and the discrete AlphaFold 3 sampler break these assumptions, so equation 8 is a practical correction (Appendix A).

4 Experimental setup

Benchmarks. We evaluate protein–protein (PP) and antibody–antigen (Ab–Ag) interfaces from FoldBench (Xu et al., 2025) and Ab–Ag interfaces from 2026ARK-AB (project, 2026) and from the antibody subset of PXM-2025 (Ma et al., 2025b). All methods are evaluated on the same interfaces, with coverage and exclusions documented in Appendix C. Although 2026ARK-AB and PXM-2025 contain recent structures, their release dates alone does not establish independence from model training data.

Models and sampling budget. We compare FKCFold with its component models, Protenix-v2, OpenDDE, and ESMFold2, using the OpenDDE-Ab checkpoint for antibody complexes. Each method generates 25 candidates from five random seeds, with five samples per seed. Each model retains its native featurization, and standalone and coupled runs use the same model-specific inputs. The *best of all models* baseline samples 25 candidates uniformly without replacement from the 75 standalone predictions and selects the candidate with the highest native confidence. We report the exact expected performance of this sampling rule (Appendix D). These comparisons match candidate counts; computational cost is assessed separately (§5.4).

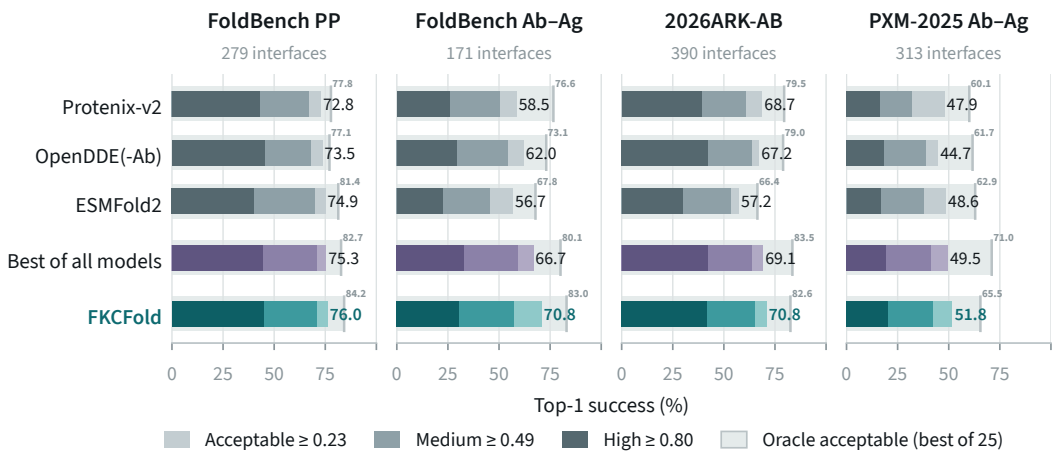

Figure 2: Top-1 success and oracle coverage at 25 candidates (PP, protein–protein; Ab–Ag, antibody–antigen). Nested bars show acceptable, medium, and high DockQ success, with the top-1 acceptable rate printed; the pale extension shows acceptable oracle coverage (best of 25), printed in grey. FKCFold uses shared-denoiser updates and the confidence-plus-density selector.

Candidate selection. Each standalone model ranks candidates using its native confidence score. For FKCFold, we evaluate two selection rules: the mean of the component models' confidence scores, and a combined score that also rewards structural density. Density measures how closely a candidate resembles other predictions in the same set after rigid alignment. Let c_i denote candidate i's mean confidence and ρ_i its structural density. We convert both quantities to percentile ranks within the candidate set, denoted by $\mathrm{pctrank}$, and combine them as

$$r_i = \lambda \,\mathrm{pctrank}(c)_i + \mathrm{pctrank}(\rho)_i. \tag{9}$$

The coefficient λ controls the relative contribution of confidence. We use a common setting across all benchmarks and report both selection rules throughout. To isolate changes in candidate generation, we also use each component model's confidence module to rank both its own candidates and FKCFold's candidates. The density estimator and selection settings appear in Appendix B.

Evaluation metrics. All benchmarks use the same FoldBench OpenStructure DockQ pipeline (Basu & Wallner, 2016; Biasini et al., 2013). Top-1 success is the fraction of interfaces for which the selected candidate meets a given DockQ threshold. Oracle success is the fraction for which any candidate meets that threshold, measuring coverage independently of selection. We use the standard acceptable, medium, and high thresholds of 0.23, 0.49, and 0.80, respectively. Reference structures are used only for evaluation and do not enter candidate selection. We additionally report mean top-1 DockQ and cluster-averaged success, giving equal weight to each interface sequence cluster. Paired 95% confidence intervals are obtained by resampling complexes while keeping their interfaces together, without multiple-comparison adjustment. Candidate-count scaling averages over all subsets of the available seeds and recomputes selection within each subset. Further evaluation details appear in Appendices B and C.

5 Results and Discussion

5.1 Coupling improves acceptable top-1 success

Coupling improves acceptable top-1 prediction on every evaluated panel. With the combined selector, FKCFold exceeds the strongest individual model on all four panels, and by the widest margin on FoldBench Ab–Ag (70.8% against 62.0%), where it also leads every individual model at the medium and high tiers (Figure 2). Composition helps most on antibody–antigen interfaces, where individual models are less accurate, without costing accuracy on protein–protein interfaces.

Table 1: Fixed confidence module, different candidate sets: acceptable top-1 success (%) when each model's own confidence module ranks its own 25 candidates or FKCFold's 25, without density. Differences are coupled minus own (pp); CIs resample complexes (5,000 replicates).

Confidence module	Own (%)	Coupled (%)	Difference (pp)	95% CI
FoldBench: protein–protein				
Protenix-v2	72.76	75.63	+2.87	[1.09, 5.05]
OpenDDE	73.48	74.55	+1.08	[−1.11, 3.40]
ESMFold2	74.91	77.78	+2.87	[1.06, 4.94]
FoldBench: antibody–antigen				
Protenix-v2	58.48	64.33	+5.85	[−0.58, 12.72]
OpenDDE-Ab	61.99	65.50	+3.51	[0.57, 7.41]
ESMFold2	56.73	67.25	+10.53	[4.05, 17.82]
2026ARK-AB: antibody–antigen				
Protenix-v2	68.72	70.77	+2.05	[−2.33, 6.53]
OpenDDE-Ab	67.18	66.92	−0.26	[−2.47, 2.15]
ESMFold2	57.18	67.69	+10.51	[5.47, 16.11]
PXM-2025: antibody–antigen				
Protenix-v2	47.92	49.84	+1.92	[−4.47, 7.79]
OpenDDE-Ab	44.73	45.69	+0.96	[−3.07, 4.79]
ESMFold2	48.56	50.48	+1.92	[−3.17, 6.69]

The improvement persists with confidence-only selection, while density changes its magnitude in a benchmark-dependent way. Confidence alone already exceeds the strongest individual model on every panel (Appendix E); density raises top-1 success on FoldBench Ab–Ag and 2026ARK-AB and lowers it on PP and PXM-2025 Ab–Ag. On FoldBench and 2026ARK-AB, no paired difference between the selection rules excludes zero; on PXM-2025 Ab–Ag, confidence alone exceeds the combined rule by 4.8 points [1.7, 8.6] (Appendix I).

Pooling the three models' outputs is a strong baseline. Best of all models also exceeds every individual model, and FKCFold's advantage over it is clearest at the acceptable threshold; with the combined selector the paired intervals include zero (Table 9), and at stricter thresholds the pool is competitive or ahead on some panels. The clearest gain over pooling is on PXM-2025 Ab–Ag with confidence-only selection, +7.0 points [+2.6, +12.3]. Distinguishing what coupling adds beyond pooling requires separating candidate generation from selection, which the next sections do.

5.2 Separating generation and selection effects

Holding the confidence module fixed tests whether coupling produces candidates that existing rankers can use more effectively. Each model's own confidence module selects better predictions from FKCFold's candidates than from its own: all six FoldBench comparisons favor coupled candidates, and four paired intervals exclude zero (Table 1). This supports a contribution from candidate generation to the overall improvement. The effect is weaker on 2026ARK-AB and PXM-2025 Ab–Ag, where most intervals include zero.

Oracle coverage points the same way: FKCFold's 25 candidates contain an acceptable interface more often than any individual model's on all four panels. Against the 75-prediction pool subsampled to 25, coverage depends on the benchmark and the quality threshold (Appendix E). Together, these controls show why generation and selection should be evaluated separately: coupling can increase the availability of acceptable structures, while the ranking rule determines how much of that improvement appears in the selected prediction.

5.3 Scaling with candidate count

More candidates create more opportunities, but selection captures only part of them. Oracle coverage rises steadily with the number of seeds on every panel, whereas top-1 success saturates earlier (Figure 3). On FoldBench Ab–Ag, FKCFold's top-1 success keeps improving up to 25 candidates; on 2026ARK-AB and PXM-2025 Ab–Ag it levels off after 10 to 15. At 25 candidates, oracle cov-

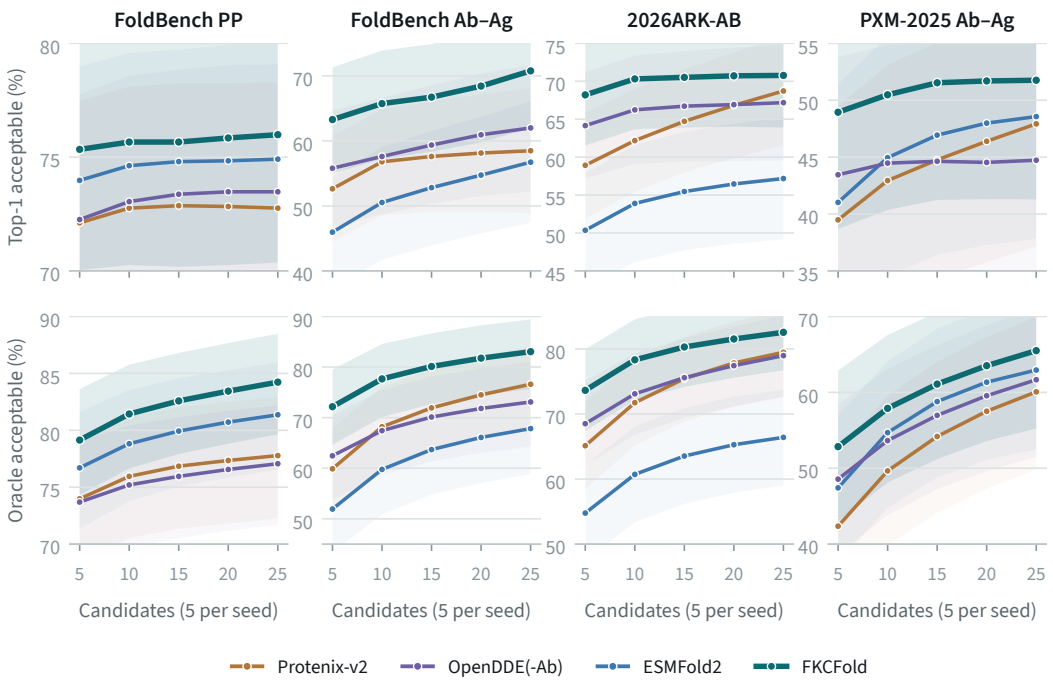

Figure 3: Scaling from one to five seeds (five candidates per seed). Points average all seed subsets; bands are 95% intervals from resampling complexes. FKCFold uses the combined selector, recomputed per subset. Vertical scales differ by panel and bands may extend beyond them. Curves compare candidate counts, not wall-clock budgets.

erage still exceeds top-1 success by 12 to 14 points on the three Ab–Ag panels. The remaining gap identifies substantial room to improve selection within the generated candidate sets.

5.4 Inference time and hardware requirements

Shared inference is practical within the profiled setup. Each model's trunk runs once per seed and its conditioning is reused along the shared trajectory, so one seed of five coupled candidates takes 248.9 s on one H100 GPU, compared with 271.1 s for 15 candidates from three standalone runs (Appendix Table 12; timings include loading, featurization, generation, and confidence evaluation). All three models fit on one 80 GB GPU for most complexes, with idle models moved to host memory during featurization; the 33 FoldBench complexes that need four GPUs are the same ones for which standalone OpenDDE needs four.

5.5 Model contributions and independent pooling

A pilot distinguishes improved candidate coverage from improved selected predictions. It uses 16 FoldBench Ab–Ag complexes sampled by size before any outcome was inspected, with 10 candidates per arm (Appendix I). Against independent sampling of the same three models, coupled candidates contain an acceptable pose for more interfaces than an equal number of pooled candidates, and the paired interval excludes zero (Table 2). Under common rescoring, the comparison does not establish a top-1 advantage: coupling enlarges the set of usable candidates, and the selector does not yet convert that gain.

Model removal on the full one-GPU cohort shows that the three models contribute unequally (Table 3). On Ab–Ag, removing Protenix-v2 lowers top-1 success, mean DockQ, and oracle coverage, the last two with intervals excluding zero; removing OpenDDE-Ab or ESMFold2 leaves Ab–Ag top-1 success essentially unchanged. Removing ESMFold2 slightly lowers PP mean DockQ. The 16-complex pilot suggested a larger ESMFold2 contribution on Ab–Ag (Appendix Table 16), which

Table 2: Coupling versus independent pooling on the pilot, ranked by *common rescoring* with the Protenix-v2 confidence module. Differences are coupled minus independent; brackets give 95% CIs over complexes.

Candidates	Oracle acc. (%)	Top-1 acc. (%)	Top-1 mean DockQ
Candidate sets			
Coupled, 10	87.5	50.0	0.438
Independent, 10 of 30 (expected)	77.5	51.1	0.435
Independent, all 30	83.3	58.3	0.504
Coupled minus independent (pp; DockQ)			
vs. 10 of 30 (expected)	$+10.0$	-1.1	$+0.004$
	$[0.2, 21.9]$	$[-11.0, 9.3]$	$[-0.074, 0.068]$
vs. all 30	$+4.2$	-8.3	-0.065
	$[0.0, 14.3]$	$[-26.1, 0.0]$	$[-0.231, 0.041]$

Table 3: Model removal on FoldBench complexes: the full three-model FKCFold run against the reduced two-model run. Top-1 uses a confidence module present in both arms (Protenix-v2, or ESMFold2 when Protenix-v2 is removed); brackets give 95% CIs over complexes.

		Top-1 acc. (%)	Reduced minus full		
Removed model	Panel	Full $\rightarrow$ reduced	Acc. (pp)	Mean DockQ	Oracle acc. (pp)
ESMFold2	PP	$74.5 \rightarrow 72.5$	-2.0	-0.014	-1.6
			$[-4.2, 0.0]$	$[-0.027, -0.002]$	$[-4.0, 0.4]$
ESMFold2	Ab–Ag	$66.0 \rightarrow 66.0$	$+0.0$	$+0.003$	-0.6
			$[-6.0, 6.1]$	$[-0.030, 0.039]$	$[-4.6, 2.6]$
OpenDDE-Ab	Ab–Ag	$66.0 \rightarrow 67.3$	$+1.3$	-0.010	-3.1
			$[-5.0, 8.2]$	$[-0.034, 0.014]$	$[-9.5, 2.6]$
Protenix-v2	Ab–Ag	$67.9 \rightarrow 61.6$	-6.3	-0.039	-5.0
			$[-13.0, 0.0]$	$[-0.077, -0.008]$	$[-9.9, -0.7]$

the full cohort does not reproduce. The contributions of OpenDDE-Ab and ESMFold2 are therefore small at this sample size, and each row uses a confidence module present in both of its arms, so rows are not directly comparable in magnitude.

5.6 Sensitivity to alignment and noise caps

These experiments characterize how sensitive the sampler is to its design choices. Alignment's clearest measured contribution is preserving PP oracle coverage: without per-step alignment of each model's prediction, fewer PP interfaces have an acceptable candidate, while top-1 changes are small on both panels. The noise cap is an activation rule that queries ESMFold2 only below a maximum noise level. Removing the cap changes acceptable top-1 success by less than one point on either panel in this ablation, with intervals that include zero.

5.7 Limitations and future directions

The findings above point to concrete next steps. The oracle–top-1 gap, visible on every panel and widest on Ab–Ag, motivates better ranking, for example confidence calibrated across models or learned rankers. The benchmark-dependent effect of density, together with selection rules developed on FoldBench, motivates selectors that transfer across benchmarks; the stored held-out partition (Appendix F) is a first check, and larger held-out benchmarks with explicit control of training overlap (Škrinjar et al., 2025) would sharpen it. The model-removal results motivate studying which models complement one another, including further co-folding models such as Boltz (Wohlwend et al., 2025; Passaro et al., 2025) and Chai-1 (Chai Discovery team et al., 2024), and other molecule types. Finally, alignment and noise-dependent activation make FKCFold an approximation of the exact product of experts; closing this gap, and comparing methods at matched compute as well as matched candidate counts, remain open.

6 CONCLUSION

FKCFold enables independently trained co-folding models to refine a shared population of structures without additional training. By reconciling atom layouts, coordinate frames, and noise ranges, it combines all-atom denoisers within a common sampler. The associated Feynman–Kac potential reduces to disagreement among the denoised structures, providing an interpretable correction that reuses the same model evaluations. Across FoldBench, 2026ARK-AB, and PXM-2025, the coupled sampler improves acceptable top-1 success over the individual models, with the largest gains on FoldBench antibody–antigen interfaces. Fixed-confidence and oracle analyses support a contribution from candidate generation. On the component pilot, coupled generation improves acceptable oracle coverage over candidate-count-matched independent pooling. These results support shared denoising as a practical form of model composition and motivate further work on reweighting and candidate selection.

ETHICS STATEMENT

This work uses only publicly released structures and model weights: FoldBench, 2026ARK-AB, and PXM-2025 are derived from the Protein Data Bank, and Protenix-v2, OpenDDE, and ESMFold2 are public checkpoints. No human subjects, personal data, or new biological materials are involved. FKCFold predicts the structures of given complexes and does not generate new sequences, so it does not add a design capability beyond that of the models it combines. Improved antibody–antigen structure prediction is dual use in principle, as is protein structure prediction generally; we see no risk specific to this method.

REPRODUCIBILITY STATEMENT

Appendices B–D specify the sampler, evaluation coverage, selection rules, and exact independent-pool calculation. Appendices H and I give the protocols of the ablations, including cohorts, seeds, and rerun variability. All methods are scored with the same OpenStructure DockQ implementation and thresholds. Code will be released upon acceptance.

LLM USAGE STATEMENT

We did not use generative AI tools for methodology design, hypothesis development, or synthetic data generation.The authors reviewed all AI-assisted work and take full responsibility for the final content of this paper, including text, claims, code, and results produced with the aid of generative AI.

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.

Junhyeok Bae, Soohyun Jo, Yeajin Kim, Dongyun Kim, Kwanwook Kim, Sanghyun Park, Sanggeun Park, Sojung Myung, Hyunho Shin, Min Hyeok Kim, et al. Thal-kak: unifying biomolecular structure predictors reveals a sampling-selection gap. *bioRxiv*, pp. 2026–08, 2026.

Sankar Basu and Björn Wallner. Dockq: a quality measure for protein-protein docking models. *PloS one*, 11(8):e0161879, 2016.

Marco Biasini, Tobias Schmidt, Stefan Bienert, Valerio Mariani, Gabriel Studer, Jürgen Haas, Niklaus Johner, Andreas Daniel Schenk, Ansgar Philippsen, and Torsten Schwede. Openstructure: an integrated software framework for computational structural biology. *Biological crystallography*, 69(5):701–709, 2013.

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*, pp. 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*, pp. 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*, pp. 2024–10, 2024.

Arnaud Doucet, Adam M Johansen, et al. A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of nonlinear filtering*, 12(656-704):3, 2009.

Yilun Du, Conor Durkan, Robin Strudel, Joshua B Tenenbaum, Sander Dieleman, Rob Fergus, Jascha Sohl-Dickstein, Arnaud Doucet, and Will Sussman Grathwohl. Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and mcmc. In *International conference on machine learning*, pp. 8489–8510. PMLR, 2023.

Samuel Fromm, Marko Ludaic, and Arne Elofsson. Evaluating deep learning based structure prediction methods on antibody–antigen complexes. *Bioinformatics*, 42(4):btag136, 2026.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Foundations of Crystallography*, 32(5):922–923, 1976.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *European conference on computer vision*, pp. 423–439. Springer, 2022.

Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, et al. Inference-time scaling for diffusion models beyond scaling denoising steps. *arXiv preprint arXiv:2501.09732*, 2025a.

Wenzhi Ma, Zhenyu Liu, Jincai Yang, Chan Lu, Hanyu Zhang, and Wenzhi Xiao. From dataset curation to unified evaluation: Revisiting structure prediction benchmarks with PXMeter. *bioRxiv*, 2025b.

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.

Aureka AI OpenDDE project. Folding, reasoning, and scaling with open-source drug discovery engine, 2026. URL `https://arxiv.org/abs/2607.03787`.

Raghav Singhal, Zachary Horvitz, Ryan Teehan, Mengye Ren, Zhou Yu, Kathleen McKeown, and Rajesh Ranganath. A general framework for inference-time scaling and steering of diffusion models. *arXiv preprint arXiv:2501.06848*, 2025.

Marta Skreta, Tara Akhound-Sadegh, Viktor Ohanesian, Roberto Bondesan, Alán Aspuru-Guzik, Arnaud Doucet, Rob Brekelmans, Alexander Tong, and Kirill Neklyudov. Feynman-kac correctors in diffusion: Annealing, guidance, and product of experts. *arXiv preprint arXiv:2503.02819*, 2025a.

Marta Skreta, Lazar Atanackovic, Joey Bose, Alexander Tong, and Kirill Neklyudov. The superposition of diffusion models using the itô density estimator. In *International Conference on Learning Representations*, volume 2025, pp. 67004–67057, 2025b.

Peter Škrinjar, Jérôme Eberhardt, Janani Durairaj, and Torsten Schwede. Have protein-ligand cofolding methods moved beyond memorisation? *BioRxiv*, pp. 2025–02, 2025.

Arnav Solanki, Neha S Maurya, Meaghan Ramlakhan, Rongbin Li, Wenbo Chen, Zhuhao Wu, and Wenjin Jim Zheng. A structural antibody benchmark of alphafold3 reveals hallucinated epitopes and a bias for orderness. *bioRxiv*, pp. 2026–07, 2026.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Brian L Trippe, Jason Yim, Doug Tischer, David Baker, Tamara Broderick, Regina Barzilay, and Tommi Jaakkola. Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem. *arXiv preprint arXiv:2206.04119*, 2022.

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*, pp. 2024–11, 2025.

Luhuan Wu, Brian Trippe, Christian Naesseth, David Blei, and John P Cunningham. Practical and asymptotically exact conditional sampling in diffusion models. *Advances in Neural Information Processing Systems*, 36:31372–31403, 2023.

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.

Yuxuan Zhang, Chengyue Gong, Jinyuan Sun, Jiaqi Guan, Milong Ren, Song Xue, Hanyu Zhang, Wenzhi Ma, Zhenyu Liu, Xinshi Chen, et al. Protenix-v2: Broadening the reach of structure prediction and biomolecular design. *bioRxiv*, pp. 2026–04, 2026.

A Denoiser-space potential and its scope

Continuous-time correction. Let $\nu = \sigma^2$ denote noise variance, and let $d\nu > 0$ denote its decrement during reverse diffusion. For a score field s, the reverse stochastic differential equation (SDE) and the corresponding probability-flow ordinary differential equation (ODE) are

$$x \leftarrow x + s(x, \sigma)\, d\nu + \sqrt{d\nu}\, \epsilon, \qquad \frac{dx}{d\sigma} = -\sigma s(x, \sigma), \tag{10}$$

with $\epsilon \sim \mathcal{N}(0, I)$. With the exact score of a Gaussian-noised data distribution, both dynamics recover its marginals as noise decreases. The product density in equation 2 generally does not follow this same forward noising process.

For fixed model weights, write $\bar{s} = \sum_e \beta_e s_e$ for the product's score. Feynman–Kac correction augments the reverse SDE driven by $\bar{s}$ with

$$d \log w = \frac{d\nu}{2} \left(\|\bar{s}\|_F^2 - \sum_e \beta_e \|s_e\|_F^2 \right). \tag{11}$$

This potential corrects the marginal evolution in the ideal continuous-time construction (Skreta et al., 2025a, Proposition D.5), given the corresponding initial distribution.

Denoiser-space identity. For constant weights summing to one, let $u_e = D_e - x$ and $\bar{u} = \sum_e \beta_e u_e = \bar{D} - x$. Then

$$\begin{aligned} \Big\| \sum_e \beta_e s_e \Big\|_F^2 - \sum_e \beta_e \|s_e\|_F^2 &= \sigma^{-4} \left(\|\bar{u}\|_F^2 - \sum_e \beta_e \|u_e\|_F^2 \right) \\ &= -\sigma^{-4} \sum_e \beta_e \|u_e - \bar{u}\|_F^2 = -\sigma^{-4} V. \end{aligned} \tag{12}$$

The identity also holds for signed weights summing to one; nonnegative weights give $V \geq 0$. For two models with weights $(\beta, 1 - \beta)$,

$$V = \beta(1 - \beta) \|D_1 - D_2\|_F^2, \qquad d \log w = \frac{d\nu}{2} \beta(\beta - 1) \|s_1 - s_2\|_F^2, \tag{13}$$

recovering the guidance potential of Skreta et al. (2025a), including the signed-weight regime of classifier-free guidance (Ho & Salimans, 2022). FKCFold uses nonnegative equal base weights. If the weights do not sum to one, the shared x term does not cancel and equation 12 does not hold in this form.

Limits of the product interpretation. The continuous-time construction assumes exact, compatible scores, a fixed model set and weights, and the prescribed SDE; finite steps and particles add error. FKCFold departs further: Kabsch alignment transforms denoiser outputs, per-atom renormalization mixes score components with different coefficients, and noise gates change the active model set. The variance identity still holds numerically, but it need no longer be the potential of a single joint product density.

The AlphaFold 3-style transition adds another approximation. The net interval in equation 8 follows from

$$(\hat{\sigma}_i^2 - \sigma_{i+1}^2) - (\hat{\sigma}_i^2 - \sigma_i^2) = \sigma_i^2 - \sigma_{i+1}^2, \tag{14}$$

but this does not cancel transition-density errors: the denoiser changes along the path, and both the added noise and the Euler step are scaled. The update in equation 8 is therefore a practical discretization evaluated by ablation. The numerical verification supplied with the paper checks the variance identity; it does not establish the distribution sampled by the implementation.

B Implementation and reproducibility

B.1 Sampler settings

The schedule interpolates in a power-transformed noise scale:

$$\sigma_i = \sigma_{\text{data}} \left[s_{\max}^{1/r} + \frac{i}{T} \left(s_{\min}^{1/r} - s_{\max}^{1/r} \right) \right]^r, \quad 0 \leq i < T, \qquad \sigma_T = 0, \tag{15}$$

with $T = 200$, $\sigma_{\text{data}} = 16\,\text{Å}$, $s_{\max} = 160$, $s_{\min} = 4 \times 10^{-4}$, and $r = 7$ ($\sigma_0 = 2560\,\text{Å}$). The sampler equation 6 uses $\kappa = 1.003$, $\alpha = 1.5$, and $\gamma_i = 0.8 \cdot \mathbf{1}[\sigma_{i+1} > 1\,\text{Å}]$: each step raises the noise level by 80% before denoising, until the next level falls below $1\,\text{Å}$. Each step applies an independent uniform rotation and a $\mathcal{N}(0, 1\,\text{Å}^2)$ translation to every centered particle. Model weights are $\beta_e = 1/3$. Activation uses the level σ_i before this increase, and the models are queried at the raised level $\hat{\sigma}_i$. Only ESMFold2 has a cap, $\sigma^{\max} = 256\,\text{Å}$, so the first 67 steps use Protenix-v2 and OpenDDE only. Alignment equation 3 is unweighted over all atoms a model represents. Main runs use $K = 5$ particles and seeds 42, 66, 101, 2024, and 8888.

The correction equation 8 is applied for $4 \le \sigma_i \le 160\,\text{Å}$ (steps 78–143), inside the three-model range and while the sampler still raises the noise level ($\sigma_{i+1} > 1\,\text{Å}$). The increment follows the coordinate update, and resampling is systematic (Doucet et al., 2009), with no increment clamping and no terminal resampling; terminal weights do not enter selection. Both settings share seeds and one random-number generator, so they diverge after the first resampling event.

B.2 Model inputs and candidate selection

OpenDDE is used on PP and OpenDDE-Ab on Ab–Ag complexes, in both standalone and coupled inference. FoldBench uses ColabFold MSAs. On 2026ARK-AB, Protenix-v2 uses Protenix-server MSAs in both settings, while the other models use ColabFold inputs.

Native ranking scores keep their implementation-specific clash and disorder terms. The confidence selector is their unweighted mean. For M candidates, structural density is

$$\rho_i = \frac{1}{M-1} \sum_{j \neq i} \exp\left(-\frac{d_{ij}^2}{2h^2}\right). \tag{16}$$

Here d_{ij} is the RMSD between the protein Cα atoms of candidates i and j in matching order after rigid alignment, and h is the Gaussian kernel width. We use $h = 2\,\text{Å}$ and confidence coefficient $\lambda = 2$ in equation 9. Percentile ranks are computed within each candidate set, with average ranks for ties. Ties in the final score go to the lower seed, then the lower sample index. Seed-subset analyses recompute ranks and density within each subset.

B.3 Evaluation and reproducibility

All DockQ scores use OpenStructure 2.8.0. Cluster-averaged success uses the published 2026ARK clusters and FoldBench clusters defined by 40% sequence identity and 80% coverage. Paired intervals use 5,000 bootstrap resamples of complexes, preserving all interfaces within each sampled complex. Seed-scaling curves average over all subsets of one to five seeds.

The package contains `audit_results.py`, `make_tables.py`, and `make_figures.py`, a frozen per-interface score snapshot, the confidence values and similarity matrices needed for aggregation, and hashes of every structure and confidence file read. The audit recomputes selection and aggregation from saved outputs. As a spot check, OpenStructure was rerun on 24 prediction files (48 interface scores, including all three restored interfaces) and reproduced the stored DockQ values. Every retained interface has a complete 25-candidate DockQ array for every method.

C Benchmark coverage and exclusions

FoldBench contains 451 reference interfaces in the two evaluated classes. The PP panel keeps all 279 interfaces from 239 complexes. The Ab–Ag panel keeps 171 of 172, excluding `8xnh-assembly1` A/F, where coupled generation did not complete; the exclusion applies to every method.

2026ARK-AB has 404 interfaces from 164 entries. Six interfaces of `9j4c` and `9n8n` are omitted because ESMFold2 did not complete. Eight more have no mapped OpenStructure score for any of the four methods: `9lz0` A/C, `9pso` A/C, `9qrv` A/B, A/B.1, A/C, A.1/B, and `9vnp` A/B.1, A.1/B. This removes `9qrv` entirely, leaving 161 entries.

Three published interfaces, `9ly2` C/E, `9ly3` D/E, and `9lz2` D/E, have complete OpenStructure scores but are absent from the legacy PXMeter index. We include them based on the complete score

records, yielding 390 interfaces in 153 published clusters. The earlier 387-interface aggregation is retained in the evaluation audit for traceability.

PXM-2025 Ab–Ag is PXMeter's antibody subset of PXM-2025: the interfaces labelled antibody–protein in its low-homology table, 314 interfaces from 131 entries. All four methods have 25 scored candidates for every interface. One interface, `9lad` A/B, has no mapped OpenStructure score for any method and is omitted, leaving 313 interfaces from 130 entries in 94 clusters (pairs of PXM-2025 entity sequence clusters). Chain copies generated by PXMeter's symmetry expansion are mapped to the corresponding assembly chains before scoring.

All benchmarks use the same OpenStructure version and cumulative thresholds ($\geq$), with recorded mapping recoveries. Rates therefore need not match published PXMeter-based numbers, which use a different implementation and strict comparisons. All reported values come from the evaluation campaigns documented in the accompanying data snapshot.

D Exact evaluation of independent pooling

The pool samples $m = 25$ of $N = 75$ standalone candidates uniformly without replacement. With h candidates above a DockQ threshold, the expected oracle success is

$$1 - \frac{\binom{N-h}{m}}{\binom{N}{m}}, \tag{17}$$

with $\binom{a}{m} = 0$ for $a < m$. For top-1 selection, group candidates by equal native confidence. A group of g candidates with b candidates of higher confidence supplies the selection with probability

$$\frac{\binom{N-b}{m} - \binom{N-b-g}{m}}{\binom{N}{m}}, \tag{18}$$

and ties within a group are equally likely. Weighting each group's mean success indicator, or mean DockQ, by these probabilities gives the exact expected top-1 metric. This replaces the earlier Monte Carlo estimate based on 2,000 draws and removes simulation error. This is a sampling baseline, not an optimal cross-model selector.

E Full benchmark results

Table 4: FoldBench protein–protein: top-1 and oracle success (%) and top-1 mean DockQ on 279 interfaces from 239 complexes, 25 candidates per method. Acc., Med., and High: DockQ ≥ 0.23, ≥ 0.49, and ≥ 0.80. Best of all models is the exact 25-of-75 expectation.

	Top-1 success (%)				Oracle success (%)		
Method	Acc.	Med.	High	Mean DockQ	Acc.	Med.	High
Protenix-v2	72.8	67.0	43.0	0.580	77.8	71.3	50.5
OpenDDE	73.5	67.7	45.2	0.594	77.1	71.0	54.5
ESMFold2	74.9	69.5	40.1	0.596	81.4	74.2	55.6
Best of all models (25 of 75)	75.3	70.8	44.4	0.609	82.7	77.7	59.2
FKCFold, confidence	76.7	71.7	45.5	0.620	84.2	77.8	56.6
FKCFold, conf. + density	76.0	70.6	44.8	0.612	84.2	77.8	56.6

Tables 4–7 give all three DockQ tiers for selected and oracle predictions, Table 8 cluster-averaged top-1 success, and Table 9 paired intervals against best of all models (whose selector differs). Figure 4 extends the comparison over the full DockQ range. Bootstrap intervals cover sampling of complexes only, not training overlap or hyperparameter selection.

Complexes whose coupled run needs four GPUs shard OpenDDE across processes (§5.4). In the saved FoldBench and 2026ARK-AB runs, the sampler state was not synchronized across these processes at every step, so small numerical differences between their copies of the other models could accumulate; the PXM-2025 runs synchronize it. Table 10 therefore repeats the main comparison on

Table 5: FoldBench antibody–antigen: top-1 and oracle success (%) and top-1 mean DockQ on 171 interfaces from 112 complexes, 25 candidates per method. Acc., Med., and High: DockQ ≥ 0.23, ≥ 0.49, and ≥ 0.80. Best of all models is the exact 25-of-75 expectation.

	Top-1 success (%)				Oracle success (%)		
Method	Acc.	Med.	High	Mean DockQ	Acc.	Med.	High
Protenix-v2	58.5	50.3	25.7	0.445	76.6	61.4	38.0
OpenDDE-Ab	62.0	54.4	29.2	0.479	73.1	62.6	41.5
ESMFold2	56.7	45.6	22.2	0.413	67.8	51.5	33.3
Best of all models (25 of 75)	66.7	59.1	32.8	0.519	80.1	69.3	47.4
FKCFold, confidence	66.1	54.4	28.7	0.490	83.0	70.8	43.3
FKCFold, conf. + density	70.8	57.3	30.4	0.522	83.0	70.8	43.3

Table 6: 2026ARK-AB: top-1 and oracle success (%) and top-1 mean DockQ on 390 interfaces from 161 complexes, 25 candidates per method. Acc., Med., and High: DockQ ≥ 0.23, ≥ 0.49, and ≥ 0.80. Best of all models is the exact 25-of-75 expectation.

	Top-1 success (%)				Oracle success (%)		
Method	Acc.	Med.	High	Mean DockQ	Acc.	Med.	High
Protenix-v2	68.7	60.5	39.0	0.537	79.5	68.7	46.7
OpenDDE-Ab	67.2	63.8	42.3	0.552	79.0	69.7	49.0
ESMFold2	57.2	53.6	29.7	0.455	66.4	58.2	40.3
Best of all models (25 of 75)	69.1	63.8	42.1	0.560	83.5	74.9	54.5
FKCFold, confidence	70.5	65.4	42.1	0.566	82.6	74.4	54.1
FKCFold, conf. + density	70.8	65.1	41.8	0.569	82.6	74.4	54.1

Table 7: PXM-2025 antibody–antigen: top-1 and oracle success (%) and top-1 mean DockQ on 313 interfaces from 130 complexes, 25 candidates per method. Acc., Med., and High: DockQ ≥ 0.23, ≥ 0.49, and ≥ 0.80. Best of all models is the exact 25-of-75 expectation.

	Top-1 success (%)				Oracle success (%)		
Method	Acc.	Med.	High	Mean DockQ	Acc.	Med.	High
Protenix-v2	47.9	31.9	16.3	0.329	60.1	38.7	23.0
OpenDDE-Ab	44.7	38.7	18.5	0.334	61.7	45.7	24.9
ESMFold2	48.6	37.7	16.6	0.351	62.9	46.3	25.6
Best of all models (25 of 75)	49.5	41.4	19.1	0.368	71.0	53.6	30.9
FKCFold, confidence	56.5	45.0	21.1	0.396	65.5	51.8	28.1
FKCFold, conf. + density	51.8	42.2	20.1	0.374	65.5	51.8	28.1

Table 8: Cluster-averaged top-1 success, with each interface sequence cluster weighted equally. FoldBench clusters follow the 2026ARK recipe; 2026ARK-AB and PXM-2025 use their published cluster assignments.

Method	Cluster-averaged top-1 success (%)		
	Acceptable	Medium	High
FoldBench: protein–protein (265 clusters)			
Protenix-v2	71.7	65.8	42.1
OpenDDE	72.3	66.4	44.3
ESMFold2	73.8	68.3	39.1
Best of all models (25 of 75)	74.4	69.6	43.6
FKCFold, confidence	75.8	70.6	45.1
FKCFold, conf. + density	75.1	69.4	43.6
FoldBench: antibody–antigen (166 clusters)			
Protenix-v2	58.4	50.3	26.5
OpenDDE-Ab	61.4	53.6	28.9
ESMFold2	57.8	46.4	22.9
Best of all models (25 of 75)	66.1	58.4	32.6
FKCFold, confidence	66.3	54.8	29.5
FKCFold, conf. + density	71.1	57.2	30.7
2026ARK-AB (153 clusters)			
Protenix-v2	59.7	47.4	25.3
OpenDDE-Ab	58.6	53.4	29.5
ESMFold2	46.7	41.2	17.0
Best of all models (25 of 75)	60.0	53.1	29.8
FKCFold, confidence	63.2	56.9	30.3
FKCFold, conf. + density	63.4	55.6	30.1
PXM-2025: antibody–antigen (94 clusters)			
Protenix-v2	51.5	38.3	18.4
OpenDDE-Ab	48.6	41.2	21.7
ESMFold2	47.1	35.7	16.1
Best of all models (25 of 75)	55.5	46.8	21.5
FKCFold, confidence	60.8	51.0	19.9
FKCFold, conf. + density	54.1	47.2	21.4

Table 9: FKCFold minus best of all models in acceptable top-1 success, 25 candidates each; CIs resample complexes (5,000 replicates). The two use different selectors, so these are not fixed-confidence-module controls.

FKCFold selector	Difference (pp)	95% CI
FoldBench: protein–protein		
Confidence	$+1.35$	$[-1.96, 4.97]$
Confidence + density	$+0.64$	$[-0.92, 2.22]$
FoldBench: antibody–antigen		
Confidence	-0.63	$[-6.17, 5.32]$
Confidence + density	$+4.05$	$[-1.78, 10.22]$
2026ARK-AB		
Confidence	$+1.43$	$[-1.28, 4.41]$
Confidence + density	$+1.68$	$[-1.43, 4.93]$
PXM-2025: antibody–antigen		
Confidence	$+7.04$	$[2.55, 12.25]$
Confidence + density	$+2.25$	$[-2.44, 7.24]$

Table 10: Main comparison restricted to complexes whose coupled run used one GPU (interfaces kept: FB PP 247 of 279, FB Ab–Ag 159 of 171, ARK 260 of 390), with the selectors and 25 candidates of the main tables. FB, FoldBench; ARK, 2026ARK-AB.

	Top-1 acc. (%)			Oracle acc. (%)		
Method	FB PP	FB Ab–Ag	ARK	FB PP	FB Ab–Ag	ARK
Protenix-v2	71.7	59.1	60.4	76.9	78.0	74.6
OpenDDE(-Ab)	72.1	64.8	59.2	76.1	73.6	74.6
ESMFold2	74.1	58.5	51.5	81.0	69.2	61.2
Best of all models (25 of 75)	74.6	68.4	60.3	82.5	81.2	79.0
FKCFold, confidence	75.7	68.6	62.7	83.8	84.3	77.7
FKCFold, conf. + density	74.9	71.7	63.5	83.8	84.3	77.7

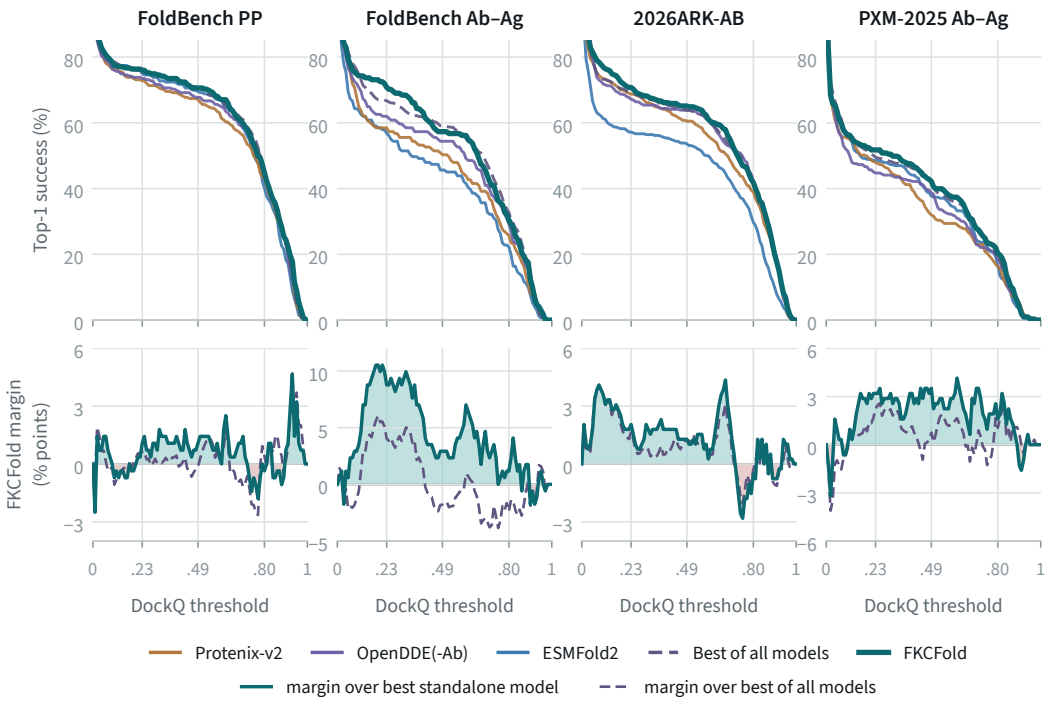

Figure 4: Top-1 success as a function of the DockQ threshold (PP, protein–protein; Ab–Ag, antibody–antigen), with 25 candidates per method. Top: percentage of interfaces whose top-1 prediction has DockQ at or above the threshold on the horizontal axis; values at 0.23, 0.49, and 0.80 equal the acceptable, medium, and high rates in the tables. The axis stops at 85%, which only clips curves below DockQ 0.05. Bottom: FKCFold's margin, its top-1 success minus that of a reference, in percentage points. The solid line and shading use as reference the highest of the three standalone curves at each threshold, so the reference model may change with the threshold; shading is positive where FKCFold leads. The dashed line uses best of all models. Each margin panel has its own vertical scale. FKCFold uses confidence plus density.

the complexes whose coupled run used one GPU (24 PP and 8 Ab–Ag FoldBench complexes and 55 2026ARK-AB complexes are excluded). With the combined selector, FKCFold's ranking relative to the individual models and to best of all models is the same as on the full panels, and on the excluded complexes FKCFold's oracle coverage is at least that of every individual model.

Table 11: Stored FoldBench held-out partition: acceptable top-1 success and mean top-1 DockQ. Differences are FKCFold minus Protenix-v2 in mean DockQ; CIs from 5,000 paired bootstrap resamples of complexes.

Panel / selector	Top-1 acc. (%)	Mean DockQ	Δ Mean DockQ	95% CI
Protein–protein: 54 interfaces, 48 complexes				
Protenix-v2	72.2	0.5731	—	—
FKCFold, confidence	68.5	0.5474	-0.0257	$[-0.0619, 0.0031]$
FKCFold, conf. + density	66.7	0.5418	-0.0313	$[-0.0685, -0.0024]$
Antibody–antigen: 20 interfaces, 16 complexes				
Protenix-v2	65.0	0.4862	—	—
FKCFold, confidence	40.0	0.3056	-0.1806	$[-0.3646, -0.0257]$
FKCFold, conf. + density	55.0	0.4118	-0.0743	$[-0.1838, 0.0024]$

F HELD-OUT SUBSET ANALYSIS

An earlier FoldBench analysis stored a target-level development/held-out partition (salt `fkcfold-20260919-heldout-v1`). Table 11 reports both FKCFold selectors on its held-out interfaces. This small subset does not confirm the full-panel gains, and because the selectors were explored during development, it is a sensitivity analysis, not a preregistered test.

G COMPUTATIONAL COST

Table 12: Mean wall-clock time for one seed on one H100 80 GB GPU, over the 21 complexes profiled in all four configurations. Generation includes startup, loading, featurization, and output; Total adds confidence evaluation.

Configuration	Candidates	Generation (s)	Confidence (s)	Total (s)
Protenix-v2	5	58.5	1.0	59.5
OpenDDE-Ab	5	88.8	2.4	91.3
ESMFold2	5	118.2	2.2	120.4
FKCFold	5	243.3	5.6	248.9
Three standalone runs (sum)	15	265.5	5.6	271.1

Table 12 covers all 21 targets with complete timing records at the audit snapshot. Each run generates five candidates through the same implementation, with the active model set determining whether generation is standalone or coupled. Timings include process startup, model loading, featurization, generation, and confidence evaluation; MSA search and DockQ scoring are excluded.

H ALIGNMENT AND NOISE-CAP ABLATIONS

Each variant changes one setting of the deployed FKCFold run and keeps all others: the same 319 FoldBench complexes (215 PP and 104 Ab–Ag): the 351 evaluated complexes minus the 32 among them whose coupled run needs four GPUs (the 33rd four-GPU complex, `8xnh-assembly1`, is already outside the Ab–Ag panel; Appendix C), five seeds with five candidates each, 200 sampling steps, and the same model inputs. *No per-step alignment* averages the models' denoised structures without first superimposing each prediction on the current sample. *No noise caps* queries every model at every noise level, so ESMFold2 is also used above its native maximum. The control is the saved FKCFold run on the same complexes. Reruns are distributional rather than bitwise matches, because GPU arithmetic is not deterministic. Variants are compared by the deployed confidence-plus-density selector and by oracle coverage, both computed from each arm's own 25 candidates.

Table 13: Design-choice ablations on the 319 one-GPU FoldBench complexes (215 PP, 104 Ab–Ag; 25 candidates per arm): each variant changes one FKCFold setting, against the saved FKCFold run. Top-1 uses the confidence-plus-density selector; brackets give 95% CIs over complexes (5,000 replicates).

		Top-1 acc. (%)	Variant minus control		
Variant	Panel	Control → variant	Acc. (pp)	Mean DockQ	Oracle acc. (pp)
No alignment	PP	74.9 → 73.7	−1.2 [−2.8, 0.0]	−0.001 [−0.006, 0.004]	−2.0 [−4.0, −0.4]
No alignment	Ab–Ag	71.7 → 70.4	−1.3 [−5.5, 3.0]	−0.003 [−0.020, 0.013]	−1.3 [−5.0, 1.9]
No noise cap	PP	74.9 → 74.5	−0.4 [−2.0, 0.8]	+0.004 [−0.001, 0.011]	+0.4 [−0.8, 1.7]
No noise cap	Ab–Ag	71.7 → 71.7	+0.0 [−3.3, 2.6]	+0.005 [−0.005, 0.014]	−1.3 [−5.8, 2.6]

I SELECTION, MODEL REMOVAL, AND POOLING CONTROLS

Table 14: Selector ablation on the same 25 coupled FKCFold candidates per complex: acceptable top-1 success (Acc.) and mean top-1 DockQ. Differences are each selector minus confidence plus density; brackets give 95% CIs over complexes (5,000 replicates).

Selector	Acc. (%)	Δ Acc. (pp)	Mean DockQ	Δ Mean DockQ
FoldBench: protein–protein: 279 interfaces, 239 complexes				
Confidence only	76.7	+0.7 [−2.5, 4.4]	0.620	+0.0083 [−0.0156, 0.0369]
Density only	76.3	+0.4 [−0.7, 1.7]	0.609	−0.0025 [−0.0107, 0.0037]
Confidence + density	76.0	—	0.612	—
FoldBench: antibody–antigen: 171 interfaces, 112 complexes				
Confidence only	66.1	−4.7 [−10.5, 0.6]	0.490	−0.0317 [−0.0670, −0.0004]
Density only	69.0	−1.8 [−6.2, 2.2]	0.510	−0.0119 [−0.0358, 0.0100]
Confidence + density	70.8	—	0.522	—
2026ARK-AB: 390 interfaces, 161 complexes				
Confidence only	70.5	−0.3 [−2.4, 1.7]	0.566	−0.0033 [−0.0179, 0.0106]
Density only	70.8	+0.0 [−3.4, 3.6]	0.574	+0.0045 [−0.0180, 0.0293]
Confidence + density	70.8	—	0.569	—
PXM-2025: antibody–antigen: 313 interfaces, 130 complexes				
Confidence only	56.5	+4.8 [1.7, 8.6]	0.396	+0.0220 [0.0043, 0.0437]
Density only	49.5	−2.2 [−5.7, 0.4]	0.360	−0.0142 [−0.0386, 0.0053]
Confidence + density	51.8	—	0.374	—

Table 14 ranks the same 25 coupled candidates per complex by confidence alone, by density alone, and by the combined rule equation 9, keeping the kernel width, confidence coefficient, tie handling, and evaluation coverage fixed. The confidence and combined rows reproduce the main tables; density alone is the only new row, and Table 15 gives all DockQ tiers.

Table 3 removes one model at a time from FKCFold on the 319 FoldBench complexes of Appendix H, with all five seeds per arm and the settings of the saved runs. ESMFold2 is removed

Table 15: All DockQ tiers for the selector ablation: top-1 success (%) and mean top-1 DockQ from the same 25 coupled candidates. Acc., Med., and High: DockQ ≥ 0.23, ≥ 0.49, and ≥ 0.80.

Selector	Acc.	Med.	High	Mean DockQ
FoldBench: protein–protein (oracle acceptable 84.2%)				
Confidence only	76.7	71.7	45.5	0.620
Density only	76.3	70.3	43.7	0.609
Confidence + density	76.0	70.6	44.8	0.612
FoldBench: antibody–antigen (oracle acceptable 83.0%)				
Confidence only	66.1	54.4	28.7	0.490
Density only	69.0	57.9	29.8	0.510
Confidence + density	70.8	57.3	30.4	0.522
2026ARK-AB (oracle acceptable 82.6%)				
Confidence only	70.5	65.4	42.1	0.566
Density only	70.8	65.9	42.3	0.574
Confidence + density	70.8	65.1	41.8	0.569
PXM-2025: antibody–antigen (oracle acceptable 65.5%)				
Confidence only	56.5	45.0	21.1	0.396
Density only	49.5	39.6	17.9	0.360
Confidence + density	51.8	42.2	20.1	0.374

on both panels; OpenDDE-Ab and Protenix-v2 are removed on Ab–Ag. The three-model control is the saved FKCFold run on the same complexes. Because each two-model arm lacks one confidence module, top-1 is selected by a module present in both arms: Protenix-v2 when ESMFold2 or OpenDDE-Ab is removed, and ESMFold2 when Protenix-v2 is removed. The 16-complex pilot below preceded these runs.

Table 16: Model removal on the 16-complex pilot (24 FoldBench Ab–Ag interfaces, 10 candidates per arm), ranked by *stored confidence*: the Protenix-v2 score recorded during each run, unlike the common rescoring of Table 2. Differences are two-model minus three-model arm; brackets give 95% CIs over complexes (10,000 replicates).

Models	Top-1 acc. (%)	Top-1 mean DockQ	Oracle acc. (%)
Model combination			
Protenix-v2	50.0	0.432	66.7
Protenix-v2 + OpenDDE-Ab	58.3	0.531	75.0
Protenix-v2 + ESMFold2	62.5	0.511	87.5
Protenix-v2 + OpenDDE-Ab + ESMFold2	75.0	0.539	87.5
Two-model minus three-model arm (pp; DockQ)			
Without ESMFold2	-16.7 $[-38.5, 0.0]$	-0.009 $[-0.086, 0.063]$	—
Without OpenDDE-Ab	-12.5 $[-30.4, 0.0]$	-0.029 $[-0.090, 0.029]$	—

The pilot in §5.5 asks whether ESMFold2 adds to the Protenix-v2 and OpenDDE-Ab composition. Eligible complexes are the 104 FoldBench Ab–Ag complexes whose saved three-model and Protenix-v2 predictions are complete and scored for seeds 42 and 66 and whose coupled run needed one GPU. They are split into four equal-count groups by total protein length, and four complexes are drawn from each group with a fixed random seed. The new arm uses FKCFold's shared-denoiser update with Protenix-v2 and OpenDDE-Ab and the settings of the saved runs. All 32 target–seed jobs completed on one H100 GPU each, taking 4,854 s in total. The three-model and Protenix-v2 arms reuse the saved seed-42 and seed-66 predictions, and selection and oracle coverage are recomputed within each arm's 10 candidates. Rerunning saved three-model seeds with the current code reproduces most candidates within 2 Å Cα RMSD, but not bit for bit: GPU arithmetic is not deterministic, and some trajectories reach different poses. The saved controls use the same targets, seeds, and sampling settings, but are not exact reruns. A second new arm couples Protenix-v2 with ESMFold2 under the same settings, keeping ESMFold2's noise cap, so Protenix-v2 acts alone at the

highest noise levels. All 32 jobs completed on one H100 GPU each, taking 5,922 s in total. Removing OpenDDE-Ab lowers acceptable top-1 success by 12.5 points and leaves acceptable oracle coverage unchanged at 87.5% (Table 16).

Table 2 asks whether coupling generates candidates that independent sampling of the same three models does not. With no selector, the 10 coupled candidates contain an acceptable pose for 87.5% of interfaces, against 77.5% expected for 10 candidates drawn from the 30 independent ones, a difference of +10.0 points [+0.2, +21.9]. All 30 independent candidates reach 83.3%. Oracle coverage needs no selector, and the 10-of-30 values are exact expectations over uniform subsets, computed as in Appendix D. An acceptable candidate exists only among the coupled candidates for 1 interface, only among the independent ones for 0, in both for 20, and in neither for 3. Top-1 selection under one confidence module does not preserve this advantage. When all 40 candidates of a complex are rescored by Protenix-v2, with each candidate's confidence averaged across two passes before ranking, coupled top-1 success is 50.0%, against 51.1% for the expected 10 of 30. The same coupled candidates reach 75.0% with the confidence values stored during generation (Table 16). The two analyses use the same structures with stored or rescored confidence values. Repeated Protenix-v2 passes shift confidence by up to 0.08, mostly uniformly within a seed, and the leading candidates of several complexes differ by less than 0.001. Small score gaps make candidate ordering sensitive to rescoring variability. Oracle coverage assesses candidate generation independently of confidence rescoring.