SEARCH-BASED INFERENCE-TIME SCALING FOR ALL-ATOM PROTEIN BINDER DESIGN

Anonymous authors
Paper under double-blind review

ABSTRACT

All-atom protein binder design with diffusion models follows a costly generate-and-filter pipeline in which candidates are sampled, redesigned, and refolded; for difficult targets, fewer than 5% of samples survive. Rather than scaling by generating more candidates, we apply diffusion inference-time search methods—steering computation within the denoising trajectory toward structures that are more likely to be designable. We introduce lightweight rewards based on confidence-head predictions and geometric-decoding/inverse-folding self-consistency, evaluated on intermediate denoised estimates without running the full refolding pipeline. Protein-specific adaptations, including noise-level gating, adaptive quantile thresholds, and failure-scaled exploration noise, address the unique failure modes of all-atom diffusion intermediates. On six difficult binder targets, our method reaches equivalent prefold confidence with 5–$10\times$ fewer function evaluations and increases the throughput of designable binders by 4–$8\times$.

1 INTRODUCTION

Diffusion-based generative models have become a leading approach for *de novo* protein design (Watson et al., 2023; Yim et al., 2023; Trippe et al., 2022; Anand & Achim, 2022). Recent methods directly generate all-atom coordinates, enabling tighter coupling between backbone geometry and sidechain placement (Chu et al., 2024; Krishna et al., 2024; Qu et al., 2024; Butcher et al., 2025). BoltzGen (Stark et al., 2025) encodes amino-acid identity geometrically, keeping sampling entirely in continuous coordinate space and enabling a single architecture that unifies design with structure prediction (Abramson et al., 2024; Zambaldi et al., 2024). In practice, binder design with Stark et al. (2025) follows an iterative workflow: sample all-atom structures, redesign sequences via conditional inverse folding (Dauparas et al., 2022), refold with Boltz-2 (Passaro et al., 2025), and filter on confidence metrics and physical measures (ΔSASA, hydrogen bonds). We say a sample is *designable* if it passes this downstream pipeline of refolding and filtering. Empirically, only a small fraction of samples are designable for difficult targets, defined here as those where fewer than 5% of baseline samples pass the pipeline.

Inference-time scaling (Ma et al., 2025) offers a principled alternative by allocating extra test-time compute within the denoising trajectory, guided by lightweight ***prefold*** **rewards**—evaluated on intermediate denoised estimates before the costly refolding stage—to preferentially complete promising partial samples. Recent work casts diffusion sampling as a search problem over intermediate denoised estimates $\hat{\mathbf{x}}_{0|s}$, using particle-based (Singhal et al., 2025; Hartman et al., 2025; Kim et al., 2025), tree-search (Guo et al., 2025; Jain et al., 2025; Zhang et al., 2025), value-based (Li et al., 2024), and noise-space (Ma et al., 2025; Lee et al., 2025) strategies. However, applying these ideas to all-atom protein generation introduces domain-specific challenges: at high noise levels, denoised estimates can decode to `UNK` residues that yield meaningless rewards; inexpensive prefold signals correlate only noisily with post-refold designability; and multimodality demands adaptive exploration strategies.

We show that inference-time search strategies, when combined with lightweight prefold rewards and protein-specific adaptations, can substantially improve the yield and throughput of all-atom binder design, avoiding the computational bottleneck of full refolding during generation. Our contributions are as follows:

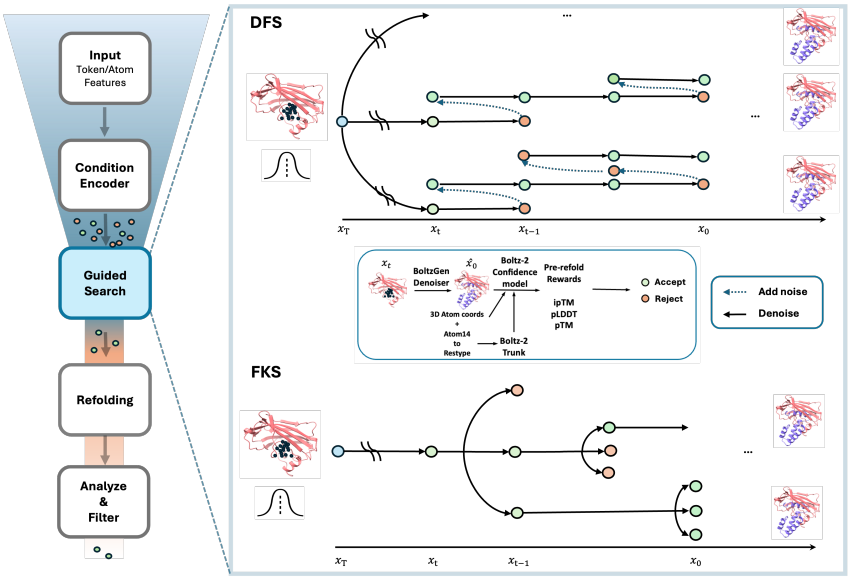

Figure 1: **Inference-time scaling for All-atom Binder Design.** Refolding every generated sample ($20 - 80$ s / sample) is costly, with cost scaling superlinearly in the sample count and average design size. Our method scores intermediate denoised estimates $\hat{\mathbf{x}}_{0|s}$ via a lightweight *prefold* reward, reserving expensive refolding only for accepted candidates. **Center:** Decoded residue identities and coordinates from $\hat{\mathbf{x}}_{0|s}$ are passed through the Boltz-2 trunk and confidence head to yield prefold confidence scores. Accepted/rejected intermediates are evaluated at sparse denoising checkpoints. **Top:** DFS resamples among parallel particles at each checkpoint. **Bottom:** FKS backtracks to earlier diffusion steps when scores fall below an adaptive threshold.

1. **Prefold steering signals.** We introduce two lightweight prefold rewards—confidence-head scores and a geometric-decoding/inverse-folding agreement signal—that guide search without running the full refolding pipeline.
2. **Protein-specific search adaptations.** We adapt existing Feynman–Kac steering (FKS) and depth-first search (DFS) methods with noise-level/`UNK`-count gating, adaptive quantile-based thresholds, and failure-scaled exploration noise to handle the unique failure modes of all-atom denoised estimates.
3. **Systematic efficiency gains.** Across difficult binder targets, DFS dominates the quality–compute Pareto frontier. By filtering low-quality trajectories before refolding, search-based steering yields up to $8\times$ higher end-to-end throughput per wall-hour even without a dilated schedule.

2 Method

Overview. We treat Stark et al. (2025)'s design-mode sampling as a search over the reverse-diffusion trajectory (Figure 1). Let $s \in \{0, \dots, T-1\}$ index diffusion steps with decreasing noise scales $\{\sigma_s\}$. At each step, the network produces a denoised estimate $\hat{\mathbf{x}}_{0|s}$ from a noisy atom14 state $\mathbf{x}_s$. Given conditioning context $\mathcal{C}$ (target structure and design specification), we define a *prefold* reward $r_s = \phi(\hat{\mathbf{x}}_{0|s}, \mathcal{C})$—evaluated on $\hat{\mathbf{x}}_{0|s}$ *before* the costly inverse-folding and refolding stages—to reallocate compute toward trajectories likely to yield designable structures. We use the standard schedule from Abramson et al. (2024) to isolate the contribution of search from schedule engineering; the two are complementary (§A.6).

2.1 Inference-Time Scaling Framework

We view denoising as a fixed-depth tree whose nodes are noisy states $\mathbf{x}_s$ and edges are reverse transitions $\tilde{p}_\theta(\mathbf{x}_{s+1} \mid \mathbf{x}_s, \mathcal{C})$ (Zhang et al., 2025). A best-of-N (BoN) baseline samples N inde-

pendent trajectories and keeps the best final sample; the methods below improve on this by scoring intermediate estimates and adapting compute online.

BFS via FKS. Because Stark et al. (2025) encodes residue identity geometrically, our rewards operate directly on denoising intermediates without ProteinMPNN or PyRosetta (Hartman et al., 2025). Following Singhal et al. (2025); Moral (2004), we maintain K particles and score each via log-potentials $\ell_s^{(k)}$:

$$w_s^{(k)} = \frac{\exp(\ell_s^{(k)})}{\sum_j \exp(\ell_s^{(j)})}, \qquad \mathrm{ESS}_s = \left(\sum_k (w_s^{(k)})^2 \right)^{-1}. \tag{1}$$

Resampling is triggered when $\mathrm{ESS}_s/K < \rho_{\min}$, duplicating high-weight and discarding low-weight trajectories; a temperature τ_s controls selection pressure.

DFS with adaptive backtracking. DFS denoises a single trajectory and evaluates r_s at sparse checkpoints. If the score falls below a step-dependent threshold δ_s and backtracking budget $B > 0$ remains, the algorithm re-noises to $s' = \max(0, s - \Delta)$ and resumes (Zhang et al., 2025). A buffer stores scored states; when the budget is exhausted, DFS falls back to the best previously observed state, yielding instance-adaptive compute.

2.2 Prefold Reward Functions

Running the full evaluation pipeline at every guidance step is prohibitively expensive. We instead define lightweight rewards on $\hat{\mathbf{x}}_{0|s}$.

Self-consistency reward. Stark et al. (2025) provides two independent readouts of residue identity from $\hat{\mathbf{x}}_{0|s}$: (i) geometric decoding via marker-atom placement, and (ii) IF prediction conditioned on backbone geometry. Let $\mathcal{D}$ denote design positions, a_i^{marker} the marker-decoded amino acid at position i, and $\mathcal{D}_{\text{valid}} = \{i \in \mathcal{D} : a_i^{\text{marker}} \neq \texttt{UNK}\}$. We measure agreement as:

$$r_{\text{SC}}^{(s)} = \frac{1}{|\mathcal{D}_{\text{valid}}|} \sum_{i \in \mathcal{D}_{\text{valid}}} \log p_{\text{IF}}\big(a_i^{\text{marker}} \;\big|\; \hat{\mathbf{x}}_{0|s}\big)\,. \tag{2}$$

This steers toward structures whose backbone naturally implies the same residues that the marker atoms encode, so the downstream IF step largely preserves the generated sequence. More information on Appendix A.5.

Confidence reward. We leverage the confidence module from Boltz-2 (Passaro et al., 2025) to estimate interface quality directly from $\hat{\mathbf{x}}_{0|s}$ via a four-stage pipeline (Figure 1, center inset): (i) decode residue identities from marker atoms, (ii) update sequence-dependent features with the decoded identities and generated coordinates, (iii) run the pairformer trunk, and (iv) query the confidence head to obtain *prefold* ipTM, pTM, and pLDDT. We use global iPTM as the default steering signal, which captures predicted interface accuracy between designed and target chains. To further reduce overhead, we decrease the number of recycling steps in the trunk cutting per-evaluation cost by approximately 30–40%.

Reward validation. Prefold iPTM is the strongest predictor of post-refold quality (r=0.39–0.69; Appendix A.1); hard sequence recovery correlates only weakly, motivating the soft formulation in Eq. 2.

2.3 Protein-Specific Adaptations

Structure-quality gating. Following Stark et al. (2025), we concentrate guidance in the $\tau \in [0.6, 0.8]$ window where residue types crystallize, skipping evaluation at high noise or when too many design-region residues decode as UNK.

Adaptive thresholds (DFS). Reward scales vary across targets, making fixed thresholds brittle. We instead set step-dependent thresholds from a quantile of the per-step score history, with forced acceptance after repeated failures to prevent budget exhaustion (details in Appendix A.2).

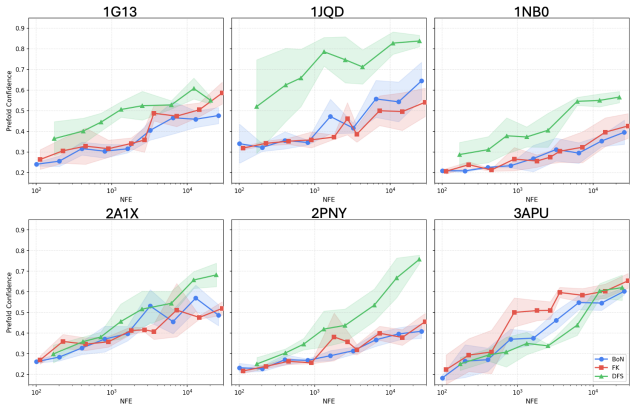

Figure 2: **Quality–compute Pareto Frontier.** Prefold confidence vs. number of function evaluations for six targets. DFS (green) dominates on five of six targets; FKS (blue) is competitive at high budgets. Shaded regions: mean ± s.d. across 10 seeds.

Exploration noise (DFS). At low noise, backtracking yields near-identical trajectories. We inject noise proportional to the number of consecutive failures at a step, scaled by the current noise level $\sigma_{s'}$, to encourage trajectory diversity (Eq. 4, Appendix A.2).

3 Results

3.1 Experimental Setup

We evaluate on six difficult binder targets introduced in Stark et al. (2025): GM2A (`1G13`), HNMT (`1JQD`), RFK (`1NB0`), PHYH (`2A1X`), IDI2 (`2PNY`), and ORM2 (`3APU`); none have a close homolog (>30% sequence identity) in a bound context in the PDB. Binder length is held constant across methods for each target. For pipeline efficiency test, we generate 20,000 baseline samples and compare against FKS (4,000 samples) and DFS (3,000 samples), both using the confidence reward (prefold iPTM) as the default steering signal (full configurations in Appendix A.3). All methods are evaluated through the full pipeline: inverse folding, refolding with Boltz-2, and filtering on post-refold metrics (iPTM > 0.5, pTM > 0.75). To trace Pareto curves, we scale compute budget via particle count (K) for FKS and backtracking budget (B) for DFS.

3.2 Scaling and Efficiency

DFS dominates the quality–computation frontier. Figure 2 plots prefold confidence against the number of function evaluations of diffusion model and reward models for all six targets. DFS dominates the Pareto frontier on five of six targets, with the largest gains on the most difficult targets (`1JQD`, `2PNY`) where baseline pass rates are lowest and adaptive backtracking redirects the most computation. FKS is competitive at high budgets but its fixed per-particle cost limits scaling efficiency.

Pipeline efficiency. Search-based steering yields throughput gains by filtering low-quality trajectories before refolding (Figure 3). FKS generates only ∼600–1,500 unique trajectories due to low sampling temperature, cutting refolding attempts roughly in half. DFS produces substantially more designable binders per hour on the hardest targets: 55/hour on `1JQD` (3.2× baseline) and 8/hour on `1NB0` (8× baseline).

3.3 Steering Analysis and Ablations

Prefold confidence tracks post-refold quality. Figure 3 (top) shows that prefold and post-refold iPTM distributions track closely: samples steered toward high prefold confidence consistently achieve higher post-refold scores, validating our choice of steering signal (Appendix A.1).

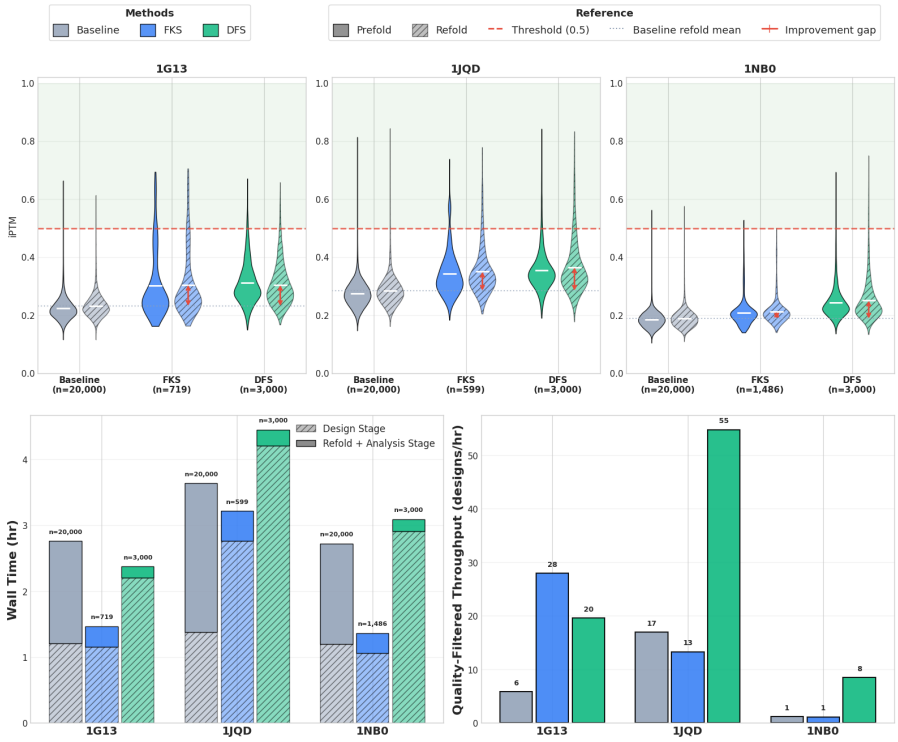

Figure 3: **Top:** Prefold (solid) and post-refold (hatched) iPTM distributions for three difficult targets. Dashed red line: iPTM $= 0.5$ designability threshold. **Bottom left:** Wall-time breakdown; search methods spend more on design (dashed) but drastically cut refolding cost (solid). **Bottom right:** End-to-end throughput of designable binders per wall-hour.

Self-consistency steering improves inverse folding compatibility. SC guidance shifts all three IF metrics favorably: sequence recovery improves from 0.578 to 0.645 (+11.6%), mean IF log-likelihood increases from -2.661 to -2.472 (+7.1%), and IF entropy decreases from 1.585 to 1.505 (-5.0%), with all differences significant ($p < 10^{-30}$; more explanation in Appendix A.5).

Search and schedule optimization compose. Under Stark et al. (2025)'s dilated schedule, combining confidence and SC rewards yields iPTM gains of up to +53.8% and RMSD reductions of up to 76.0%, confirming orthogonal gains (Appendix A.6). Both methods are robust to hyperparameter choices: compute budget is the dominant factor, while other hyperparameters show broad near-optimal plateaus (Appendix A.4).

4 Conclusion

We presented an inference-time scaling framework for all-atom protein binder design that steers diffusion sampling toward designable structures before the costly refolding stage. Our approach introduces two lightweight prefold rewards—confidence-head scoring and a self-consistency signal based on geometric-decoding/inverse-folding agreement—together with protein-specific adaptations that address failure modes unique to all-atom denoised estimates. Across six difficult binder targets, DFS with adaptive backtracking dominates the quality–compute Pareto frontier and yields up to $8\times$ higher throughput of designable binders on the hardest targets. The two reward signals provide complementary benefits, and search-based gains compose with schedule optimization, confirming that inference-time search is orthogonal to improvements in the underlying generative process.

Limitations and future work. Rewards for physical properties(binding free energy or aggregation propensity) will be filled in full paper. Extending the framework to multi-state or specificity-constrained design—where the reward must penalize off-target binding—is a natural next step.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

Namrata Anand and Tudor Achim. Protein structure and sequence generation with equivariant denoising diffusion probabilistic models. *arXiv preprint arXiv:2205.15019*, 2022.

Jasper Butcher, Rohith Krishna, Raktim Mitra, Rafael I Brent, Yanjing Li, Nathaniel Corley, Paul T Kim, Jonathan Funk, Simon Mathis, Saman Salike, et al. De novo design of all-atom biomolecular interactions with rfdiffusion3. *bioRxiv*, 2025.

Alexander E Chu, Jinho Kim, Lucy Cheng, Gina El Nesr, Minkai Xu, Richard W Shuai, and Po-Ssu Huang. An all-atom protein generative model. *Proceedings of the National Academy of Sciences*, 121(27):e2311500121, 2024.

Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022.

Yingqing Guo, Yukang Yang, Hui Yuan, and Mengdi Wang. Training-free guidance beyond differentiability: Scalable path steering with tree search in diffusion and flow models. *arXiv preprint arXiv:2502.11420*, 2025.

Erik Hartman, Jonas Wallin, Johan Malmström, and Jimmy Olsson. Controllable protein design through feynman-kac steering. *arXiv preprint arXiv:2511.09216*, 2025.

Vineet Jain, Kusha Sareen, Mohammad Pedramfar, and Siamak Ravanbakhsh. Diffusion tree sampling: Scalable inference-time alignment of diffusion models. *arXiv preprint arXiv:2506.20701*, 2025.

Sunwoo Kim, Minkyu Kim, and Dongmin Park. Test-time alignment of diffusion models without reward over-optimization. *arXiv preprint arXiv:2501.05803*, 2025.

Rohith Krishna, Jue Wang, Woody Ahern, Pascal Sturmfels, Preetham Venkatesh, Indrek Kalvet, Gyu Rie Lee, Felix S Morey-Burrows, Ivan Anishchenko, Ian R Humphreys, et al. Generalized biomolecular modeling and design with rosettafold all-atom. *Science*, 384(6693):eadl2528, 2024.

Gyubin Lee, Bao N Nguyen Truong, Jaesik Yoon, Dongwoo Lee, Minsu Kim, Yoshua Bengio, and Sungjin Ahn. Adaptive inference-time scaling via cyclic diffusion search. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Xiner Li, Yulai Zhao, Chenyu Wang, Gabriele Scalia, Gokcen Eraslan, Surag Nair, Tommaso Biancalani, Shuiwang Ji, Aviv Regev, Sergey Levine, et al. Derivative-free guidance in continuous and discrete diffusion models with soft value-based decoding. *arXiv preprint arXiv:2408.08252*, 2024.

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

Pierre Moral. *Feynman-Kac formulae: genealogical and interacting particle systems with applications*. Springer, 2004.

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

Wei Qu, Jiawei Guan, Rui Ma, Ke Zhai, Weikun Wu, and Haobo Wang. P (all-atom) is unlocking new path for protein design. *bioRxiv*, pp. 2024–08, 2024.

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

Hannes Stark, Felix Faltings, MinGyu Choi, Yuxin Xie, Eunsu Hur, Timothy O'Donnell, Anton Bushuiev, Talip Uçar, Saro Passaro, Weian Mao, et al. Boltzgen: Toward universal binder design. *bioRxiv*, pp. 2025–11, 2025.

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

Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

Jason Yim, Brian L Trippe, Valentin De Bortoli, Emile Mathieu, Arnaud Doucet, Regina Barzilay, and Tommi Jaakkola. Se (3) diffusion model with application to protein backbone generation. *arXiv preprint arXiv:2302.02277*, 2023.

Vinicius Zambaldi, David La, Alexander E Chu, Harshnira Patani, Amy E Danson, Tristan OC Kwan, Thomas Frerix, Rosalia G Schneider, David Saxton, Ashok Thillaisundaram, et al. De novo design of high-affinity protein binders with alphaproteo. *arXiv preprint arXiv:2409.08022*, 2024.

Xiangcheng Zhang, Haowei Lin, Haotian Ye, James Zou, Jianzhu Ma, Yitao Liang, and Yilun Du. Inference-time scaling of diffusion models through classical search. *arXiv preprint arXiv:2505.23614*, 2025.

A Additional Figures and Tables

A.1 Reward Validation

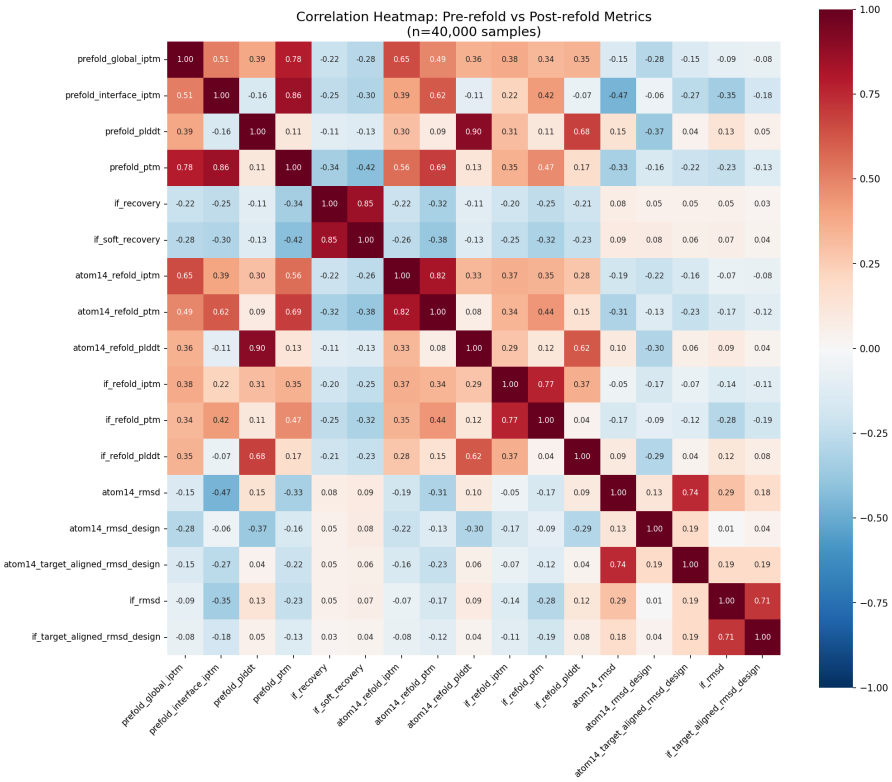

Figure 4: **Correlation heatmap** (n=40,000)**.** Prefold iPTM/pTM are strong predictors of post-refold quality.

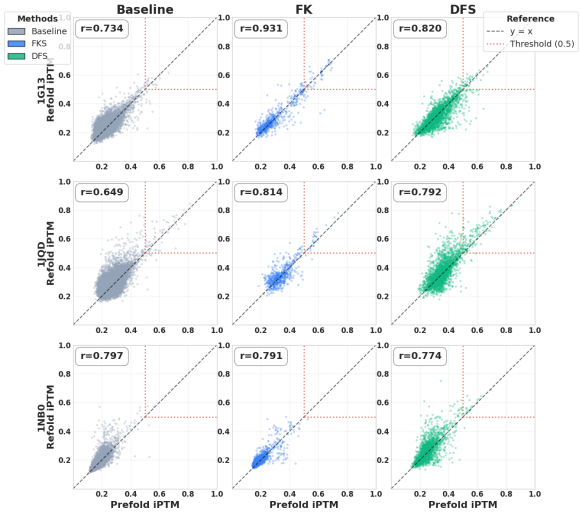

Figure 5: **Prefold vs. post-refold iPTM.** Validates prefold confidence as a steering signal.

A.2 PROTEIN-SPECIFIC ADAPTATION DETAILS

Structure-quality gating. We skip reward evaluation when (i) $\sigma_s > \sigma_{\max}$ ($\sigma_{\max} = 20$, i.e., $\tau \approx 0.6$), or (ii) more than $k = 5$ design-region residues decode as UNK.

Adaptive thresholds (DFS). We set thresholds from step-specific score history:

$$\delta_s = \text{quantile}_q\Big(\{r_1^{(s)}, \ldots, r_m^{(s)}\}\Big) + \epsilon, \tag{3}$$

with $q = 0.25$, $\epsilon = 0.05$, falling back to a global window before sufficient history accumulates. After $M = 5$ consecutive failures at a step, we force acceptance to prevent budget exhaustion.

Exploration noise (DFS). We inject noise proportional to consecutive failures:

$$\mathbf{x}'_{s'} = \mathbf{x}_{s'} + \min(\gamma \cdot n_{\text{fail}},\ \gamma_{\max}) \cdot \sigma_{s'} \cdot \boldsymbol{\epsilon}', \quad \boldsymbol{\epsilon}' \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{4}$$

with $\gamma = 0.1$ and $\gamma_{\max} = 1.0$; n_{fail} resets upon acceptance.

A.3 EXPERIMENTAL CONFIGURATIONS

For FK steering, we use K=64 particles with guidance activated when $\sigma_s < \sigma_{\max}$, temperature τ=0.01, and immediate potentials. For DFS, we run 64 trajectories with backtracking budget B=100, rollback depth Δ=20, and adaptive thresholds with q=0.25. To trace Pareto curves (Figure 2), we scale FK by increasing K and DFS by increasing B; curves are reported for all six targets. Wall-time efficiency (Figure 3) is measured on three representative targets (1G13, 1JQD, 1NB0).

Copy

A.4 PARAMETER SENSITIVITY

We perform one-at-a-time ablations over all hyperparameters for both search methods, holding remaining parameters at their defaults (Appendix A.3), with each setting evaluated across 10 random seeds. Figure 6 ranks parameters by iPTM range (max − min across settings) and Figure 7 shows the full ablation grids. For FKS, the dominant factor is particle count K (iPTM range $\sim$0.20), consistent with the theoretical dependence of SMC estimator variance on particle count (Moral, 2004); remaining standard parameters—resampling frequency (Singhal et al., 2025), potential mode, temperature, and resampling method—have comparatively modest effects (iPTM range $\leq$ 0.08). We note that *immediate* potentials outperform the *difference* potentials recommended by Singhal et al. (2025), likely because prefold confidence is already well-calibrated as an absolute quality indicator. Our protein-specific addition, guidance start $\sigma_{\max}$ (§2.3), shows moderate sensitivity, confirming that gating reward evaluation to the residue-crystallization window (Stark et al., 2025) is important but not fragile.

For DFS, the compute-controlling parameters—recursion depth and backtracking budget B (Zhang et al., 2025)—are the most sensitive (iPTM range $\sim$0.11–0.12), mirroring the dominance of particle count in FKS. Our protein-specific adaptations—adaptive window, adaptive percentile threshold, and adaptive margin (§2.3)—show moderate sensitivity (iPTM range $\sim$0.05–0.09), while standard parameters (step size, start step) are comparatively stable. In both methods, non-compute hyperparameters exhibit broad plateaus of near-optimal performance, indicating that practitioners can adopt the default configurations with minimal tuning effort.

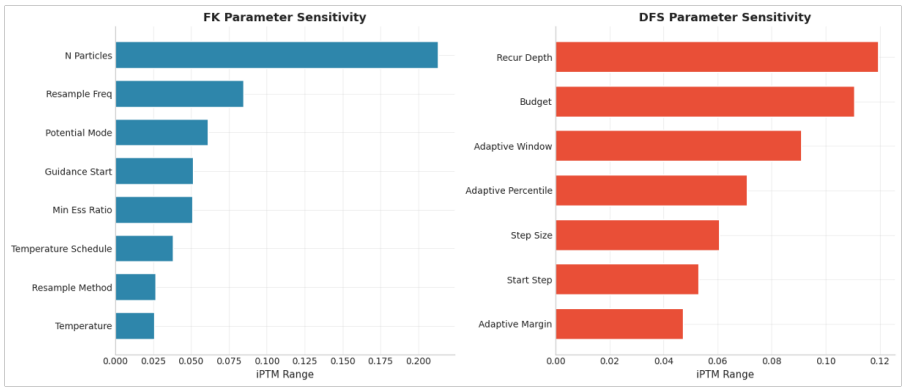

Figure 6: **Sensitivity ranking.** iPTM range for FKS (left) and DFS (right). Compute-controlling parameters (particle count, budget, recursion depth) dominate; non-compute hyperparameters show broad near-optimal plateaus.

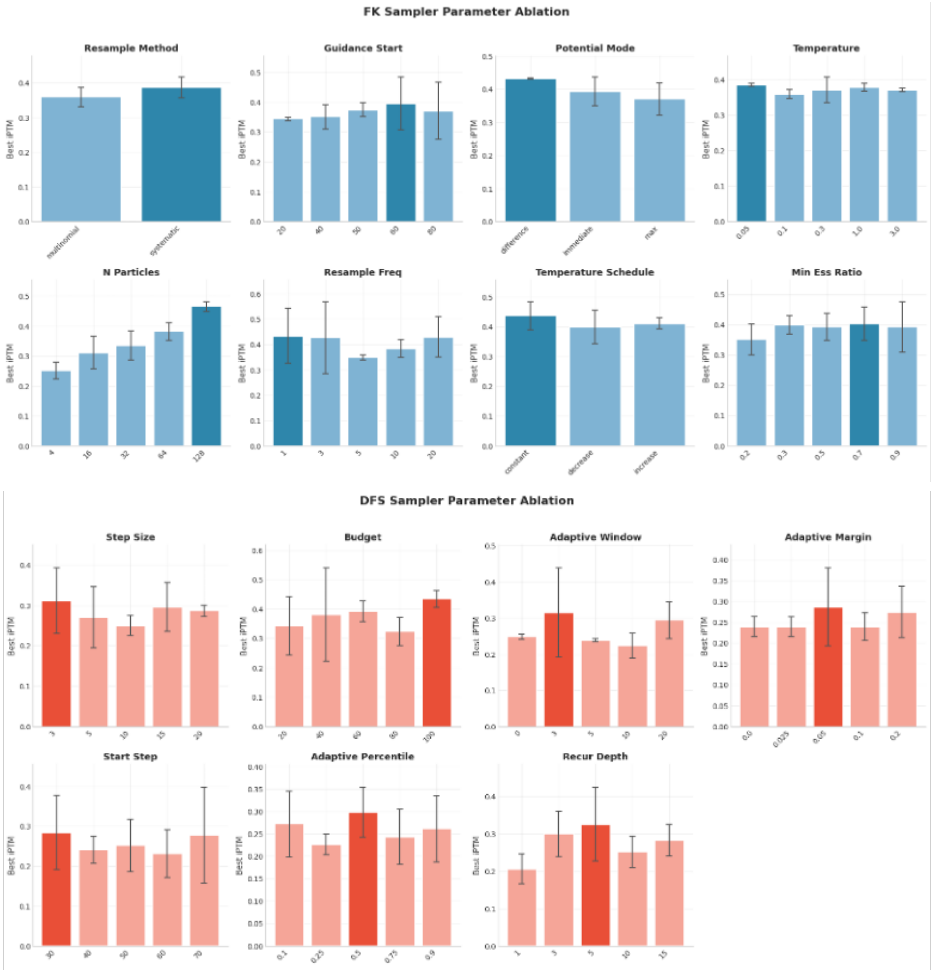

Figure 7: **Full parameter ablations.** Best iPTM per setting (FKS top, DFS bottom). Error bars: mean $\pm$ s.d. across 10 seeds.

A.5 Self-Consistency Steering

The self-consistency (SC) reward (Eq. 2) is designed to steer generation toward structures whose backbone geometry naturally implies the same residue identities encoded by marker atoms. The motivation is to improve compatibility with downstream inverse folding: if the generated backbone already "wants" the designed sequence, the IF model will preserve it rather than proposing alternatives, leading to higher sequence recovery and more consistent folding. We evaluate SC steering using FKS (K=64 particles) on four targets and measure its effect on both IF metrics and post-refold quality. As shown in Table 1, SC steering improves sequence recovery by 5–9 percentage points across all targets, and these gains propagate downstream: RMSD drops consistently, indicating that self-consistent sequence–structure pairs refold more faithfully.

Metric definitions.

- Sequence Recovery (Seq. Rec.): Fraction of design positions where the IF model's most-likely amino acid matches the marker-decoded residue, measuring hard designability.
- Log-Likelihood (Log-Lik.): Mean log-probability assigned by the IF model to the marker-decoded amino acid at each design position, providing a soft measure of sequence–structure consistency.
- Entropy: Shannon entropy of the IF predicted distribution over amino acids, averaged across design positions; lower values indicate more confident, structurally constrained predictions.
- Max-Prob: Average maximum probability across amino acids at each design position, quantifying the IF model's per-residue prediction confidence.

Table 1: **Effect of SC steering on inverse folding and post-refold metrics.** SC steering improves IF compatibility (left) and downstream quality (right) across most targets. RMSD refers to all-atom RMSD. Means ± std reported; best values bolded.

		Inverse Folding Metrics				Post-Refold Metrics			
Target	Method	Seq. Rec. ↑	Log-Lik. ↑	Entropy ↓	Max-Prob ↑	iPTM ↑	Design pTM ↑	RMSD (Å) ↓	Min PAE ↓
1G13	Baseline	0.578±0.086	−2.661±0.287	1.585±0.159	0.500±0.056	0.470±0.125	0.599±0.132	2.935±2.997	5.428±3.089
	SC Guided	**0.645±0.066**	**−2.472±0.136**	**1.505±0.102**	**0.525±0.041**	**0.496±0.100**	**0.650±0.085**	**1.650±1.656**	**4.504±2.292**
1JQD	Baseline	0.544±0.095	−2.825±0.351	1.692±0.222	0.469±0.078	0.432±0.217	0.662±0.134	5.895±6.676	9.335±7.201
	SC Guided	**0.623±0.053**	**−2.610±0.236**	**1.616±0.107**	**0.490±0.033**	**0.491±0.186**	**0.727±0.052**	**3.718±5.377**	**7.120±5.812**
1NB0	Baseline	0.544±0.082	−3.295±0.983	**1.331±0.428**	**0.581±0.139**	0.342±0.112	0.680±0.098	3.468±3.054	8.052±3.600
	SC Guided	**0.647±0.059**	**−2.760±0.314**	1.441±0.114	0.543±0.040	**0.346±0.098**	**0.710±0.042**	**2.685±2.365**	**7.496±3.118**
2A1X	Baseline	0.595±0.100	**−4.382±1.636**	0.886±0.586	0.727±0.183	0.281±0.192	0.737±0.153	12.40±9.11	14.36±6.04
	SC Guided	**0.688±0.062**	−4.419±0.698	**0.550±0.270**	**0.831±0.084**	**0.341±0.208**	**0.819±0.061**	**8.628±9.043**	**12.49±5.95**

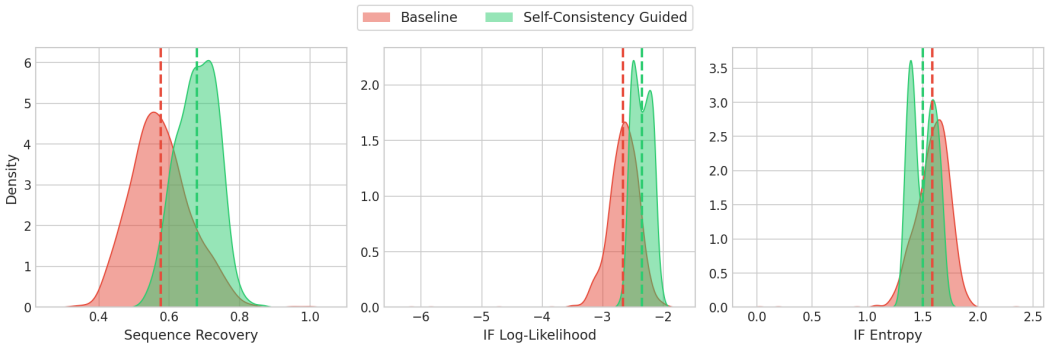

Figure 8: **SC steering shifts IF metrics favorably.** Density plots for 1G13 comparing baseline (red) and SC-guided (green) samples. SC steering increases sequence recovery and log-likelihood while reducing entropy.

A.6 DILATED SCHEDULE RESULTS

The main experiments (§2.1) use the standard AlphaFold3 noise schedule to isolate the contribution of search-based steering from schedule engineering. Here we verify that inference-time search remains beneficial when combined with BoltzGen's default dilated schedule, which already concentrates function evaluations in the critical $\tau \in [0.6, 0.8]$ window. Figure 9 shows that all three FKS reward variants (confidence(Conf), self-consistency(SC), and their combination(Conf+SC)) shift the joint iPTM–RMSD distribution toward the high-confidence, low-RMSD region relative to the dilated-schedule baseline, confirming that the two sources of improvement are complementary rather than redundant. Table 2 quantifies the gains: the combined reward (Conf+SC) yields the largest improvements across most metrics and targets, with iPTM gains of up to +53.8% (2A1X) and RMSD reductions of up to 76.0% (1JQD).

Table 2: **FKS under BoltzGen dilated schedule.** Mean post-refold metrics for three FKS reward variants versus baseline on 1JQD, 1G13, 1NB0, and 2A1X. Conf: confidence reward (prefold iPTM); SC: self-consistency reward (Eq. 2); Conf+SC: weighted combination of both. Δ columns report relative improvement (%) over baseline. Best result per metric is **bolded**.

			Method			Δ (%) vs. Baseline		
Target	Metric	Baseline	Conf	SC	Conf+SC	Conf	SC	Conf+SC
1JQD	RMSD (Å) ↓	5.89	1.83	3.72	**1.41**	+69.0	+36.9	**+76.0**
	$\text{RMSD}_{\text{design}}$ (Å) ↓	4.17	1.93	2.82	**1.61**	+53.8	+32.5	**+61.4**
	$\text{RMSD}_{\text{target}}$ (Å) ↓	0.80	0.59	0.61	**0.53**	+26.6	+23.7	**+33.4**
	iPTM ↑	0.432	0.591	0.491	**0.617**	+37.0	+13.7	**+42.9**
	pTM ↑	0.640	0.679	0.653	**0.694**	+6.0	+1.9	**+8.4**
	Design pTM ↑	0.662	0.722	0.727	**0.770**	+9.0	+9.9	**+16.3**
	Target pTM ↑	0.710	0.715	0.711	**0.730**	+0.7	+0.1	**+2.8**
1G13	RMSD (Å) ↓	2.94	1.43	1.65	**1.28**	+51.3	+43.8	**+56.2**
	$\text{RMSD}_{\text{design}}$ (Å) ↓	3.50	1.50	1.86	**1.37**	+57.1	+46.7	**+60.8**
	$\text{RMSD}_{\text{target}}$ (Å) ↓	0.97	0.87	**0.85**	0.85	+10.7	**+12.4**	+11.9
	iPTM ↑	0.470	0.551	0.496	**0.576**	+17.1	+5.4	**+22.6**
	pTM ↑	0.633	0.665	0.640	**0.670**	+4.9	+1.0	**+5.7**
	Design pTM ↑	0.599	0.684	0.650	**0.707**	+14.1	+8.5	**+17.9**
	Target pTM ↑	0.729	**0.735**	0.727	0.735	**+0.8**	−0.4	+0.7
1NB0	RMSD (Å) ↓	3.47	2.73	2.69	**1.94**	+21.3	+22.6	**+44.1**
	$\text{RMSD}_{\text{design}}$ (Å) ↓	2.92	2.38	1.95	**1.79**	+18.4	+33.2	**+38.7**
	$\text{RMSD}_{\text{target}}$ (Å) ↓	1.12	1.10	1.05	**0.91**	+1.9	+5.8	**+18.5**
	iPTM ↑	0.342	0.331	0.347	**0.389**	−3.4	+1.2	**+13.6**
	pTM ↑	0.538	0.524	0.536	**0.555**	−2.5	−0.3	**+3.1**
	Design pTM ↑	0.680	0.692	**0.710**	0.704	+1.7	**+4.4**	+3.5
	Target pTM ↑	0.658	0.645	0.654	**0.661**	−2.0	−0.7	**+0.5**
2A1X	RMSD (Å) ↓	12.40	6.88	8.63	**4.40**	+44.5	+30.4	**+64.5**
	$\text{RMSD}_{\text{design}}$ (Å) ↓	3.85	1.94	2.34	**1.91**	+49.7	+39.3	**+50.4**
	$\text{RMSD}_{\text{target}}$ (Å) ↓	0.56	0.56	0.55	**0.54**	−0.2	+0.2	**+2.0**
	iPTM ↑	0.281	0.391	0.341	**0.432**	+39.2	+21.2	**+53.8**
	pTM ↑	0.618	0.645	0.628	**0.651**	+4.4	+1.7	**+5.5**
	Design pTM ↑	0.738	0.825	0.819	**0.840**	+11.8	+11.1	**+13.9**
	Target pTM ↑	0.719	**0.724**	0.716	0.722	**+0.8**	−0.4	+0.4

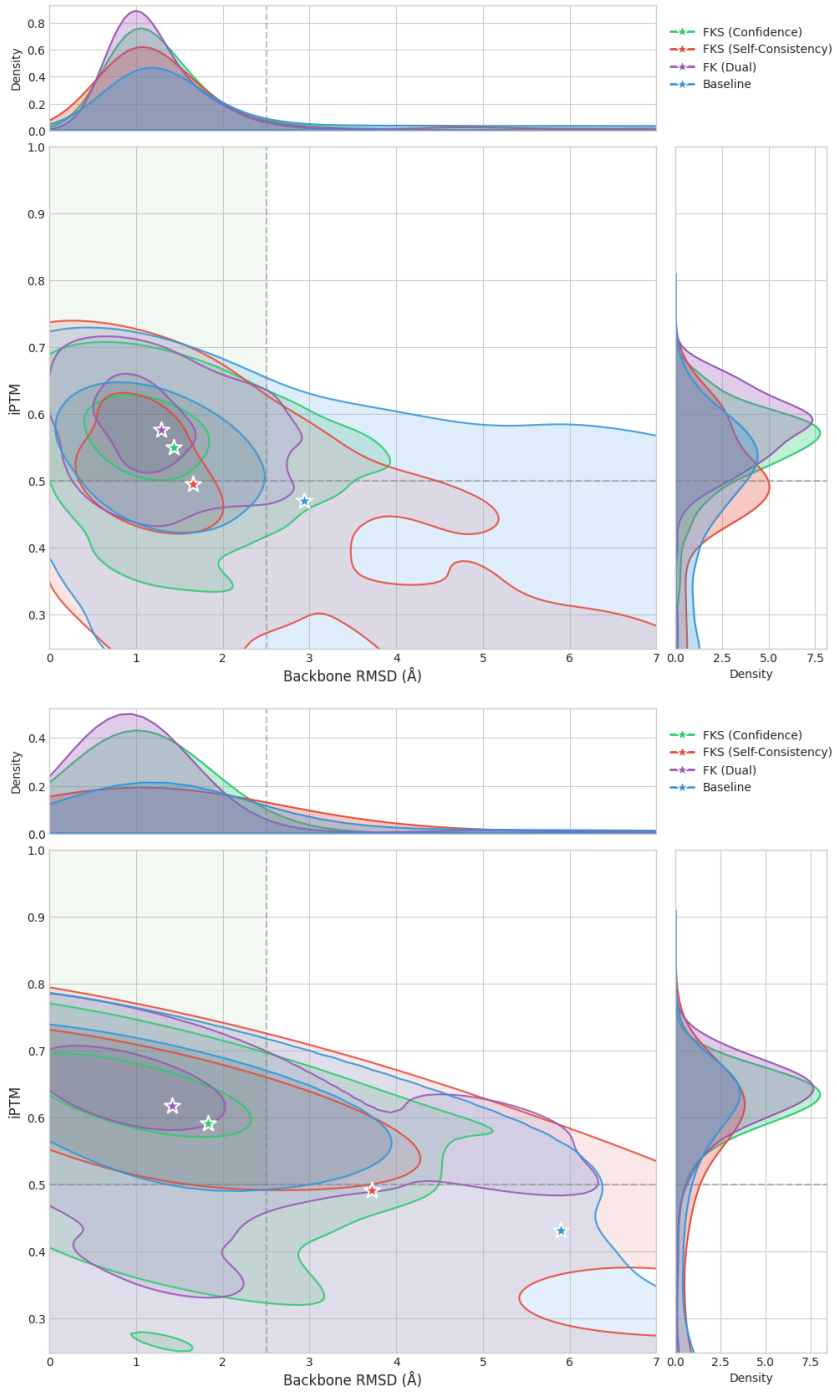

Figure 9: **FKS under BoltzGen's dilated schedule.** Joint distribution of post-refold iPTM vs. backbone RMSD on 1G13 (top) and 1JQD (bottom). Contours show kernel density estimates; stars mark distribution means. The dashed lines indicate the designability thresholds (iPTM>0.5, RMSD<2 Å). All three FKS variants—**confidence**, **self-consistency**, and **dual**—**concentrate mass in the high-iPTM / low-RMSD region** relative to the unguided **baseline**, with the dual reward achieving the tightest clustering near the designable zone.