GUIDING WHERE THE ANTIBODY BINDS: AMORTISED CROSS-SUBSAMPLE PARTICLE STEERING FOR ANTIBODY–ANTIGEN CO-FOLDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Co-folding models have advanced antibody-antigen structure prediction, but a dominant failure mode persists. They often dock the antibody onto the wrong epitope. Docking has two problems. A sampler must produce an acceptable complex at the correct interface, and a verifier must then rank it above the alternatives. We address the first, *site coverage*, since a verifier cannot select an epitope the sampler never proposed. Drawing more seeds or samples adds variety around the pose the model already favours, but does not relocate it. We introduce **CAST** (*Cross-subsample Amortised Site Tilting*), which steers diffusion trajectories toward a binding *site* using a reward built from a candidate epitope map, training-free and with no reference structure. The sampler approximates an ideal target whose tilt provably changes the site marginal but not the pose at a site. Amortising K alignment-subsample latents into one steered pool cuts the cost of exploring them from $\mathcal{O}(KN)$ to $\mathcal{O}(N+K)$ for a pool of N particles. At a matched 720 structures and 4 conditionings, CAST holds an acceptable pose on 36 of the 50 difficulty-balanced CFA50 targets against 26 for Best-of-N, and raises mean max DockQ from 0.322 to 0.373, a gain of 15.8%. On 21 targets that a $1{,}000$-seed run fails to reach CAPRI-medium quality, CAST reaches mean max DockQ 0.370 at 720 structures and brings 18 of the 21 to an acceptable pose against 9 at the same budget. Best-of-N flattens, spending $5.3\times$ the compute and trailing by 20.1%. Raising the latent count alongside the samples carries this to 0.457, beyond every Best-of-N budget. On CFA50 CAST also matches the best Best-of-N allocation of 720 structures, at $1.5\times$ less compute. Such steering is a reallocation, changing little where the base model already succeeds and much where it does not.

1 INTRODUCTION

A predicted antibody–antigen complex is useful only if the antibody is docked at the right epitope, the patch a design campaign builds against (Mille-Fragoso et al., 2026) and an affinity model reads (Liu et al., 2025), and the system can recognise it. Co-folding therefore has two bottlenecks. The sampler must produce an acceptable complex at the correct interface, and a verifier must then rank it above the incorrect alternatives.

This work addresses the first, which we call *site coverage*. A verifier can reorder structures already present in a pool. It cannot, however, select a correct epitope that the sampler never proposed. The two problems are also separate in practice, since the interface confidence scores used to rank such structures are themselves unreliable (Dunbrack, 2025). We therefore isolate sampling coverage and report what the pool makes available, the number of targets holding an acceptable complex and the best structure present, $\mathrm{DockQ}^{\max}_{\mathrm{abag}}$ (Mirabello & Wallner, 2024), not what a ranker returns.

Coverage is where co-folding currently fails. Models predict antibody–antigen complexes directly from sequence (Abramson et al., 2024; Wohlwend et al., 2024; Chen et al., 2025; Candido et al., 2026; Zhang et al., 2026b), yet AlphaFold3 fails to dock 65% of them at one seed (Hitawala & Gray, 2025), and accuracy drops further without a close structural analogue in training (Fromm et al., 2026). Across $1{,}628$ complexes, methods place the antibody incorrectly and converge on the *same redundant binding mode* (Clifford & Nielsen, 2026). That wrong site is not a near miss. A

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.

structure on a different patch shares almost no native contact, so its DockQ is near zero. No local move reaches a distant patch and repairs it, and the model will not offer an alternative. This calls for two capabilities, naming a site and steering the sampler to it.

Spending more supplies neither. Fromm et al. (2026) find accuracy rises roughly linearly in the *logarithm* of the sample count, because independent draws deepen the mode the model favours instead of relocating it. Of the 52 targets we collected, 18 still hold nothing acceptable at 1,000 unsteered structures each, and 7 recover no native contact at all (§C.3).

Related work. *Conditioning names a site, but pays for each one.* A Chai-1 (Chai Discovery et al., 2024) pocket restraint from a B-cell epitope predictor improves accuracy and diversity (Clifford & Nielsen, 2026), so the modes are reachable inside a frozen model. But the restraint enters before the trunk and is fixed for the trajectory, so S candidate sites cost S trunk passes. Inside the diffusion loop a site costs no trunk pass of its own. *Steering works there, but names no site.* It diversifies the *conformation* a complex adopts, not the *site* where it binds, by weakening the alignment signal (Kalakoti & Wallner, 2025; 2026), rescaling trunk latents (Suzuki & Amagasa, 2026a; Lee et al., 2026), or guiding and repelling in the sampler (Richman et al., 2025; Corso et al., 2024; Suzuki & Amagasa, 2026b). Those that repel score by deviation from a reference, so the mass they move lands wherever the base model ranks next, never on a site we name. The few that aim somewhere use a target correct by construction that we lack, an experimental measurement (Li et al., 2026) or physical prior knowledge (Wang et al., 2024). Twisted sequential Monte Carlo (Wu et al., 2023) and Feynman–Kac steering (Singhal et al., 2025) are training-free and backbone-agnostic but prescribe *how* to steer, not *what* to steer toward. We are not aware of a method that carries a binding site as an objective inside the diffusion loop (§A).

We introduce **CAST**, Cross-subsample Amortised Site Tilting, which supplies that objective. It is training-free, builds on the twisted sampler, and steers it toward a binding site from inside the loop. Without a reference structure the only site signal is a predicted epitope, itself an open problem (Liu et al., 2024). CAST therefore treats its map as a direction, not an answer, and sweeps a grid of candidates. A site is a destination carried by a particle, and so is the conditioning latent, so site hypotheses and the K MSA subsamples share one pool instead of K pools. At the same 720 structures this brings 36 of 50 targets to an acceptable pose against 26.

Contributions. **(i) A site collective variable and its tilt** (§2.2), a reward on the paratope's position alone, hinged at an arrival radius and swept over *umbrellas*, one per target separation radius $d^\star$. It is a soft prior, so CAST reaches epitopes a displaced map candidate would exclude. **(ii) Cross-subsample amortisation** (§2.2). We make the conditioning latent part of the particle state, so the K subsamples share one pool. That cuts the cost from $\mathcal{O}(KN)$ to $\mathcal{O}(N+K)$, and stratification keeps the candidate grid occupied at any K. **(iii) A reference-free epitope map** (§2.4) from pool binding modes plus a farthest-point surface tiling, without ground truth. **(iv) A pose-invariance guarantee** (§2.3). Under a site-measurable reward the ideal target preserves the pose conditional at a fixed site, at every temperature. The tilt thus controls the site marginal alone, and leaves the pose at that site to the base model. **(v) Empirical gains and the axis they scale on** (§4). At a matched 720 structures the gain sits entirely on the targets Best-of-N does not reliably solve, where CAST brings 16 of 29 to threshold against 6. It reaches threshold on 36 targets against 30, 27, 24 and 22 for four steered baselines. The gain scales as latents and samples rise together, 0.370 to 0.457, while Best-of-N flattens at every budget. CAST also transfers to a second backbone without retuning (§C.9).

2 METHOD

2.1 PRELIMINARIES

A co-folding model maps sequences to atomic coordinates in two stages (Abramson et al., 2024; Zhang et al., 2026b). A *trunk* embeds the input, including a multiple-sequence alignment, and produces a conditioning representation z. A diffusion *decoder* then samples $x \sim p_\theta(x \mid z)$ by integrating a reverse process from noise. The decoder is variance-exploding (Karras et al., 2022), with $x_t = x_0 + \sigma_t \varepsilon$ for $\varepsilon \sim \mathcal{N}(0, \mathbb{I})$. Writing $p_t(x_t \mid z)$ for the marginal at noise level σ_t and D_θ for the denoiser, Tweedie's identity relates the denoiser to the score (Vincent, 2011; Karras et al., 2022),

$$\hat{x}_0 \;=\; D_\theta(x_t, \sigma_t \mid z) \;\approx\; \mathbb{E}[x_0 \mid x_t, z] \;=\; x_t + \sigma_t^2 \nabla \log p_t(x_t \mid z), \tag{1}$$

with equality in the first step for a Bayes-optimal denoiser. The sampler is the stochastic update of Karras et al. (2022). A step recenters and randomly rotates the coordinates (Abramson et al., 2024), injects noise to raise the level from σ_t to $\hat{\sigma}_t \geq \sigma_t$,

$$\tilde{x}_t = x_t + \beta_t \varepsilon_t, \qquad \beta_t = \sqrt{\hat{\sigma}_t^2 - \sigma_t^2}, \qquad \varepsilon_t \sim \mathcal{N}(0, \mathbb{I}), \tag{2}$$

denoises once at that raised level, then descends deterministically to $x_{t-1} = \tilde{x}_t + (\sigma_{t-1} - \hat{\sigma}_t)(\tilde{x}_t - \hat{x}_0)/\hat{\sigma}_t$, along $-\hat{\sigma}_t \nabla \log p_{\hat{\sigma}_t}(\tilde{x}_t \mid z)$ by equation 1. The rigid-frame augmentation is an isometry, so after it the injection in equation 2 is the only stochastic perturbation of the diffusion state.

Reward tilting. Biasing such a decoder toward high-reward structures is usually posed as sampling from a reward-tilted target, $q(x) \propto p_\theta(x \mid z) \exp(\lambda r(x))$ for a reward r and a strength $\lambda > 0$. Reaching it needs the *time marginals* of q, which carry an intermediate potential $h_t(x_t) = \mathbb{E}[\exp(\lambda r(x_0)) \mid x_t, z]$ under the endpoint posterior, since $q_t \propto p_t\, h_t$ makes the tilted score the base score plus $\nabla \log h_t$. That is exact and intractable, and a structural reward is in any case defined on clean coordinates not on a noisy iterate. Evaluating it on the denoised estimate answers both at once (Chung et al., 2023; Wu et al., 2023), giving the plug-in potential $\hat{h}_t = \exp(\lambda r(\hat{x}_0))$, whose gradient carries the denoiser Jacobian $J_{D_\theta}(x_t) = \partial \hat{x}_0 / \partial x_t$,

$$\nabla_{x_t} \log \hat{h}_t = \lambda\, J_{D_\theta}(x_t)^\top \nabla_{\hat{x}_0} r(\hat{x}_0). \tag{3}$$

Twisted sequential Monte Carlo treats such a tilt as a *proposal*, not as the answer and corrects it by reweighting a population of particles (Wu et al., 2023; Singhal et al., 2025). Reserving q for the target, a guided sampler draws the injection of equation 2 from a shifted mean $x_t + g$ instead of x_t and changes nothing else. Writing $r_t = r(\hat{x}_0)$ for the reward read on the denoised estimate at step t, both factors of the weight are then closed-form,

$$w_t \;\propto\; \exp\left(-\left[\|g\|^2 + 2\beta_t \langle \varepsilon_t, g \rangle\right] / 2\beta_t^2\right) \;\cdot\; \exp\left(\lambda(r_{t-1} - r_t)\right), \tag{4}$$

an exact likelihood ratio for the Gaussian *injection kernel*, paying for the shifted injection drawn in place of the decoder's own, times a *potential ratio* imposing the reward tilt. Reverse sampling runs from t to $t-1$, so the potential ratio exceeds one exactly when the reward improves. Particles are resampled in proportion to these weights when their effective sample size falls too low, which turns weight into pool composition.

2.2 CAST

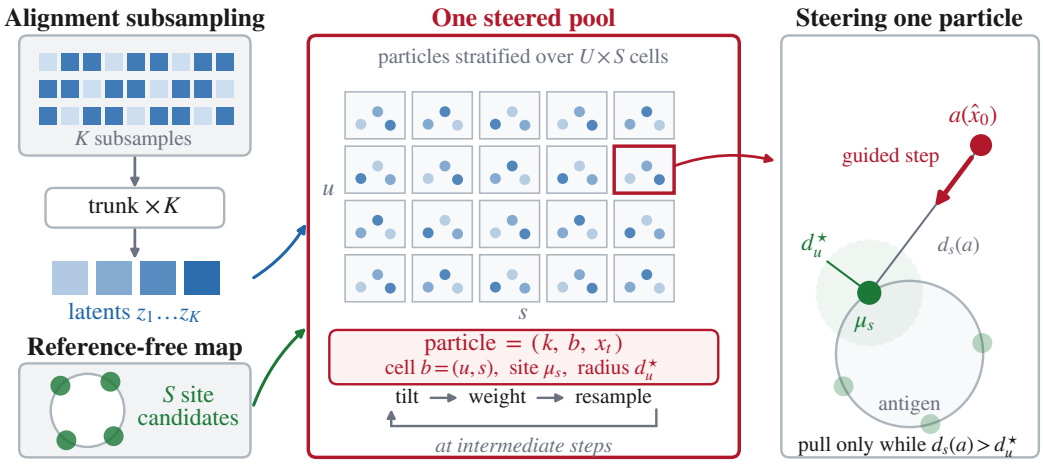


Figure 1: **CAST in one pool.** K alignment subsamples and the reference-free map of §2.4 feed one steered pool, whose particles are stratified over the $U \times S$ cells. **Right**, one cell opened up. The guided step displaces the paratope centroid toward the candidate centroid μ_s, and vanishes once it is within the arrival radius $d_u^\star$.

CAST augments the particle state with a destination. Each particle carries the candidate site it targets and the conditioning it was drawn from, both fixed for the whole trajectory. One pool therefore covers the full candidate grid across all K MSA subsamples. A particle is the triple $(k,\ b,\ x_t)$, with diffusion state x_t, an index k into the conditionings $z_1, \dots, z_K$, and cell $b = (u, s)$ pairing a candidate site s with an arrival radius u, or *umbrella*. Guidance displaces each particle toward its cell, and resampling selects among them (Figure 1). Resampling may reallocate k, whereas stratification fixes the cells. Both act on the antibody's position, not its pose.

Site and pose. The *paratope* is fixed before sampling as the IMGT CDR residues under ANARCI numbering (Dunbar & Deane, 2016), so it is defined on any partly denoised structure. We parameterise a complex by a site coordinate a, an orientation R and internal coordinates c. The site a is the paratope centroid in the antigen frame, so a global rigid motion of the complex leaves it fixed, and c is expressed about that centroid, so a change of internal geometry cannot move it. The chain rule then gives each latent its own factorisation

$$x = (a, R, c), \qquad p(a, R, c \,|\, z_k) = p(a \,|\, z_k)\, p(R, c \,|\, a, z_k). \tag{5}$$

The first factor is *where* the antibody sits, the second is *how* it is arranged once placed. Proposition 1 shows that a reward depending on x only through a leaves the second factor invariant.

Joint target over structures and conditionings. Different alignment subsamples produce different conditioning representations (Kalakoti & Wallner, 2025; 2026), so z_k indexes a genuinely *different* conditional and not another draw from a single one, so the choice of conditioning becomes part of the sampling problem instead of a setting fixed beforehand. Given candidate centroids μ_s from the map of §2.4, arrival radii $d_u^\star$, and a uniform prior $\rho_k = 1/K$ over latents, each cell b defines a reward-tilted target on the augmented state (Wu et al., 2023; Singhal et al., 2025),

$$q_b(k, x) \ \propto\ \rho_k\, p_\theta(x \,|\, z_k) \exp\big(\lambda\, r_b(a)\big), \qquad r_b(a) = -\,\ell_\delta\big(\max(0,\ d_s(a) - d_u^\star)\big), \tag{6}$$

where ℓ_δ is a Huber function and $d_s(a) = \|a - \mu_s\|$ is the paratope-to-candidate separation. We write $r_b(x) = r_b(a(x))$ for the reward read on a structure, so r_b sees the site coordinate and nothing else, the property Definition 1 names. The hinge makes $d_u^\star$ an arrival radius, not a target point, so a particle already inside it feels neither reward nor gradient, and it is also what keeps the strong-tilt limit a conditioning instead of a collapse (Corollary 1). We sweep u over several paratope-to-site distances (§B.6). Marginalising the latent tilts the *mixture* of the K conditionals,

$$q_b(x) \ \propto\ p_{\mathrm{mix}}(x)\ \exp\big(\lambda\, r_b(a)\big), \qquad p_{\mathrm{mix}}(x) = \textstyle\sum_{k=1}^{K} \rho_k\, p_\theta(x \,|\, z_k), \tag{7}$$

so a single particle system can reallocate decoder trajectories across conditionings. K separate pools do not perform this reallocation, because their allocations are fixed in advance, and matching equation 7 would additionally require weighting the pools by their per-latent normalising constants. Particles are partitioned over the cell grid $\mathcal{U} \times \mathcal{S}$, indexing U arrival radii against S candidates, each cell holding a fixed share of the pool and targeting its own q_b. Reserving budget cell by cell is what lets the pool reach sites the base model would not visit on its own.

One pool over K latents. Both designs evaluate the trunk K times. Steering separately runs K pools of N particles each, whereas CAST defers the decoder until all K latents are in hand and runs one pool of N that every latent shares. Separate steering then needs KN decoder trajectories where CAST needs N, so a latent is an addition for CAST, $\mathcal{O}(N + K)$, and a multiplication for a separate-model arm, $\mathcal{O}(KN)$. The multiplication is stratification's, since each separate pool must occupy the whole cell grid on its own. A cell must also hold enough particles to resample with its occupancy divided among the latents, so N and K are floored and capped together (§B.4, with structure counts in §C.1).

Guided proposal. CAST steers by displacing the mean of the Gaussian injection in equation 2 for particle i by $g_t^i = \mathrm{clip}\big[\eta_t\, \nabla_{\hat{x}_0} r_b(\hat{x}_0^i)\big]$, under a step η_t annealed with the noise level and a size-invariant cap (§B.5). Because the churn precedes this step's denoise, it is read on the preceding step's denoised estimate and adds no decoder call of its own. Each scoring event below costs four denoiser passes per particle (§C.1). Algorithm 1 in §B.1 states the whole sampler. The reward reads a centroid separation, so $\nabla_{\hat{x}_0} r_b$ is available in closed form, and we take the Jacobian-free straight-through approximation $J_{D_\theta} \approx \mathbb{I}$ in equation 3 (Bengio et al., 2013; Moufad et al., 2026), so the guided step needs no backward pass through the decoder or the reward. The injection is the step's only stochastic perturbation and the shift is applied only there, so equation 4 is exact at $g = g_t^i$.

Weights and resampling. Weights accumulate both factors of equation 4, the potential ratio only at a sparse set of *scoring events* $m = 0, \dots, M$, with r_m^i the reward of particle i at event m. The potential enters through increments, and these telescope to $e^{\lambda(r_M^i - r_0^i)}$. We add the initial term back once, giving $e^{\lambda r_M^i}$, which keeps the baseline differences between latents and needs $K > 1$ (§B.3). Resampling runs *within* cells and only inside an early window $\mathcal{W}$, and the two restrictions protect different diversity. Within cells protects *site* diversity, since a global resample would let one cell take another's budget, at the price of capping each candidate at a $1/S$ share of the pool. Early protects *pose* diversity, since once the site has settled a resample can only clone, and under the ideal target a site-only reward prefers no pose over another. Weights are consumed only inside $\mathcal{W}$ and the ensemble is returned unweighted, while guidance continues past $\mathcal{W}$ as an unweighted drift, so the asymptotic guarantee of Wu et al. (2023) does not apply.

2.3 A SITE TILT PRESERVES THE POSE CONDITIONAL

We now make the separation between site and pose precise. Definition 1 names the class of rewards it requires, Proposition 1 states the invariance, and Corollary 1 gives the strong-tilt limit.

Definition 1 (site-measurable reward)**.** *A reward is* site-measurable *if it depends on a structure $x = (a, R, c)$ only through its site coordinate, $r(x) = r(a)$.*

The reward of equation 6 is site-measurable. A reward reading interface contacts, clashes, orientation or internal geometry is not site-measurable, and none of what follows applies to those.

Proposition 1. *Let q be the tilt of equation 6 with any site-measurable reward. For every λ the pose conditional is unchanged both within each latent and, after marginalising k, for the mixture,*

$$q(R, c \mid a, z_k) = p(R, c \mid a, z_k) \ \textit{for every } k, \qquad \textit{and} \qquad q(R, c \mid a) = p(R, c \mid a).$$

Corollary 1 (the $\lambda \to \infty$ limit is a conditioning, not a collapse)**.** *Write $\mathcal{B}_b = \{a \in \mathbb{R}^3 : \|a - \mu_s\| \le d_u^\star\}$ for the arrival ball, and assume $p_{\mathrm{mix}}(a \in \mathcal{B}_b) > 0$. With the hinge of equation 6 the tilt factor tends to an indicator on $\mathcal{B}_b$, so the mixture target of equation 7 tends to a conditioning of the same mixture, $q_b(x) \xrightarrow[\lambda\to\infty]{} p_{\mathrm{mix}}\big(x \mid a \in \mathcal{B}_b\big)$. The same holds at a fixed latent whenever $p_\theta(a \in \mathcal{B}_b \mid z_k) > 0$, with $q_b(x \mid k) \propto p_\theta(x \mid z_k) e^{\lambda r_b(a)}$ tending to $p_\theta(x \mid a \in \mathcal{B}_b, z_k)$.*

Both proofs are in §B.2. The hinge creates the zero-reward region inside the ball, and without it a larger λ would favour ever smaller separations instead of conditioning on an arrival region. The finite particle system approximates q_b through a plug-in potential, a straight-through gradient, and a truncated window. Only the last costs exactness, and §B.1 states the cost of each.

2.4 REFERENCE-FREE CANDIDATE MAP

The map is built without ground truth from two ingredients with complementary coverage. Pool *modes* are the binding-site modes of an unsteered pilot pool, antibody-conditioned by construction and requiring no reference structure. *Tiles* come from a farthest-point covering of the antigen surface, so every residue lies within the covering radius of some seed. The map lists the leading modes, then tiles ranked by their overlap with those modes, up to a fixed candidate budget. §B.7 gives the construction, and sizes are in §3.

3 EXPERIMENTAL SETUP

Evaluation set. CFA50 comprises 50 antibody- and nanobody–antigen complexes from the CoFold Arena panel (CoFold Arena, 2026), which admits only structures released after 9 September 2024 and so gives a clean time split. Two different quantities are at work. CoFold Arena scores a target by $\mathrm{DockQ}_{\mathrm{abag}}$ on the model's single *best-ranked* structure. We screened 80 panel entries, stratifying on format, size and antigen identity (§C.2). We report the pool *maximum* instead, so we then ran an unsteered pool of 200 structures per candidate and kept the 50 that *balance* whether any structure reaches $\mathrm{DockQ}_{\mathrm{abag}} \ge 0.23$. **Collected 21-hard** is a second set of 21 complexes, used to see how performance scales with budget on hard cases. We curated it from PDB depositions released after 30 June 2025 and novel against earlier entries, under the structure and interface filters of §C.2. CDR-sequence clustering then keeps only a cluster representative, so they are not near-duplicates of one

another. From these we kept the targets a 100-seed run of the base model leaves unsolved, then the 21 that a 1000-seed run still leaves below the CAPRI-medium threshold 0.49 (Méndez et al., 2003). Selection and evaluation use different runs. The 1000-seed run selects the targets and a separate 50-conditioning pool supplies every Best-of-N number, so the draws scored are disjoint from those that chose the set.

Baselines. Every baseline gets CAST's budget and conditioning count, 720 structures as 4 MSA subsamples $\times$ 180 samples. *Best-of-N* (BoN) is the unsteered control and the reference for all comparisons. Since a matched budget can still be spent badly, we also sweep its allocation (§C.6). Four further baselines each steer by a different mechanism. *Feynman-Kac steering* (FK) (Singhal et al., 2025) applies the same reward and resampling without the gradient, at a single $d^\star = 12$ Å. *Noise search* (NS) (Jajal et al., 2025) runs an evolutionary search over the initial noise. We run their CoSyNE at a population of 12 over 7 generations, so that it spends the same 180 evaluations per conditioning. *Boltz-Perturb* (BP) (Jung et al., 2026) is conditioning perturbation with no objective and no resampling. *AF3-ReD* (ReD) (Ohnuki & Okazaki, 2026) repels each conditioning from the structures already produced, over whole-protein Cα, also with no objective and no resampling.

Implementation. All arms run on Protenix-v2 (Zhang et al., 2026b), one backbone and one checkpoint, 10 recycles and 200 diffusion steps, with MSAs subsampled to depth 1000 independently per conditioning. Steering uses the reference-free hybrid map at a 12-tile budget, giving at most $S = 15$ candidates, crossed with $U = 4$ umbrella radii $9, 11, 13, 15$ Å, so the $\mathcal{U} \times \mathcal{S}$ grid holds at most 60 resampling cells. The map's modes are read from a 150-structure unsteered pilot pool. Runs use A100 GPUs and ground truth is used for scoring only.

Metrics. $\mathrm{DockQ}^{\max}_{\mathrm{abag}}$, the pool maximum of DockQ restricted to the antibody–antigen interface (Basu & Wallner, 2016; Mirabello & Wallner, 2024), is the primary measure and states what a pool makes available, since ranking is a separate failure and existing confidence estimates often cannot identify the best structure a pool contains (Fromm et al., 2026). A target *reaches* 0.23 when its $\mathrm{DockQ}^{\max}_{\mathrm{abag}}$ clears the CAPRI-acceptable threshold (Méndez et al., 2003), and the tables count targets that do, which measures how many targets hold a usable structure at all. Compute is reported as PFLOPs per target, counting one trunk pass per conditioning plus the denoiser and confidence passes each structure costs (§C.1), so two arms at the same structure count can differ in compute.

Reporting protocol. Every $\mathrm{DockQ}^{\max}_{\mathrm{abag}}$ we compare is at a matched 720 structures, and a BoN rung uses the K lowest conditioning indices with $720/K$ samples from each. At $K = 4$ a rung holds every sample of its 4 conditionings, so the value is exact. Above $K = 4$ a rung takes a subset of each conditioning's samples, and we report the exact expectation of the maximum over those subsets (§C.5). The scaling comparison (§4.1, Figure 2(c,d)) states each budget, and averages over every conditioning subset of that size instead of fixing the lowest indices. CFA50 targets are grouped by the fraction of 4-conditioning draws in which BoN alone reaches $\mathrm{DockQ}^{\max}_{\mathrm{abag}} \geq 0.23$, computed exactly over every subset. *Reliably fails* at $\leq 10\%$, *reliably solves* at $\geq 90\%$, and *borderline* between (§C.4). Differences between arms are reported as means over targets, with paired win counts.

4 Results

4.1 CAST against BoN, at a matched budget and under scaling

At a matched budget CAST solves 10 more targets. Table 1 and Figure 2(a) give the comparison. Means are 0.373 against 0.322, a difference of $+0.051$ ($+15.8\%$), with 31 wins out of 50 at Wilcoxon signed-rank $p = 0.012$, and 10 extra targets reaching 0.23, 36 against 26. On the 16 targets BoN *reliably fails*, CAST wins 12 at $+0.071$ ($+63.9\%$) and reaches 0.23 on 4 against 0, which BoN does under no 4-conditioning draw. On the 13 *borderline* targets it wins 10 at $+0.131$ ($+51.8\%$) and reaches 0.23 on 12 against 6. On the 21 BoN *reliably solves* the difference is -0.015 (-2.8%), which the Wilcoxon test cannot separate from zero ($p = 0.99$).

CAST reaches binding sites the base model rarely samples. The mechanism is budget reallocation, by construction. CAST stratifies its pool over the $U \times S = 60$ cells and so spends most of it on wrong sites, while BoN spends everything where the base model already goes (SITE, Table 1). On the 16 targets BoN reliably fails, its closest structure over the whole 720 sits 5.52 Å from the true epitope on average against CAST's 3.02 Å, and on 4 of them CAST lands within 6 Å where

Table 1: $\mathrm{DockQ}^{\max}_{\mathrm{abag}}$ over the 50 CFA50 targets for CAST and BoN, both at a *matched* budget of 720 structures and the same 4 conditionings, BoN spending it as 4×180. Rows are grouped by how reliably the *BoN alone* solves a target (§C.4): *reliably fails* is $\leq 10\%$ of draws reaching $\mathrm{DockQ}^{\max}_{\mathrm{abag}} \geq 0.23$, *reliably solves* is $\geq 90\%$, and *borderline* is the rest, sizes in brackets. SITE measures how close an arm gets to the true site, using the centroid distance between the predicted and true epitopes for its closest structure. Lower values are better. REACH 0.23 counts targets on which that arm attains $\mathrm{DockQ}^{\max}_{\mathrm{abag}} \geq 0.23$, CAST WINS is a paired count spanning both rows, and p is a two-sided Wilcoxon signed-rank test on the per-target differences within that subset.

SUBSET	ARM	MEAN $\mathrm{DockQ}^{\max}_{\mathrm{abag}}$	SITE (Å)	REACH 0.23	CAST WINS	p
All targets (50)	CAST	**0.373**	**1.75**	**36**	31	0.012
	BoN	0.322	2.96	26		
BoN reliably fails (16)	CAST	**0.182**	**3.02**	**4**	12	0.020
	BoN	0.111	5.52	0		
Borderline (13)	CAST	**0.385**	**1.67**	**12**	10	0.040
	BoN	0.253	3.41	6		
BoN reliably solves (21)	CAST	0.511	0.83	20	9	0.99
	BoN	**0.525**	**0.72**	20		

no structure in BoN's pool does. 2 more sit in the borderline group, and where BoN already solves there is no difference, 0.83 Å against 0.72 Å. Every candidate site receives particles whatever the base model prefers, so CAST has the chance to explore the right site even where the base model puts little mass, which BoN at the same budget does not.

CAST at $K = 4$ matches BoN at $K = 20$ from $5\times$ fewer conditionings and $1.5\times$ less compute. Holding the budget at 720 structures, we vary only how it splits between conditionings and samples. BoN improves with conditioning count, from 0.322 at $K = 4$ to 0.366 at $K = 20$. CAST returns 0.373 from $K = 4$, within $+0.007$ of the top rung (Wilcoxon $p = 0.77$), which the test does not separate. Cost is what separates them, since that rung pays 20 trunk passes and 15.40 PFLOPs against CAST's 8.72 (Figure 2(b), §C.6). Charging the map's 150-structure pilot pool takes CAST to 10.04. CAST therefore reaches the same level from $5\times$ fewer conditionings and $1.5\times$ less compute.

On collected 21-hard CAST leads at every budget. We vary the budget on collected 21-hard. BoN can spend more two ways, by scaling conditionings or by scaling samples (Figure 2(c,d)). The two axes agree in structures and diverge in compute. Both baseline curves keep climbing and flatten as they go, so each added structure buys less than the last, and neither reaches 0.31 at any budget we ran. CAST passes that level at 720 structures and 9.73 PFLOPs. At that matched 720 it brings 18 of the 21 to CAPRI-acceptable DockQ against BoN's 9, and no BoN allocation of the same budget passes 11. The baseline's whole 2500-structure pool costs $5.3\times$ that compute and still sits 0.062 below, a 20.1% lead for CAST. Scaling CAST further needs both of its axes at once. We raise samples and latents together along $N = 180K$, which holds each latent's share of every cell fixed. Each rung is a separate run, reaching 0.439 at $K = 8$ and 0.457 at $K = 16$. The lead over the best baseline widens from 0.062 to 0.149, a 48.3% gain, while still spending three quarters of compute.

The gain transfers to a second backbone. We repeat the matched-budget and scaling comparisons on OpenDDE (Aureka AI OpenDDE project, 2026), with no hyperparameter retuned. Over 40 difficulty-balanced targets CAST reaches 0.454 against BoN's 0.372 ($p < 10^{-4}$), bringing 27 to an acceptable pose against 20. The gain again sits where BoN is unreliable, $+0.111$ on the 16 it reliably fails, where CAST also outscales BoN and exceeds every allocation of the budget (§C.9).

4.2 Steered baselines and ablations

FK, NS, BP and ReD (§3) are four other ways to spend the same budget on steering, by resampling without a gradient, by searching over the initial noise, by perturbing the conditioning with no objective, and by repelling from the structures already produced. All four run at the same $K = 4$ conditionings and the same 4×180 allocation as CAST and BoN, so every arm in Table 2 sees 720 structures at the same compute to within 7%.

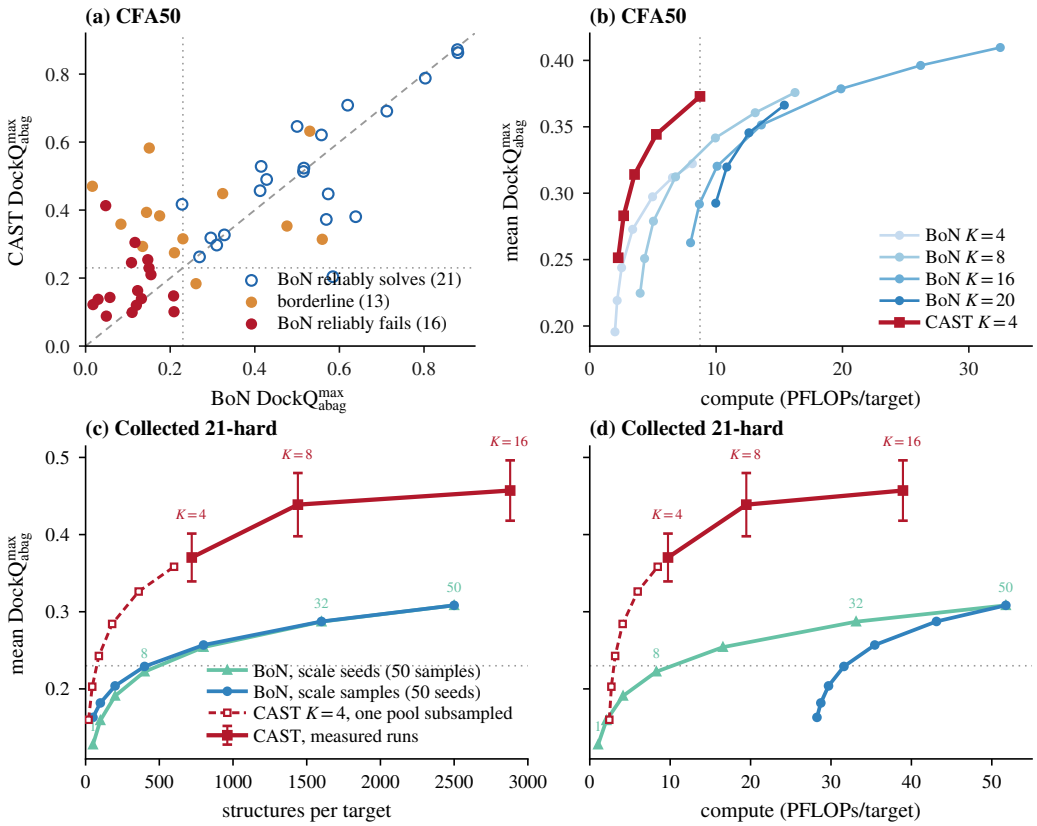


Figure 2: **(a)** Per-target $\mathrm{DockQ}^{\max}_{\mathrm{abag}}$ on CFA50, both arms at 720 structures and the same 4 alignment subsamples, BoN spending it as 4×180. Points above the dashed diagonal are CAST wins, dotted lines mark the CAPRI-acceptable threshold 0.23, and fill encodes the grouping of Table 1. **(b)** The same set against compute, BoN drawn separately at each conditioning count K. **(c)** and **(d)** show the collected 21-hard targets, against structures and against compute. Best-of-N appears twice, scaling conditionings at 50 samples each and scaling samples at 50 conditionings. Every Best-of-N point in (b), (c) and (d) is the exact expectation over subsets of that size (§C.5). Filled squares are separate CAST runs at $K = 4$, 8 and 16 along $N = 180K$, labelled by K and with bars giving the standard error over the 21 targets. Open squares on the dashed line, and CAST's points below 720 in (b), are its single $K = 4$ pool of 720 drawn at fewer particles.

Only CAST converts more failed targets. CAST reaches 36 of the 50 targets, against 30 for FK, 27 for ReD, 26 for BoN, 24 for BP and 22 for NS. Exact McNemar on the paired threshold crossings gives $p = 0.11$, 0.012, 0.013, 0.004 and $< 10^{-3}$ respectively. The separation is sharpest on the 16 BoN reliably fails, where CAST reaches 0.182 and brings 4 to 0.23, against ReD's 0.131 and 1, FK's 0.119 and 1, and none for NS and BP. CAST wins 13 of 16 over ReD ($p = 0.025$). What separates them is transport. FK resamples on the same site reward, so it can only keep particles that already drift toward a candidate. CAST's gradient moves them there. Where BoN already solves, FK is best at 0.540, since it uses one arrival radius to CAST's four, so it holds 48 particles per candidate against 12, and concentration wins where the site is already found. The two cancel, so CAST's lead over FK across all 50 is $+0.036$ with a bootstrap 95% interval of $[-0.003,\ 0.079]$ ($p = 0.104$), while on the 29 BoN does not reliably solve it is $+0.082$ ($p = 0.002$), reaching 16 against 9.

Diversity alone does not explain the gain. BP perturbs the conditioning with no objective, and lands at 0.313 against BoN's 0.322. ReD, which spreads structures apart by repelling from earlier ones, lands at 0.311. Neither is separable from the unsteered control over the 50 paired targets, at $p = 0.92$ and $p = 0.48$, so variety on its own does not account for the gain. NS is weakest of all at 0.238. Searching the initial noise lets it score only finished structures, and by then the site is decided, so a site reward can no longer tell candidates apart. It is at its maximum by the first generation on 46

of the 50 targets, and the rest of the search shrinks the pool without improving it. Over all 50 targets CAST has the highest mean, 0.373 against 0.337, 0.313, 0.311 and 0.238, gains of 10.7%, 19.2%, 19.9% and 56.7%, and reaches 0.23 on 36 targets against 30, 24, 27 and 22.

Table 2: Steered baselines at a matched 720 structures, against CAST and BoN on CFA50. Mean $\mathrm{DockQ}^{\max}_{\mathrm{abag}}$ over all 50 targets and within each group of Table 1, and the number of targets reaching 0.23. Groups are named by BoN, so the column headed *fails* is the subset BoN reliably fails. p is a two-sided Wilcoxon signed-rank test of CAST against that arm on per-target differences.

ARM	**PFLOPs/TGT**	**MEAN** $\mathrm{DockQ}^{\max}_{\mathrm{abag}}$				**REACH** 0.23	p
		all (50)	fails (16)	border (13)	solves (21)		
CAST	8.72	**0.373**	**0.182**	**0.385**	0.511	**36**	
BoN	8.12	0.322	0.111	0.253	0.525	26	0.012
FK	8.69	0.337	0.119	0.279	**0.540**	30	0.104
NS	8.12	0.238	0.076	0.180	0.398	22	$< 10^{-9}$
BP	8.12	0.313	0.103	0.222	0.529	24	0.013
ReD	8.12	0.311	0.131	0.262	0.478	27	0.003

Ablations. Six ablations each change one field of the reference arm at $K = 4$ and $N = 720$, over all 50 CFA50 targets (§C.8, Table 5). Three carry the method. Removing the gradient costs 0.063, shuffling the map 0.054 and collapsing the umbrella sweep 0.012, so the gain is the tilt and the site it names. A separate run per conditioning moves the mean by only 0.006, and its spread across conditionings is ± 0.075, so the pool buys certainty and cost, not accuracy. Resampling instead trades between the groups, gaining $+0.034$ on borderline and paying 0.039 on fails, with no change in the pooled mean. CAST restricts resampling to two early events to limit pool collapse. An oracle map adds 0.111 and takes reach from 36 to 42, the headroom from better maps.

5 Discussion and conclusion

CAST carries a site hypothesis and its conditioning latent in the particle state, so many hypotheses and K subsamples share one pool. On CFA50 at a matched 720 structures it brings 36 of 50 targets to an acceptable pose against 26, and raises mean $\mathrm{DockQ}^{\max}_{\mathrm{abag}}$ from 0.322 to 0.373, a gain of 15.8%. That gain sits on the 29 targets BoN does not reliably solve, and it matches BoN's best allocation of the budget from $1.5\times$ less compute. On collected 21-hard it brings 18 of the 21 to threshold against 9, where BoN's largest pool spends $5.3\times$ the compute and still trails by 20.1%. Raising latents and samples together carries it to 0.457, where Best-of-N flattens. Both datasets are selected using the unsteered model alone, CFA50 on a 200-structure pool and collected 21-hard on a 1,000-seed reference, so difficulty here is a property of the base model.

Two parts of this generalise beyond CAST. The mechanism is a site-measurable reward and a twisted sampler, so any co-folding model with a diffusion decoder can carry the same objective, which the OpenDDE transfer confirms. The pose-invariance guarantee comes from that class of reward, so any site variable on the interface separates exploring sites from selecting poses at one.

CAST does not beat BoN on the targets BoN already solves, where the difference is not separable from zero. This trade-off is by construction. CAST stratifies its pool over the candidate cells and so spends most of it on wrong sites, while BoN spends everything where the base model already goes. Exploring sites is paid for out of the budget that would otherwise exploit one. The pose-invariance result then bounds what steering can reach. CAST moves the antibody to a site and leaves everything after that to the base model, so it cannot correct a pose the base model gets wrong at the true site.

CAST addresses one of co-folding's two failures. It improves site coverage, and what the pool contains bounds the result before ranking. Selection sits one stage beyond, and ranking a pool without a reference remains open. CAST runs on a frozen model and takes a site as an argument, so a hypothesis from an epitope predictor or an experiment turns directly into structures at that site. It needs no training and no reference structure, so testing candidate epitopes costs only a single sampling run. Better epitope prediction and ranking each improve what it returns without changing the method.

AI USE STATEMENT

We used generative AI tools in three ways, all under the authors' direction. For *writing*, they drafted and polished prose, including parts of the abstract, introduction, related work and conclusion. For *retrieval and discovery*, they were used to find related work, and the authors read and verified every paper cited. For *execution*, they acted as a coding assistant for the implementation, for experiment orchestration, and for the analysis code that produces the tables and figures.

The research questions, the hypotheses, the methodology and the claims are the authors' own, as are Proposition 1 and its proof. The authors directed every revision and edited or rejected what the tools produced. Every reported number was generated from the recorded runs by scripts, and the analysis code was read by the authors. We have reviewed all AI-assisted work and take responsibility for the final content of this work.

ETHICS STATEMENT

This work studies structure prediction for antibody–antigen complexes using public PDB data and involves no human subjects and no personally identifying data. Methods improving antibody–antigen modelling are dual-use in the general sense that applies to structural biology, and we do not consider the present contribution to raise concerns beyond that baseline.

REPRODUCIBILITY STATEMENT

The method is defined in §2.2, the map in §2.4 with its construction in §B.7, and the run configuration in §3. Algorithm 1 in §B.1 gives the sampler in full. The guidance schedule and cap are in §B.5, the latent assignment and cell budget in §B.4, the umbrella sweep in §B.6, and the level correction that lets one pool carry K latents in §B.3. Proposition 1 is proved in §B.2 under a single stated hypothesis. Both evaluation sets are built from public structures. CFA50 comes from the CoFold Arena panel (CoFold Arena, 2026), and collected 21-hard from PDB depositions released after 30 June 2025. Both selection rules are stated in §3, with the composition and filters in §C.2. The compute accounting is in §C.1 and the target grouping rule in §C.4. The allocation sweep is in §C.6, the subset expectation in §C.5, the ablations in §C.8 and the map reach in §C.7. Every number in Tables 1–5 therefore traces to a configuration stated in the paper. Every run uses a single frozen, publicly available Protenix-v2 checkpoint, and CAST adds no trained parameters. We will make the code publicly available for the camera-ready version.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016):493–500, 2024. doi: 10.1038/s41586-024-07487-w.

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

Sankar Basu and Björn Wallner. DockQ: A quality measure for protein-protein docking models. *PLoS ONE*, 11(8):e0161879, 2016. doi: 10.1371/journal.pone.0161879.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013. No published version.

Salvatore Candido, Thomas Hayes, Alexander Derry, Roshan Rao, Zeming Lin, Robert Verkuil, Bryan Z. Wu, Jin Sub Lee, et al. Language modeling materializes a world model of protein biology. *bioRxiv*, 2026. doi: 10.64898/2026.06.03.729735.

Chai Discovery, Jacques Boitreaud, Jack Dent, Matthew McPartlon, Joshua Meier, Vinicius Reis, Alex Rogozhnikov, and Kevin Wu. Chai-1: Decoding the molecular interactions of life. *bioRxiv*, 2024. doi: 10.1101/2024.10.10.615955.

Xinshi Chen, Yuxuan Zhang, Chan Lu, Wenzhi Ma, Jiaqi Guan, Chengyue Gong, Jincai Yang, Hanyu Zhang, Ke Zhang, Shenghao Wu, Kuangqi Zhou, Yanping Yang, Zhenyu Liu, Lan Wang, Bo Shi, Shaochen Shi, and Wenzhi Xiao. Protenix – advancing structure prediction through a comprehensive AlphaFold3 reproduction. *bioRxiv*, 2025. doi: 10.1101/2025.01.08.631967.

Hyungjin Chung, Jeongsol Kim, Michael Thompson McCann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=OnD9zGAGT0k.

Joakim Nøddeskov Clifford and Morten Nielsen. Pocket restraints guided by B-cell epitope prediction improve Chai-1 antibody–antigen structure modeling. *Protein Science*, 35(8):e70730, 2026. doi: 10.1002/pro.70730.

CoFold Arena. CoFold Arena leaderboard and scoring methodology. https://cofoldarena.ai/, 2026. Public co-folding benchmark, continuously updated. Methodology page accessed 4 September 2026.

Gabriele Corso, Yilun Xu, Valentin de Bortoli, Regina Barzilay, and Tommi Jaakkola. Particle guidance: non-I.I.D. diverse sampling with diffusion models. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.13102.

James Dunbar and Charlotte M. Deane. ANARCI: antigen receptor numbering and receptor classification. *Bioinformatics*, 32(2):298–300, 2016. doi: 10.1093/bioinformatics/btv552.

Roland L. Dunbrack. Rēs ipSAE loquuntur: What's wrong with AlphaFold's ipTM score and how to fix it. *bioRxiv*, 2025. doi: 10.1101/2025.02.10.637595.

Samuel Fromm, Marko Ludaic, and Arne Elofsson. Evaluating deep learning based structure prediction methods on antibody–antigen complexes. *Bioinformatics*, 42(4):btag136, 2026. doi: 10.1093/bioinformatics/btag136.

Erik Hartman, Jonas Wallin, Johan Malmström, and Jimmy Olsson. Controllable protein design with particle-based Feynman–Kac steering. *arXiv preprint arXiv:2511.09216*, 2025.

Fatima N. Hitawala and Jeffrey J. Gray. What does AlphaFold3 learn about antibody and nanobody docking, and what remains unsolved? *mAbs*, 17(1):2545601, 2025. doi: 10.1080/19420862.2025.2545601.

Purvish Jajal, Nick John Eliopoulos, Benjamin Shiue-Hal Chou, George K. Thiruvathukal, James C. Davis, and Yung-Hsiang Lu. Inference-time alignment of diffusion models via evolutionary algorithms. *arXiv preprint arXiv:2506.00299*, 2025.

Bowen Jing, Bonnie Berger, and Tommi Jaakkola. AlphaFold meets flow matching for generating protein ensembles. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 22277–22303, 2024. arXiv:2402.04845.

Hyeyun Jung, Alan C. Cheng, and BoRam Lee. Boltz-Perturb: Probing generalization in co-folding models via inference-time perturbation. In *ICML 2026 Workshop on Foundations of Deep Generative Models*, 2026. URL https://openreview.net/forum?id=DTrnBoEGID.

Yogesh Kalakoti and Björn Wallner. AFsample2 predicts multiple conformations and ensembles with AlphaFold2. *Communications Biology*, 8:373, 2025. doi: 10.1038/s42003-025-07791-9.

Yogesh Kalakoti and Björn Wallner. AFsample3: Generating and selecting multiple conformational states with AlphaFold3. *bioRxiv*, 2026. doi: 10.64898/2026.01.16.699904.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 26565–26577, 2022. arXiv:2206.00364.

Minji Lee, Colin Kalicki, Minkyu Jeon, Aymen Qabel, Alisia Fadini, and Mohammed AlQuraishi. ConforNets: Latents-based conformational control in OpenFold3. *arXiv preprint arXiv:2604.18559*, 2026.

Minhuan Li, Jiequn Han, Pilar Cossio, and Luhuan Wu. Robust inference-time steering of protein diffusion models via embedding optimization. *arXiv preprint arXiv:2602.05285*, 2026.

Chu'nan Liu, Lilian Denzler, Yihong Chen, Andrew Martin, and Brooks Paige. AsEP: Benchmarking deep learning methods for antibody-specific epitope prediction. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 37, pp. 11700–11734, 2024. arXiv:2407.18184.

Chunan Liu, Aurelien Pelissier, Yanjun Shao, Lilian Denzler, Andrew C. R. Martin, Brooks Paige, and María Rodríguez Martínez. AbRank: A benchmark dataset and metric-learning framework for antibody-antigen affinity ranking. *arXiv preprint arXiv:2506.17857*, 2025.

Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, and Saining Xie. Scaling inference time compute for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2523–2534, 2025. doi: 10.1109/CVPR52734.2025.00241. arXiv:2501.09732, as *Inference-Time Scaling for Diffusion Models beyond Scaling Denoising Steps*.

Raúl Méndez, Raphaël Leplae, Leonardo De Maria, and Shoshana J. Wodak. Assessment of blind predictions of protein-protein interactions: current status of docking methods. *Proteins: Structure, Function, and Bioinformatics*, 52(1):51–67, 2003. doi: 10.1002/prot.10393.

Luis S. Mille-Fragoso, Claudia L. Driscoll, John N. Wang, Haoyu Dai, Talal Widatalla, Jim L. Zhang, Xiaowei Zhang, Bing Rao, Liang Feng, Brian L. Hie, and Xiaojing J. Gao. Efficient generation of epitope-targeted antibodies with Germinal. *Nature Biotechnology*, 2026. doi: 10.1038/s41587-026-03187-0.

Claudio Mirabello and Björn Wallner. DockQ v2: improved automatic quality measure for protein multimers, nucleic acids, and small molecules. *Bioinformatics*, 40(10):btae586, 2024. doi: 10.1093/bioinformatics/btae586.

Badr Moufad, Yazid Janati, Navid Bagheri Shouraki, Alain Oliviero Durmus, Thomas Hirtz, Eric Moulines, and Jimmy Olsson. Efficient zero-shot inpainting with decoupled diffusion guidance. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2512.18365.

Jun Ohnuki and Kei-ichi Okazaki. Enhanced sampling of protein conformations in AlphaFold3 with repulsive bias in the diffusion generative model. *JACS Au*, 2026. doi: 10.1021/jacsau.6c00596. Articles ASAP.

Daniel D. Richman, Jessica Karaguesian, Carl-Mikael Suomivuori, and Ron O. Dror. Unlocking hidden biomolecular conformational landscapes in diffusion models at inference time. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 167662–167682, 2025. arXiv:2512.03312.

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. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 55810–55827, 2025. arXiv:2501.06848.

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, 2026a. doi: 10.1021/acs.jcim.6c02094.

Shosuke Suzuki and Toshiyuki Amagasa. ConforFlux: Particle-guided trunk repulsion for diverse protein conformations. *bioRxiv*, 2026b. doi: 10.64898/2026.05.16.725138.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011. doi: 10.1162/NECO_a_00142.

Yan Wang, Lihao Wang, Yuning Shen, Yiqun Wang, Huizhuo Yuan, Yue Wu, and Quanquan Gu. Protein conformation generation via force-guided SE(3) diffusion models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 56835–56859, 2024. arXiv:2403.14088.

Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Noah Getz, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, Liam Atkinson, Tally Portnoi, Itamar Chinn, Jacob Silterra, Tommi Jaakkola, and Regina Barzilay. Boltz-1: Democratizing biomolecular interaction modeling. *bioRxiv*, 2024. doi: 10.1101/2024.11.19.624167.

Luhuan Wu, Brian L. Trippe, Christian A. Naesseth, David M. Blei, and John P. Cunningham. Practical and asymptotically exact conditional sampling in diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 31372–31403, 2023. arXiv:2306.17775.

Xiangcheng Zhang, Haowei Lin, Haotian Ye, James Zou, Jianzhu Ma, Yitao Liang, and Yilun Du. Inference-time scaling of diffusion models through classical search. In *International Conference on Learning Representations (ICLR)*, 2026a. arXiv:2505.23614.

Yuxuan Zhang, Chengyue Gong, Jinyuan Sun, Jiaqi Guan, Milong Ren, Song Xue, Hanyu Zhang, Wenzhi Ma, Zhenyu Liu, Xinshi Chen, and Wenzhi Xiao. Protenix-v2: Broadening the reach of structure prediction and biomolecular design. *bioRxiv*, 2026b. doi: 10.64898/2026.04.10.717613.

A Extended related work

Two axes organise the work we build on. The first is *what* a method diversifies, the POSE a molecule adopts, meaning its orientation and internal conformation, or the SITE on a partner at which it binds. The second is *where* the intervention enters, the trunk that produces the conditioning representation, which we call *conditioning-side* and which includes restraints, or the diffusion loop that consumes it. Pose work is crowded into both, whereas site-directed work has so far intervened at the model input. Cutting across both axes, inference-time scaling is usually cast as search, a sampler proposing candidates and a verifier ranking them (Ma et al., 2025; Zhang et al., 2026a; Jajal et al., 2025). CAST changes what the sampler proposes by steering and leaves the ranking to a separate verifier (§5).

Pose, conditioning-side. These methods perturb the conditioning and let a frozen decoder follow. Masking or subsampling the alignment weakens the coevolutionary signal that pins the model to one state (Kalakoti & Wallner, 2025; 2026). Rescaling the pair latent by a scalar does the same one layer later (Suzuki & Amagasa, 2026a), and trained channel-wise affine transforms of the pre-Pairformer latent learn a reusable direction instead of a random one (Lee et al., 2026). The first two are gradient-free, which is why alignment subsampling supplies the latents CAST amortises. None of the three can *request* a named change at inference. Two perturb without a destination, and the third fixes its direction before sampling begins.

Pose, diffusion-side. Here the intervention enters the sampler. One group aims at a target it can trust. Li et al. (2026) fits a conditional embedding to experimental measurements, so the data fixes the state it is steering to. Wang et al. (2024) guides with a learned force network carrying physical prior knowledge, so it aims at an equilibrium ensemble rather than one named state. Both targets are correct by construction in a way a predicted epitope is not, and neither of them is a location on the antigen. A second group needs no target at all. Particles are coupled so that they repel, on coordinates (Corso et al., 2024; Ohnuki & Okazaki, 2026) or on trunk embeddings (Suzuki & Amagasa, 2026b). Elsewhere, trajectories are driven *away* from the model's own default prediction, which distorts the marginal sampled. Both buy spread without aim. Richman et al. (2025) is the closest work in method and the furthest in aim. It runs the twisted sequential-Monte-Carlo scheme of Wu et al. (2023) on a co-folding model, under RMSD potentials that repel from the default structure. Retraining under flow matching (Jing et al., 2024) moves the same axis at a training cost we do not pay. CAST also couples particles, but its reward reads the site alone, so the pose at a site is left to the base model (Proposition 1), where an RMSD potential moves it.

Pose diversity is not site diversity. The potentials above are RMSD collective variables read off a superposed structure, Kabsch-aligned and masked to secondary structure in Richman et al. (2025), superposed Cα against earlier seeds in Ohnuki & Okazaki (2026). Such a variable rewards any

departure from the reference, so the mass it moves follows the base model's own ranking of the alternatives. A different site can come out of that, but the potential never chooses which, and it cannot be pointed at a particular epitope. Naming an epitope needs a variable that reads interface placement with the antigen held fixed, which is the site reward of §2.2.

Site, conditioning-side. The closest neighbour supplies the site directly. Clifford & Nielsen (2026) turn a B-cell epitope prediction into a soft pocket restraint on Chai-1 (Chai Discovery et al., 2024) and report better accuracy and diversity. The restraint is an input, fixed before the trunk runs, so each site hypothesis needs its own run. Protenix-v2 (Zhang et al., 2026b) exposes the same interface but ships no weights for its pocket embedder. A restraint is only as good as the epitope it encodes, and prediction is open (Liu et al., 2024). Boltz-Perturb instead injects time-annealed noise into the denoiser's conditioning (Jung et al., 2026). The noise lands on the per-token representation, the pair representation, or both, masked in the source to the pairs touching the ligand. Boltz-Perturb is training-free and establishes that the correct binding-mode basins already exist inside a frozen model, but it was built for small-molecule binding modes and the perturbation carries no objective.

Site, diffusion-side. What exists here is a sampler, not a destination. Twisted Diffusion Sampling steers a diffusion model toward a conditional target with a gradient-guided proposal, importance weights and resampling, and is asymptotically exact (Wu et al., 2023). Feynman–Kac steering reweights trajectories from potentials alone (Singhal et al., 2025), and has since been carried to protein design (Hartman et al., 2025). Both are backbone-agnostic and training-free, and both prescribe *how* to steer, not *what* to steer toward.

B METHOD DETAILS

B.1 THE SAMPLER IN FULL

Algorithm 1 CAST. Every line acts on each particle, whose index is suppressed. Each particle carries a latent k and a cell b. The guided step aims at that cell, the weight corrects the reward level between scoring events, and resampling is confined within cells.

input latents $z_1, \dots, z_K$, candidates $\{\mu_s\}$, radii $\{d_u^\star\}$, pool size N, window $\mathcal{W}$
give each particle a latent k and a cell b, spread evenly over $\mathcal{U} \times \mathcal{S}$ (§B.4)
$x_T \sim \mathcal{N}(0, \sigma_T^2 \mathbb{I})$, $\log w \leftarrow 0$, $g_T \leftarrow 0$
for $t = T, \dots, 1$ **do**
    recenter and randomly rotate x_t, carrying g_t through the same transform
    if $\beta_t > 0$ **then** ▷ the sampler stops churning at low σ_t
        $\varepsilon \sim \mathcal{N}(0, \mathbb{I})$, $\tilde{x}_t \leftarrow x_t + g_t + \beta_t \varepsilon$, $\log w \mathrel{+}= -\big(\|g_t\|^2 + 2\beta_t \langle \varepsilon, g_t \rangle \big)/(2\beta_t^2)$ ▷ equation 4
    else
        $\tilde{x}_t \leftarrow x_t$
    end if
    $\hat{x}_0 \leftarrow D_\theta(\tilde{x}_t, \hat{\sigma}_t \mid z_k)$, $x_{t-1} \leftarrow \tilde{x}_t + (\sigma_{t-1} - \hat{\sigma}_t)(\tilde{x}_t - \hat{x}_0)/\hat{\sigma}_t$
    $g_{t-1} \leftarrow \mathrm{clip}\big[\eta_{t-1} \nabla_{\hat{x}_0} r_b(\hat{x}_0)\big]$ ▷ reused next step, §B.5
    if t is scoring event m **then**
        $\bar{x}_0 \leftarrow$ deterministic 4-step Euler rollout of $\tilde{x}_t$ down the remaining σ ▷ 4 denoiser passes, §C.1
        $r_m \leftarrow r_b(\bar{x}_0)$, $\log w \mathrel{+}= \lambda(r_m - r_{m-1})$ ▷ $r_{-1} := r_0$
        if first event with non-zero reward spread **then**
            solve λ for a target ESS of 0.4, capped, then hold it, $\log w \mathrel{+}= \lambda r_0$ ▷ level correction, §B.3
        end if
        if $t \in \mathcal{W}$: resample each cell's N_b particles among themselves when $\mathrm{ESS}_b < \gamma N_b$, at $\gamma = 0.5$
    end if
end for
return the N structures x_0, ranked by confidence

The sampler reads two different estimates of x_0. Guidance uses the one-step $\hat{x}_0$ the reverse step already produced, which is free. The reward at a scoring event uses $\bar{x}_0$, a deterministic 4-step Euler rollout from the current state to $\sigma \approx 0$ with no churn, because a one-step estimate at mid-σ is still diffuse enough to blur the contact set the reward reads. Eight events at four passes each are the 32 extra denoiser passes charged in §C.1.

At a resampling event the weights reset, and both the difference baseline r^i_{m-1} and the pending displacement are reindexed by ancestry, since a clone continues its ancestor's trajectory and not that of the particle previously in its slot.

The window $\mathcal{W}$ sits early because the site is decided early. Once a particle's site is fixed, resampling can only duplicate or discard particles, not redirect them. By Proposition 1 the reward also says nothing about the pose, which is the only thing still free. Resampling late would therefore replace distinct structures with copies of one another, spending pose diversity and buying nothing.

Three approximations separate this sampler from the ideal target of §2.3, and they are not of the same kind. The plug-in potential and the straight-through Jacobian both act through the displacement g^i_t, which moves the mean of the Gaussian injection and nothing else. Whatever displacement they produce, equation 4 is the exact likelihood ratio for that injection, so the weights price it in full. A poor g^i_t therefore costs proposal quality, and so efficiency, while leaving the target alone. The window is the one approximation that costs exactness rather than efficiency.

We apply weights only inside $\mathcal{W}$ and return the ensemble unweighted, which forfeits the asymptotic exactness of Wu et al. (2023), since that holds only if every weight is carried to the end. Exactness would buy a correctly weighted distribution over the pool, and we never use one. Every number we report is $\mathrm{DockQ}^{\max}$, the best single structure the pool contains. A maximum depends only on whether the pool reaches the right region, not on how much probability mass sits on each member of it.

At the deployed schedule, scoring events fall every 20 of 200 steps from $0.25T$, and $\mathcal{W} = [0.25T, 0.35T]$ admits two. The first cannot be tilted, because the difference potential sets the baseline r^i_0 there, so its spread is zero, λ does not calibrate, and the reward increment and level correction are both skipped. It still resamples on the injection ratio of equation 4 alone, which selects mildly against the particles guidance moved furthest. λ is solved at the second event and then held, so that is the one tilted resample of the run. The tilt itself is carried continuously, the displacement being applied at 159 of 200 steps. Only the steps inside $\mathcal{W}$ carry a weight against it, so the remainder is an unweighted heuristic drift rather than part of the sequential correction.

B.2 Proofs

Proof of Proposition 1. Fix k, and write $Z = \sum_k \rho_k \int p(a \mid z_k) e^{\lambda r(a)} \, \mathrm{d}a$ for the normalising constant of equation 6. By equation 5 and equation 6, $q(k, a, R, c) = Z^{-1} \rho_k \, p(a \mid z_k) \, p(R, c \mid a, z_k) \, e^{\lambda r(a)}$. Marginalising the pose leaves $q(k, a) = Z^{-1} \rho_k \, p(a \mid z_k) \, e^{\lambda r(a)}$, because $p(R, c \mid a, z_k)$ is a normalised conditional and the tilt does not depend on (R, c). Dividing gives the first claim. For the second, $q(k \mid a) \propto \rho_k \, p(a \mid z_k) \, e^{\lambda r(a)}$, and the tilt is a function of a alone, so it cancels in the normalisation over k and $q(k \mid a) = p(k \mid a)$. Hence $q(R, c \mid a) = \sum_k q(k \mid a) \, p(R, c \mid a, z_k) = \sum_k p(k \mid a) \, p(R, c \mid a, z_k) = p(R, c \mid a)$. □

Proof of Corollary 1. The hinged reward r_b is zero on $\mathcal{B}_b$ and strictly negative off it, so $e^{\lambda r_b} \to \mathbf{1}\{a \in \mathcal{B}_b\}$ pointwise as $\lambda \to \infty$. Since $r_b \leq 0$ the family is bounded by 1, so dominated convergence carries the limit through both the product and its normaliser, and the normaliser tends to $p_{\mathrm{mix}}(a \in \mathcal{B}_b)$, which is positive by assumption. Normalising the product of that limit with p_{mix}, or with $p_\theta(\cdot \mid z_k)$ for the per-latent statement, gives the conditioning. The event is measurable with respect to a by Definition 1. Without that positivity the limit is a collapse, not a conditioning, since the tilt would then concentrate on a set the base mixture never visits. □

Proposition 1 constrains q_b, a distribution over finished structures, not the path to one. No step along that path is pose-preserving, and none should be. The pose conditional $p(R, c \mid a)$ depends on a, so a particle moving to a new site must adapt its pose to land on a structure the model recognises. A rigid translation would carry the old site's pose to the new one. Since r_b reads a only through

the separation $a - \mu_s$, a difference of the paratope centroid and the candidate centroid, $\nabla_{\hat{x}_0} r_b$ is non-zero only on the paratope and candidate-epitope atoms, so a step stretches each chain instead of moving it rigidly. The proposition asks nothing of the proposal, and the denoiser repairs the stretch at the next step.

B.3 Consequences of sharing one pool

The potential ratio of equation 4 is taken in *difference* form, each event contributing $\exp(\lambda(r^i_m - r^i_{m-1}))$, because the absolute reward is miscalibrated at high noise where $\hat{x}_0$ is a blurred average. At constant λ the increments telescope to $\exp(\lambda(r^i_M - r^i_0))$ not to $\exp(\lambda r^i_M)$, dividing out each particle's own starting level.

Across multiple latents in one pool, r^i_0 records where the trunk under $z_{k(i)}$ already places the antibody, and dividing it out leaves the cross-latent comparison at chance, with pairwise AUC 0.484 to 0.510 over the eight scoring events. Adding λr^i_0 back once restores the level. At $K = 1$ there is no cross-latent comparison to restore, so the correction has nothing to act on. It is a joint-pool setting rather than a free parameter.

B.4 Latent assignment and the cell budget

For particle index i, with $U = |\mathcal{U}|$ umbrellas and $S = |\mathcal{S}|$ sites, the cell and the latent are

$$b(i) = u(i)\, S + s(i), \quad u(i) = i \bmod U, \quad s(i) = \lfloor i/U \rfloor \bmod S, \qquad k(i) = \big\lfloor i/(US) \big\rfloor \bmod K,$$

a full cross-product of distance against site, so every cell receives all K latents. Umbrella ids are read from the $d^\star$ each particle actually carries rather than from its position, so a sweep repeating a value groups by that value. For a sweep of distinct $d^\star$ this is the $i \bmod U$ above up to relabelling, and only the partition matters downstream. The split is exactly even when K divides the repeat count $N/|\mathcal{U} \times \mathcal{S}|$, which holds at the budget we report, 12 repeats over 60 cells at $K = 4$. Otherwise the first few latents take one extra repeat each, and a comparison across K is confounded with that imbalance.

Two floors bound K. A cell needs about six particles to resample at all, measured at four particles over four umbrellas where the ancestors came back unpermuted at every event, and about three per latent, or one resample can drop a latent by chance. At our budget a cell holds 12, which clears the first floor and allows $12/3 = 4$ latents, the $K = 4$ we report.

B.5 Guidance schedule and cap

The gradient is taken with respect to the denoised *coordinates* of a reward that reads only the site, so it is supported on the paratope and candidate-epitope atoms, with equal and opposite coefficients that draw the two centroids together. Its step is

$$\eta_t = \kappa \, \frac{\sigma_t^2}{\sqrt{1 + \sigma_t^2}},$$

at $\kappa = 75$, the variance-exploding image of the twist coefficient $(1 - \bar{\alpha}_t)/\sqrt{\bar{\alpha}_t}$ that Wu et al. (2023) use on a variance-preserving schedule with signal retention $\bar{\alpha}_t$. It grows with the noise level, so the permitted move is large while the endpoint is still undecided and vanishes as the structure commits, and it is therefore fixed by the schedule and never tuned. The constant κ is the proposal's own strength and is set independently of the potential's λ, so the displacement need not equal the tilted score of equation 3. Nothing is lost by that, since equation 4 is exact for whichever displacement is applied, so a mismatch costs proposal quality while leaving correctness intact. The clip is a per-particle cap of 6 Å, expressed as a root-mean-square displacement per coordinate, so the same setting transfers across targets of different size. The displacement is carried through each step's rigid augmentation, since r is rotation-invariant but ∇r is a vector field and a shift computed in the previous step's frame would point the wrong way after a fresh random rotation.

B.6 What the umbrella sweep corrects

With the Huber knee of ℓ_δ at $\delta = 4$ Å, the reward is linear, not quadratic across most of the pool. Over all cell evaluations of a completed run 94.4% are active ($d > d^\star$), and 93.4% of those lie

beyond δ, where the gradient magnitude is the constant 2δ. Every active particle therefore receives the same push magnitude whatever its distance, so the transport carries direction but no distance information, and a single $d^\star$ cannot be corrected by the pull weakening near arrival. Sweeping u supplies that distance by construction. This predicts a refinement, and A3 (Table 5) matches that shape. The gain is largest where BoN already solves, $+0.021$, and smallest on the targets it fails, $+0.004$, so the sweep adjusts an arrival distance once the site is roughly right.

B.7 Building the candidate map

Pool modes. For each structure in an unsteered pool we superpose the antigen onto a common reference. A contact residue is one whose Cα lies within 10 Å of any antibody Cα, and each structure contributes the centroid of its own contacts. Modes are then found greedily on that cloud. We seed at the centroid with the most neighbours inside one 12 Å epitope footprint, claim every centroid within it, and repeat. A mode's patch is the set of residues contacted by at least half of the structures assigned to it, and its score is its population share. We keep the top three, since the first carries most of the pool mass but is not reliably the right site, and the second and third recover many of the targets where it is wrong. Each mode takes its consensus independently, instead of assigning every residue to its strongest mode. That matters because modes overlap at their rims, and stripping a shared rim shifts the weaker centroid off the site.

Surface tiles. Tiles come from a farthest-point covering of the antigen. We seed at the residue farthest from the antigen centroid, claim everything unclaimed within a radius, and repeat. That radius is binary-searched to a fixed tile count and is not fixed in advance. A fixed radius makes the tile count scale with antigen size, so a large antigen produces far more tiles than the candidate budget S can hold, and truncating drops most of its surface. Fixing the count instead keeps the budget per target constant, and simply makes the tiles coarser on a large antigen. No surface filter is applied, since dropping the inner shell assumes the epitope is exposed, and groove-binding epitopes would then fall in no tile at all.

The hybrid. Modes are listed first and kept verbatim, because they come from real poses and are better centred, whereas tiles are comprehensive but arbitrarily centred. Every tile is then appended, ranked by the fraction of its residues carrying mode support. That ranking uses only the pool's own mode mass, so it is ground-truth free, and it concentrates budget sharply because most tiles have no mode support at all. Ties break by a stable sort on the tiling's own order. That matters because the map is truncated to the candidate budget S, so an arbitrary tie-break would silently change which sites receive particles. Patches are finally de-duplicated, and any patch of fewer than 3 residues is dropped, since a smaller one has no meaningful centroid. At a 6 Å footprint the hybrid then reaches more targets than either ingredient alone.

C Experimental details

C.1 Compute accounting

Deferring the decoder to the K-th subsample leaves the first $K-1$ to return one unsteered draw each, so a joint run emits $N+K-1$ structures while its pool is N. We exclude those $K-1$ from every metric and still charge all K trunk passes, so the method pays for them in cost and collects nothing from them in quality. The exclusion changes no reported number. Across 71 CAST units no such draw has ever held its unit's maximum, and the closest any came was 0.014 below it.

Reported FLOPs are per target, averaged over all targets, and three terms enter the count. The **trunk** is charged once per conditioning and stays outside the structure count, which is why arms at the same 720 structures differ in compute, 8.12 PFLOPs at $K=4$ against 15.40 at $K=20$. The **denoiser** is charged per structure, CAST's at its measured $1.16\times$ a BoN structure's, 232 passes against 200, the excess being the mini-rollout at each scoring event, so the gradient CAST adds is paid for and never assumed free. The **confidence head** is charged once per finished structure and every arm pays it alike.

The pilot pool the map needs. The map's modes are read from an unsteered pool, which a new target must have or pay for. Ours are built from 150 structures drawn from an unsteered BoN pool.

Every CAST figure we report is the steered run alone, on targets whose pilot pool already existed and was reused. A fresh target pays for those 150 structures at the BoN per-structure rate of 0.0088 PFLOPs, so a fixed 1.31 PFLOPs, taking CAST from 8.72 to 10.04 PFLOPs per target on CFA50. The trunk is not charged twice, since the pilot draws the same $K = 4$ conditionings. Charged, CAST's margin over the 20×36 rung is $1.5\times$ rather than $1.8\times$, and its 0.373 still exceeds the best BoN allocation at or below 10.04 PFLOPs, 0.342 at 8×90 (Table 3, Figure 2(b)). CAST's advantage on hard cases, in both performance and scaling, survives the charge.

C.2 Dataset composition and filters

CFA50 is screened from the CoFold Arena panel as described in §3. The 80 candidates entering that screen span 58 Fab and 22 VHH, 177–1024 tokens (median 583) and 79 antigens, and the 50 retained span 38 Fab and 12 VHH, 177–880 tokens (median 554).

Collected 21-hard is curated from PDB depositions instead, so it carries its own structural filters. A deposition is kept when it is an X-ray or cryo-EM structure below 3.5 Å and its variable domains type under IMGT numbering with ANARCI. An antibody–antigen interface then counts when the two chains share a heavy-atom contact within 5 Å over at least three antigen residues, and more than half the paratope falls inside the IMGT CDRs. The 21 that survive the narrowing span 15 Fab and 6 VHH, 177–1155 tokens (median 545).

C.3 What a deeper unsteered pool still misses

Figure 3 characterises the failure this paper targets on a deeper pool. Every point is one target under the unsteered base model at 20 alignment subsamples $\times$ 50 diffusion samples, so 1,000 structures, $1.4\times$ the 720 used for the matched-budget comparisons.

The x-axis is how far apart the pool places the antibody, the median pairwise distance between epitope centroids after each structure is put in a common antigen frame. The y-axis is the best DockQ in the pool, i.e. $\mathrm{DockQ}^{\max}$. A target below the dotted line has no acceptable structure at all, so no verifier applied to that pool can return one.

Three readings matter. First, 18 of the pools sit below the line even at 1,000 structures, so the failure is not a budget that was too small. Second, the marker shape splits those 18 by what limits the best structure, and on 7 it recovers no native contact at all, five of them with not one draw in a thousand reaching the true epitope. That is a site failure, not a geometry one. Third, the two sides of the dashed line separate two different site failures. Targets at the left, such as 9r0j_nb1 and 9hls_nb11 near 2 Å, place the whole pool on one wrong patch, the redundant binding mode of Clifford & Nielsen (2026). Targets at the right, such as 9rxi_nb2 and 9jno_ab1 above 15 Å, spread across more than one epitope footprint and still miss. Spread alone therefore does not fix a site, which is the case for a named destination, not more diversity.

C.4 CFA50 target grouping

Let p be the fraction of 4-conditioning draws whose $\mathrm{DockQ}^{\max}_{\mathrm{abag}}$ reaches 0.23, at the same 4×180 allocation the comparison uses. A draw takes 4 of the target's conditionings and all 180 samples of each, so p is exact over all $\binom{20}{4} = 4845$ subsets. Then *reliably fails* is $p \leq 0.10$, *reliably solves* is $p \geq 0.90$ and *borderline* the rest, giving 16, 13 and 21 targets.

C.5 The exact expectation over subsets

A rung that takes m of the $n = 180$ samples from each of K conditionings could be drawn many ways, and its $\mathrm{DockQ}^{\max}_{\mathrm{abag}}$ depends on which. We average over every draw exactly, instead of picking one.

Fix a target and let $v_1 < \cdots < v_M$ be the distinct $\mathrm{DockQ}_{\mathrm{abag}}$ values in its pool. Let $n_k(v)$ count the samples of conditioning k at or below v. The maximum is at or below v only if every drawn sample

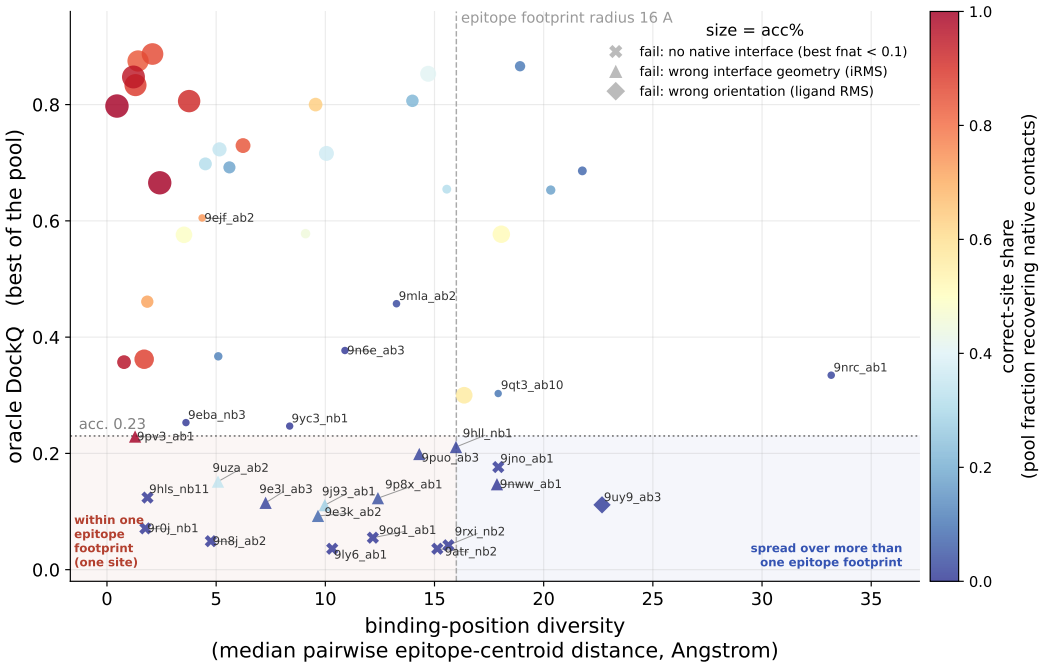


Figure 3: Binding-position spread against the best DockQ in the pool, for the collected 52 targets under the unsteered base model at 1,000 structures each (20 subsamples $\times$ 50 samples). Colour is the fraction of the pool recovering native contacts, marker area the fraction reaching DockQ ≥ 0.23, and marker shape what limits the best structure. The legend names the three failure shapes, a cross for no native interface (fnat < 0.1), a triangle for wrong interface geometry (iRMS) and a diamond for wrong orientation (ligand RMS), and a circle marks a target with no such limit. Points below the dotted line hold nothing acceptable, so selection cannot help there.

is, and the conditionings are drawn independently, so

$$F(v) \;=\; \Pr\left[\mathrm{DockQ}^{\max}_{\mathrm{abag}} \leq v\right] \;=\; \prod_{k=1}^{K} \binom{n_k(v)}{m} \Big/ \binom{n}{m},$$

$$\mathbb{E}\left[\mathrm{DockQ}^{\max}_{\mathrm{abag}}\right] \;=\; \sum_{j=1}^{M} v_j \big(F(v_j) - F(v_{j-1})\big),$$

with $F(v_0) = 0$ and $\binom{a}{b} = 0$ for $a < b$. The sum is finite, so the rung carries no sampling error. At $m = n$ the formula returns the pool maximum. The scaling curves of Figure 2(c,d) draw which conditionings to use instead, keeping every sample of each. The same argument applies with $F(v) = \binom{C(v)}{c} \Big/ \binom{C}{c}$, where c of the C conditionings are drawn and $C(v)$ counts those whose own maximum is at or below v.

C.6 BoN allocation sweep

A matched budget can still be spent badly, and BoN in particular benefits more from seed scaling, so we ask what the best use of 720 unsteered structures is. Table 3 holds the budget fixed and varies only the split. More seeds/conditionings is monotonically better for BoN, worth $+0.044$ ($+13.7\%$) from 4 to 20. Table 1 compares against the 4×180 arm because it matches CAST's $K = 4$ conditioning count. The higher-K rungs score higher at more compute (Figure 2b). The rungs up to 20 conditionings come from each target's own ladder at the matched 720 structures.

Table 3: Allocating a fixed 720-structure budget for BoN, mean $\mathrm{DockQ}^{\max}_{\mathrm{abag}}$ over the 50 CFA50 targets. The structure budget is matched and the compute budget is not, since the trunk runs once per conditioning, so PFLOPs rise with K at a constant 720 structures. Both CAST and BoN use the same trunk term, 1.82 PFLOPs at $K = 4$. Each rung uses a fixed conditioning set, the K lowest indices. The 4×180 rung takes every sample of its 4 conditionings and is exact. Where a rung takes only part of a conditioning we report the exact expectation of the maximum over all such subsets, a product of hypergeometric terms per observed value (§C.5). A fixed subset such as the first $720/K$ samples would bias the rung, because sample index tracks confidence rank and the confidence-best structures are the DockQ-worst.

ARM	$K\times$ **SAMPLES**	**PFLOPs/TGT**	**MEAN** $\mathrm{DockQ}^{\max}_{\mathrm{abag}}$
BoN	4×180	8.12	0.322
BoN	8×90	9.94	0.342
BoN	16×45	13.58	0.351
BoN	20×36	15.40	**0.366**
CAST	$K = 4$, $N = 720$	8.72	**0.373**

Table 4: Reach of the at most $S = 15$ candidate map we ship, over the 50 CFA50 targets, counting targets where *some* candidate centroid lands within δ of the true epitope centroid, both taken on the same structure. Ground truth scores reach and never builds the map. The two ingredient rows are the map's own modes and tiles, scored identically, so the hybrid row is their combination and not a separate construction.

MAP	< 6 **Å**	< 12 **Å**	< 14 **Å**	**MEDIAN (Å)**
Pool modes only	18	31	31	9.39
Surface tiles only	25	44	48	5.89
Hybrid, modes + ranked tiles	**32**	**48**	**49**	**4.59**

C.7 Candidate map reach

The shipped map is the hybrid, pool modes first and then surface tiles at a 12-tile budget, ranked by their overlap with those modes. Its closest candidate is a median 4.59 Å from the true epitope (Table 4). Modes and tiles fail on different targets, which is why the hybrid uses both. Scored apart, the tiles carry the coverage and the modes the centring. Tiles alone reach 25 targets within 6 Å against the modes' 18, and adding the modes takes the hybrid to 32 and the median from 5.89 to 4.59 Å. The hybrid is closer than tiles alone on 26 of the 50 targets and further on 3.

Figure 4 separates a map failure from a method failure. Every one of the 32 targets whose map lands within 6 Å achieves non-zero site reach, so a good map always puts some of the pool on the true site. Reach reports only that the site was visited. The pose found there must also be acceptable, and 7 of those 32 fall short of 0.23. The targets to the right of the line fail for the other reason, that the map itself misses.

The map does not cap where CAST arrives. On one target CAST produced a structure 4.3 Å from the true epitope through a map whose best candidate was 12.3 Å out, while BoN never came within 10 Å in 720 draws. The tilt supplies a direction, and diffusion supplies the rest.

C.8 Ablations

Each ablation changes one field of the reference arm, holding $K = 4$ and $N = 720$. **A1** replaces the joint pool with a separate model per conditioning, the arrangement TDS (Wu et al., 2023) uses. Each conditioning gets its own 720-particle pool, so A1 draws 4×720 structures where CAST draws 720. Splitting one 720 budget four ways would leave three particles per cell instead of twelve, so the separate arrangement pays per conditioning to keep the site sweep. **A2** removes the gradient and **A3** collapses the umbrella sweep to $d^\star = 12$ Å. **A4** substitutes the oracle map, the true epitope read from the ground-truth complex. That leaves one candidate per target in place of the fifteen, so every

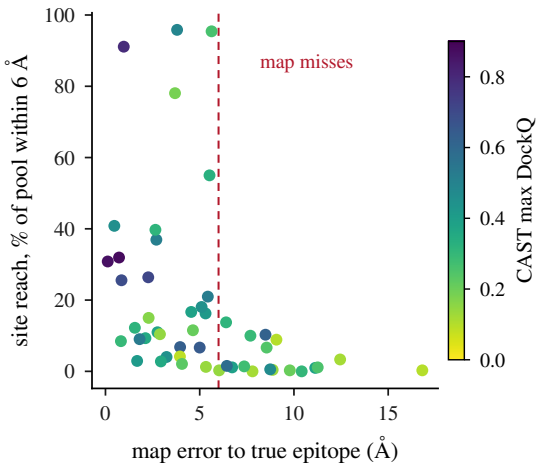


Figure 4: Site reach, the share of the pool whose predicted epitope centroid lands within 6 Å of the true one, against map error, the distance from the best map candidate to the true epitope. Colour is CAST $\mathrm{DockQ}^{\max}_{\mathrm{abag}}$, darker is higher. The dashed line at 6 Å is where the map is counted as missing the epitope.

particle is aimed at the right site. **A5** substitutes a site-shuffled map and **A6** disables resampling. All run on all 50 CFA50 targets.

Table 5: Ablations at $K = 4$, $N = 720$ over all CFA50 targets, split by the groups of Table 1. A1 is the mean over its four conditionings with the per-target standard deviation beneath, a spread CAST avoids by pooling all four into one run.

ARM	**CHANGE**	**MEAN** $\mathrm{DockQ}^{\max}_{\mathrm{abag}}$ all (50)	fails (16)	border (13)	solves (21)	**REACH** 0.23
CAST	reference	0.373	0.182	0.385	0.511	36
A1	separate pools	0.367	0.193	0.397	0.481	33.2
	sd	±0.075	±0.057	±0.107	±0.068	
A2	no gradient	0.310	0.127	0.200	0.518	28
A3	one radius	0.361	0.178	0.378	0.490	35
A4	oracle map	**0.484**	**0.345**	**0.447**	**0.614**	**42**
A5	shuffled map	0.319	0.167	0.249	0.477	28
A6	no resampling	0.373	0.221	0.351	0.502	38

The joint pool buys robustness and cost. A1 matches CAST on every group, so the level comes from the tilt, which both arms share. What the joint pool supplies is certainty and cost. A separate run commits to one conditioning, and its spread across conditionings is ±0.075, where pooling them moves the mean by only 0.006. The draw dominates by an order of magnitude. On 33 of 50 targets it decides whether such a run matches CAST. On 15 of 50 it decides whether the run returns an acceptable structure at all. CAST returns the pooled result every time, from one pool where the separate arrangement needs K, at 8.72 PFLOPs against about 29 (§2.2).

The gradient and the map act where the site is uncertain. The gradient is worth +0.185 (+92.5%) on borderline and +0.055 (+43.3%) on fails, the candidate map against shuffled (A5) +0.136 (+54.6%) and +0.015 (+9.0%). Each carries 8 of the pooled reach gain, 6 of it on borderline. The gradient also supplies 3 of the 4 fails targets CAST reaches at all, where the candidate map supplies 1. On the targets BoN already solves both are small, −0.007 (−1.4%) and +0.034 (+7.1%).

Resampling trades the fails group for the borderline one. Resampling concentrates the pool where the reward already scores well. It gains +0.034 (+9.6%) on borderline and +0.009 (+1.7%)

on solves. It pays 0.039 (-17.5%) on fails, reaching 4 of 16 there against 7 without it. The three cancel in the pooled mean, which a paired test over the 50 cannot separate from zero ($p = 1.00$). FK, which resamples without the gradient, makes the same trade (Table 2). CAST therefore resamples early and only twice, with a single tilted event. That takes the borderline gain while leaving the fails group a pool that is still exploring.

The umbrella sweep helps consistently and by a small amount. 4 radii beat one in all three group means and in reach, 36 targets against 35, and on 29 of the 50 targets, the largest gain being $+0.021$ ($+4.2\%$) where BoN already solves. No group is harmed. The pooled gain of $+0.012$ ($+3.3\%$) is small enough to trade away when the budget is tight, since one umbrella holds the same 12 particles per cell at $N = 180$ that four need $N = 720$ to reach. We keep the sweep on the consistency of its sign.

The map sets the ceiling without capping it. A4 substitutes the oracle map and adds $+0.163$ ($+89.6\%$) on fails, taking reach from 36 to 42, more than any change to the search. CAST's ceiling is set by where it is told to look. The map is not a bound, though. The shipped one reaches 32 of the 50 targets within 6 Å and 48 within 12 Å, yet CAST solves targets whose best candidate is further out, so the tilt biases the search while leaving it free to depart (§C.7).

C.9 Transfer to a second co-folding model

CAST is backbone-agnostic by construction, so the question is whether the gain survives a change of base model. We repeat the matched-budget comparison on **OpenDDE** (Aureka AI OpenDDE project, 2026), an independently trained all-atom co-folding model with an antibody-antigen fine-tuned checkpoint, over 40 difficulty-balanced targets from the public PDB. *No hyperparameter is retuned*. The guidance scale and cap, the umbrella sweep, the resampling scheme and window, the level correction and the 720 structures as 4×180 are the values CFA50 uses, and the map is built by the same reference-free construction. What changes is the base model, which is the experiment, so its weights are its own and the map is a different instance over a different target set. Targets are grouped by BoN's own reliability exactly as in §C.4, p exact over all $\binom{20}{4} = 4845$ four-conditioning draws, giving 16 reliably-failing, 7 borderline and 17 reliably-solving targets.

At the matched budget CAST reaches 0.454 against BoN's 0.372 (Table 6), a difference of $+0.082$, and brings 27 of the 40 targets to an acceptable pose against 20. The gain sits where it does on CFA50. On the 16 targets BoN reliably fails it is $+0.111$, improving 16 of 16 and taking 6 targets across the threshold from none. Where BoN already solves, the difference is $+0.011$ and a paired test cannot separate it from zero. The borderline group is 7 targets, too few to carry a test of its own, and its mean rises $+0.187$ where its median rises $+0.087$, with two targets supplying most of that.

Table 6: $\mathrm{DockQ}^{\max}_{\mathrm{abag}}$ over the 40 OpenDDE40 targets for CAST and BoN, both at a *matched* budget of 720 structures and the same 4 conditionings, BoN spending it as 4×180. The counterpart of Table 1, with the grouping built the same way (§C.4) from BoN alone. SITE is as in Table 1. REACH 0.23 counts targets attaining $\mathrm{DockQ}^{\max}_{\mathrm{abag}} \geq 0.23$, CAST WINS is a paired count spanning both rows, and p is a two-sided Wilcoxon signed-rank test on the per-target differences within that subset.

SUBSET	ARM	MEAN $\mathrm{DockQ}^{\max}_{\mathrm{abag}}$	SITE (Å)	REACH 0.23	CAST WINS	p
All targets (40)	CAST	**0.454**	**1.75**	**27**	29	$< 10^{-4}$
	BoN	0.372	6.41	20		
BoN reliably fails (16)	CAST	**0.211**	**3.07**	**6**	16	$< 10^{-4}$
	BoN	0.101	13.88	0		
Borderline (7)	CAST	**0.372**	**1.75**	**4**	6	0.11
	BoN	0.186	2.87	3		
BoN reliably solves (17)	CAST	0.715	**0.51**	17	7	0.98
	BoN	0.704	0.85	17		

The compute axis separates the two backbones. On CFA50 CAST matches BoN's best allocation of the budget, whereas here it exceeds every allocation at 11.68 PFLOPs per target, and BoN does

not close the gap by spreading the same budget over more conditionings. Even at 20×36, $5\times$ the conditionings and $2\times$ the compute at 23.43 PFLOPs per target, BoN reaches 0.389 against CAST's 0.454 (Table 7, Figure 5b).

Table 7: Allocating a fixed 720-structure budget on OpenDDE40, mean $\mathrm{DockQ}^{\max}_{\mathrm{abag}}$ over the 40 targets, the counterpart of Table 3. Conditionings are the K lowest indices and partial rungs are the exact expectation over sample subsets (§C.5).

ARM	$K\times$ **SAMPLES**	**PFLOPs/TGT**	**MEAN** $\mathrm{DockQ}^{\max}_{\mathrm{abag}}$
BoN	4×180	11.12	0.372
BoN	8×90	14.20	0.377
BoN	16×45	20.36	0.389
BoN	20×36	23.43	0.389
CAST	$K = 4$, $N = 720$	11.68	**0.454**

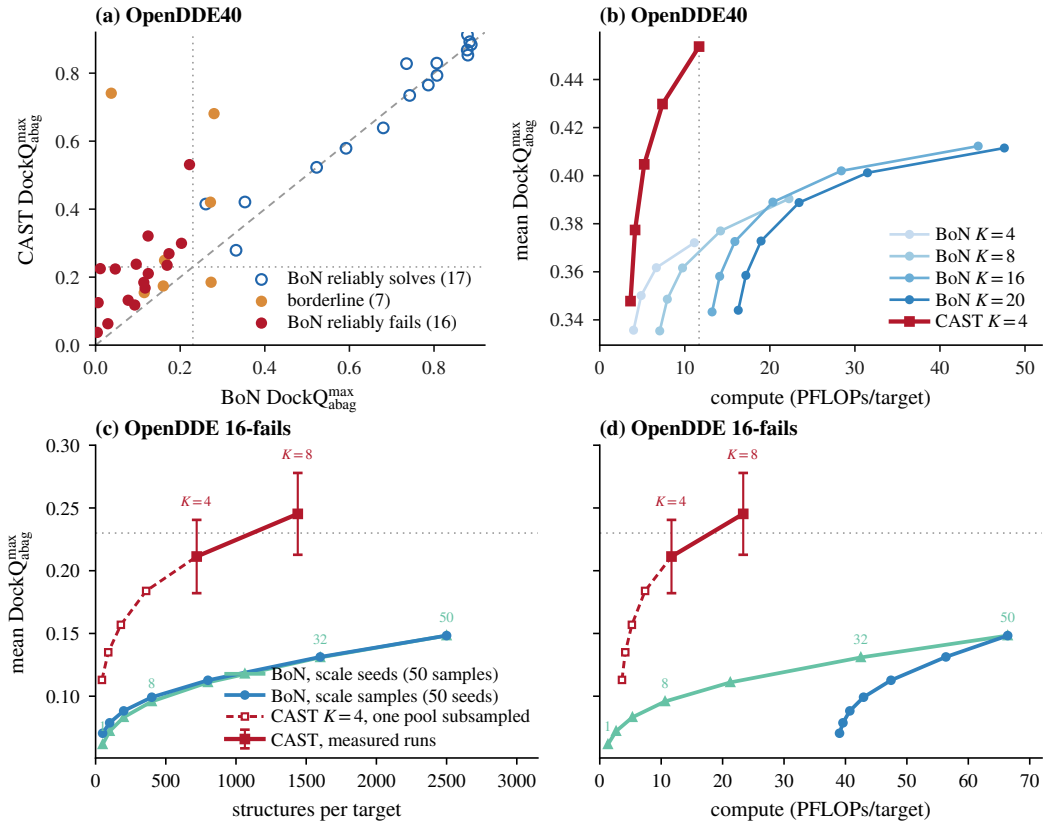


Figure 5: OpenDDE40, the counterpart of Figure 2. **(a)** Per-target $\mathrm{DockQ}^{\max}_{\mathrm{abag}}$, both arms at 720 structures and the same 4 conditionings, fill encoding the grouping. **(b)** The same set against compute, BoN drawn at each conditioning count. **(c)** and **(d)** are the scaling ladder on the 16 targets BoN reliably fails, against structures and compute, drawn as in Figure 2(c,d). Best-of-N appears twice, scaling conditionings at 50 samples each and scaling samples at 50 conditionings, each the exact expectation over subsets of that size. Filled squares are measured CAST runs at $K = 4$ and $K = 8$ along $N = 180K$, with bars giving the standard error over the 16 targets. Open squares on the dashed line are the $K = 4$ pool of 720 drawn at fewer particles.

On the 16 targets BoN reliably fails, CAST's smallest rung already beats the largest Best-of-N pool we ran (Figure 5(c,d)). CAST reaches 0.211 at 720 structures and crosses the 0.23 acceptable line at 1440 with 0.245, a step of $+0.034$ at its standard error, where Best-of-N climbs logarithmically along either axis to only 0.148 at 2500 structures and never reaches it. Compute widens the gap,

since Best-of-N pays a trunk pass per conditioning and needs 66.4 PFLOPs for that 0.148 while CAST spends 11.7 for 0.211. No Best-of-N allocation on either axis reaches CAST at any budget in the figure. This reproduces the pattern Protenix shows in Figure 2(c,d), where CAST's $K = 4$ rung also passes Best-of-N's largest pool at a fifth of the compute.