Few-step Cofolding with All-Atom Flow Maps

Anonymous Authors[1]

Abstract

All-atom generative modeling of 3D biomolecular complexes has emerged as the dominant paradigm for predicting the structure of proteins and protein-ligand systems. Generating structures at the atomic level of fidelity, however, typically requires expensive iterative diffusion rollouts, making both conventional deployment and inference-time search techniques computationally costly. In this paper, we introduce the **DENOISER COFOLDING ALL-ATOM FLOWMAP** (DECAF) framework for distilling state-of-the-art all-atom cofolding models into all-atom flow maps that produce high-quality samples in only a few inference steps. We build DECAF on a denoiser-based formulation of flow maps with endpoint losses that naturally support $\mathrm{SE}(3)$ rigid alignment, which we show is critical for training accurate models. We further derive a simple change of variables that lets DECAF operate in the σ-space noise schedule of EDM-style architectures, enabling direct distillation from pretrained cofolding diffusion models. Equipped with DECAF's flowmap lookahead, we introduce a purpose-built inference-time framework that improves sampling through reward-guided search. Empirically, DECAF statistically improves over Boltz-1x in both accuracy (RMSD) and physical validity scores of protein-ligand poses at strict NFE budgets on the challenging Runs N' Poses, while also showing a more optimal Pareto frontier across all inference compute budgets on PoseBusters.

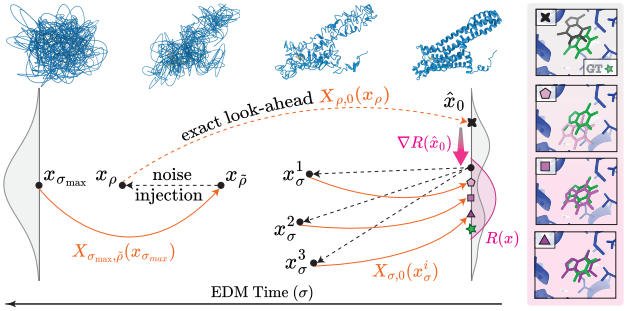

Figure 1. DECAF accelerates all-atom biomolecular structure prediction with a few-step flow-map lookahead across EDM noise levels. Atom-resolution guidance enables candidate search toward high-reward configurations.

1. Introduction

The accurate and efficient computational modeling of biological complexes has the potential to transform both our understanding of biomolecular mechanisms and our ability to catalyze the rational design of novel therapeutics (Chevalier et al., 2017; Ebrahimi & Samanta, 2023). At the core of this challenge is the need to model the intricate 3D intermolecular structures that govern processes such as protein folding, protein–ligand binding, and biocatalysis. This perspective is the thesis of structure-based drug design (SBDD), which seeks to engineer molecular structure to impart a desired downstream biological *function*. Despite its promise, progress in traditional SBDD is limited by the cost and latency of experimental structure determination. In contrast, AI-driven design offers a promising alternative by instead leveraging scalable computational approaches to unlock discoveries of novel therapies (Silva et al., 2019; Murray et al., 2022; Strauch et al., 2017; Gainza et al., 2020; Cao et al., 2020).

The modern *de facto* standard for AI-based SBDD is underpinned by large-scale generative models that represent biomolecular complexes of experimentally resolved structures directly at all-atom resolution, beginning with AlphaFold 3 (Abramson et al., 2024) and followed by other luminary works (Boitreaud et al., 2024; Wohlwend et al., 2025; Passaro et al., 2025; Team et al., 2026; Genesis Research, 2025). This perspective is well-suited to learning the global geometry of the target distribution, reflected in structural features such as relative pose, backbone arrangement, and secondary structure. However, unlike in classical generative modeling domains, success in biomolecular generation fundamentally requires modeling fine-grained local structure. Indeed, in high-impact application settings like protein-ligand cofolding, inaccurate local structure model-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Submitted to the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026). Do not distribute.

ing leads to catastrophic failure modes, yielding physically invalid generations that often include steric clashes, incorrect bond lengths and angles, strained side chain placements, and other stereochemical artifacts (Wohlwend et al., 2025).

Strict physical and biological constraints have shifted much of the compute burden to *inference time methods*. In particular, generating structures requires expensive and fine-grained numerical simulation of the learned dynamics, which are needed to accurately predict local structure. In addition, for downstream testing, it remains essential to generate a diverse pool of candidates that increases the transfer rate from in-silico design to wet lab success. Furthermore, refining samples via inference-time search with proxy physical reward models is a critical aspect of the evolving protein generative pipeline and plays an important role in facilitating utility in high-impact downstream applications. Despite this appeal, scaling inference is not a silver bullet and compounds the already expensive cost of numerical integration. For instance, reward functions in the biological setting can often be expensive to query and are only applicable to fully denoised 3D structures. Moreover, employing popular inference-time techniques that leverage multiple particles, such as Sequential Monte Carlo (SMC) (Del Moral et al., 2006), Feynman-Kac steering (FK) (Singhal et al., 2025), and Monte-Carlo Tree Search (MCTS) (Jain et al., 2025), doubly inflict the inference tax as they require multiple reward queries over each particle during the inference trajectory. This raises the natural motivating research question:

Q. *Can we train an all-atom cofolding model that can generate reward-optimized samples efficiently, using only a small number of neural function evaluations (NFEs) at inference time?*

Main contributions. In this paper, we answer the above question affirmatively and introduce the **DENOISER COFOLDING ALL-ATOM FLOWMAP** (DECAF) framework. DECAF is built on the flow map framework (Boffi et al., 2025b;a) (Figure 1) and efficiently distills a pretrained Boltz-1 model into the first all-atom cofolding model. Overall, we summarize our contributions as follows:

1. **Training methods**. We design the first all-atom protein flow-map with DECAF, and outline key methodological innovations that enable effective distillation of a pre-trained Boltz-1 teacher. In particular, we construct a novel reparameterization of the flow map in σ space that allows an easy conversion of standard flow map objectives for all-atom modeling. DECAF further exploits a denoiser-based flow map parametrization that crucially enables an SE(3) weighted rigid alignment of the ground truth structure—softly enforcing SE(3) symmetry—that is critical for stable training.
2. **Inference methods**. We introduce DECAF-SEARCH, an inference-time framework that leverages DECAF's flow map lookahead that enables higher fidelity reward estimation in comparison to all-atom diffusion models. DECAF-SEARCH shares the benefits of stochastic sampling (Kim et al., 2023), diffusion-SMC samplers (Singhal et al., 2025), Diffusion MCTS (Jain et al., 2025), Diamond maps (Holderrieth et al., 2026b; Potaptchik et al., 2026a), and FMRG (Huang et al., 2026) while being fundamentally an *inference-time search* method generating physically valid structures.
3. **Empirical performance**. We find that DECAF outperforms Boltz-1x *at low NFE budgets* on Runs N' Poses and matches the full-budget Boltz-1x (600 NFE) on PoseBusters with 20× less compute at inference.

2. Background and Preliminaries

2.1. All-Atom Biomolecular Diffusion Modeling

The dominant paradigm for all-atom biomolecular generative modeling is centered around diffusion-based cofolding models such as AlphaFold 3 (Abramson et al., 2024), which operates directly on the native Euclidean coordinates of structures. Through learning, AF3 like models can perform the task of *cofolding*—i.e., simultaneously predict the structure of a protein and a bound ligand. Standard implementations follow the EDM parameterization (Karras et al., 2022), which we review below.

VE Process. Given a target distribution of all-atom structures $p_{\text{data}}(x) \in \mathcal{P}(\mathbb{R}^d)$, the variance-exploding noising process corrupts a sample $x_0 \sim p_0(x_0) := p_{\text{data}}(x)$ with additive Gaussian noise,

$$x_\sigma = x_0 + \sigma\epsilon,\ \epsilon \sim \mathcal{N}(0, I),\ p_\sigma(x_\sigma) = p_0(x_0) \cdot \mathcal{N}(0, \sigma^2 I), \quad (1)$$

where $\sigma \in [\sigma_{\min}, \sigma_{\max}]$ indexes the noise level such that at times $t = 0$ and $t = 1$ there is $\sigma_{\min}$ and $\sigma_{\max}$ amounts of corruption added to x_0 respectively. Generation proceeds by solving for the time-reversal of the forward VE dynamics, which can be numerically simulated by following an ODE or SDE from high to low noise. For instance, we can simulate the probability-flow ODE: $dx_\sigma/d\sigma = -\sigma^2 \nabla_{x_\sigma} \log p_\sigma(x_\sigma)$. These reverse ODE dynamics require the Stein score $\nabla_{x_\sigma} \log p_\sigma(x_\sigma)$, which can be estimated through a diffusion model's denoiser D via Tweedie's formula (Tweedie, 1957):

$$s_\sigma(x_\sigma) := \nabla_{x_\sigma} \log p_\sigma(x_\sigma) \approx \frac{D_\sigma(x_\sigma) - x_\sigma}{\sigma^2}. \quad (2)$$

Eq. 2 can then be substituted in the flow ODE to simulate the reverse dynamics. The denoiser itself can be learned by a *simulation-free* ℓ_2-regression objective that performs denoising score matching across all noise levels ith a noise-dependent weighting function $\lambda(\sigma)$ against a clean sample x_0,

$$\mathcal{L}(\theta) = \mathbb{E}_{\sigma,\, x_0,\, \epsilon}\left[\lambda(\sigma)\, \|D_{\theta,\sigma}(x_0 + \sigma\epsilon) - x_0\|_2^2\right], \quad (3)$$

with $\epsilon \sim \mathcal{N}(0, I)$. **Structure prediction head**. For biomolecular all-atom diffusion models, the denoiser commonly leverages the EDM parametrization (Karras et al., 2022) that designs a σ-dependent preconditioning:

$$\hat{x}_0^{\text{EDM}} = c_{\text{skip}}(\sigma)\, x_\sigma + c_{\text{out}}(\sigma)\, F_{\theta,\sigma}(c_{\text{in}}(\sigma)\, x_\sigma, c_{\text{noise}}(\sigma))\,. \quad (4)$$

where F_θ is the raw network, c_{skip} controls the skip connection, c_{in} and c_{out} normalize input and output magnitudes, and c_{noise} embeds the noise level. In practice, the denoiser is implemented as a structure prediction head, typically a diffusion transformer with atom-attention encoder-decoder blocks.

The basic loss in Eq. 3 is augmented through SE(3) weighted rigid alignment of the ground truth structures to the prediction $\hat{x}_0^{\text{EDM}}$. This crucial step serves to simultaneously enforce soft global SE(3) symmetry and also reduce the variance of the diffusion loss. Finally, additional loss terms, including smooth-LDDT or bond-geometry penalties, encourage generating chemically plausible structures

2.2. Flow Maps

To accelerate inference in diffusion models, rather than simulating infinitesimal dynamics, one can learn a jump operator that directly traverses the probability-flow ODE associated with the diffusion model (Song et al., 2023; Song & Dhariwal, 2024; Boffi et al., 2025b;a). This operator is known as the *flow-map* and constitutes a map $X_{s,t} : [0,1]^2 \times \mathbb{R}^d \to \mathbb{R}^d$, which is the unique solution to the ODE that evolves the state dynamics between times s and t—i.e., the jump condition satisfies $X_{s,t}(x_s) = x_t$, for all $(s,t) \in [0,1]^2$. This leads to a natural parametrization of the flow-map as a displacement using the *average velocity* $u_{s,t}(x_s)$, between the two-time points s and t:

$$\begin{aligned} X_{s,t}(x_s) &= x_s + (t-s)\, u_{s,t}(x_s), \\ u_{s,t}(x_s) &= \tfrac{1}{t-s}\int_s^t v_\tau(x_\tau)\, d\tau, \end{aligned} \quad (5)$$

where $v_\tau(x_\tau)$, is the instantaneous velocity. It is clear from Eq. 5, that in the limit where the two time points converge the average velocity recovers the instantaneous velocity of the ODE, $\lim_{s\to t} \partial_t X_{s,t}(x_s) = u_{t,t}(x_t) := v_t(x_t)$. This is known as the tangent condition (Boffi et al., 2025b), and demonstrates that the flow-map contains an implicit instantaneous velocity in its parametrization.

At optimality, the flow map simultaneously enforces a set of Lagrange, Euler, and progressive consistency conditions; we refer the reader to Boffi et al. (2025a) for the construction of consistency models from these rules. Each condition naturally gives rise to a PINN-style loss that facilitates learning the flow-map (Boffi et al., 2025b;a; Frans et al., 2025; Kim et al., 2024). Furthermore, these losses can be employed to either self-distill or distill a pre-trained diffusion model into a flow-map by computing the RHS of each consistency condition with a frozen pre-trained model.

3. Method

We now introduce **DENOISER COFOLDING ALL-ATOM FLOWMAP** (DECAF), a novel σ-space flow-map framework that constructs an all-atom protein flow map by distilling a pre-trained all-atom teacher in the vein of AF3 (Abramson et al., 2024). In particular, DECAF outputs a few-step generative model that captures the 3D structure of protein-ligand interactions for the task of cofolding. Critically, DECAF offers two principal advantages over its pre-trained all-atom teacher:

1. *Few-step inference:* DECAF offers accelerated inference that compresses the full simulation of a diffusion trajectory into a few denoising steps without compromising sample quality.
2. *Flowmap lookahead:* DECAF by construction defines a lookahead map over end points that allows higher fidelity terminal reward estimation than a denoiser of an EDM model. As a result, any inference-time technique for reward optimization benefits not only fewer simulation steps that reduce simulation latency but also improved reward estimation and alignment.

3.1. DENOISER COFOLDING ALL-ATOM FLOWMAP

Two properties of the EDM parametrization complicate distillation of all-atom teachers. First, EDM's residual form Eq. 4 is conditioned on a single σ and does not extend cleanly to dual-time flow maps, motivating prior simplifications when distilling to flow maps (Nie et al., 2026). Second, the practical noise schedule $\sigma : [0,1] \to [\sigma_{\min}, \sigma_{\max}]$ is non-linear, so loss terms involving $\partial\sigma/\partial t$ become numerically unstable at the endpoints (Karras et al., 2022; Sabour et al., 2025b).

To circumvent the numerical instability of flow-map training using a pre-trained EDM protein model, we directly define a flow-map in the σ-noise space, allowing us to redefine all objectives and sampling directly over σ-steps—thus eliminating the problematic factor $\partial\sigma/\partial t$. This leads to our notion of σ-velocity, which represents the σ-instantaneous velocity along the PF-ODE,

$$\begin{aligned} v_\sigma(x_\sigma) &\triangleq \tfrac{dx_\sigma}{d\sigma}, \\ v_\sigma^{\text{EDM}}(x_\sigma) &= \tfrac{x_\sigma - D_\sigma(x_\sigma)}{\sigma} = \tfrac{x_\sigma - \hat{x}_0^{\text{EDM}}}{\sigma}. \end{aligned} \quad (6)$$

Our reparametrization of time extends, in a natural way, to now a two-noise level map that denoises $\rho \to \sigma$, for $\rho > \sigma$, using the analogous notion of *average* σ-velocity and flow-map parametrization $X_{\rho,\sigma}(x_\rho)$. This forms the

basis of the sampling update at inference:

$$u_{\rho,\sigma}(x_\rho) = \frac{1}{\rho-\sigma}\int_\rho^\sigma v_{\bar\sigma}(x_{\bar\sigma})\,d\bar\sigma, \qquad X_{\rho,\sigma}(x_\rho) = x_\rho - (\rho-\sigma)\,u_{\rho,\sigma}(x_\rho). \tag{7}$$

Denoiser Parametrization. To train our all-atom protein flow map, we follow the mean-flow objective (Geng et al., 2025a), which is also equivalent to the Eulerian objective (Boffi et al., 2025a). This requires the construction of the instantaneous σ-velocity *implied* by the average σ-velocity of Eq. 7. Explicitly, we efficiently compute this quantity using Jacobian-vector products `jvp`,

$$v_\rho(x_\rho) = u_{\rho,\sigma}(x_\rho) + (\rho-\sigma)\frac{\mathrm{d}}{\mathrm{d}\rho}u_{\rho,\sigma}(x_\rho). \tag{8}$$

In Eq. 8, $\mathrm{d}/\mathrm{d}\rho$ denotes the total derivative of the average σ-velocity. In practice, we implement this term as $\mathtt{sg}(\mathtt{jvp}(u_{\rho,\sigma}(x_\rho), (x_\rho,\rho,\sigma), (v,1,0)))$—a stop-gradient applied to the Jacobian–vector product of $u_{\rho,\sigma}$ along the tangent direction $(v,1,0)$—so that gradients flow only through the explicit $u_{\rho,\sigma}$ factor. When distilling a pre-trained EDM teacher, substituting the ground-truth σ-velocity with v_σ^{EDM} yields the standard mean-flow objective that matches Eq. 8 to Eq. 6 (Geng et al., 2025a;b; Lu et al., 2026; Potaptchik et al., 2026b). For all-atom protein models, however, we instead exploit a standard practice from AF3 that yields a lower-variance loss: predict the endpoint $\hat x_{\text{tgt}}$ and align it to $\hat x_0^{\text{EDM}}$ under $\mathrm{SE}(3)$ rigid (Kabsch) alignment. We therefore parametrize the flow map as a two-noise-level denoiser $D(x_\rho,\rho,\sigma)$ and recover the $\hat x_{\text{tgt}}$ prediction from the σ-instantaneous velocity of Eq. 8:

$$\hat x_{\text{tgt}}^{\text{DECAF}} := D(x_\rho,\rho,\sigma) = x_\rho - \rho\cdot V(x_\rho,\rho,\sigma) \tag{9}$$

The above equation gives rise to a *two-time* denoiser (Lu et al., 2026; Lee et al., 2026; Potaptchik et al., 2026b; Roos et al., 2026) that can be leveraged to construct an endpoint loss:

$$\mathcal{L} = \mathbb{E}_{x_0,x_\rho,x_\sigma}\left[\frac{1}{\sigma^2}\min_{g\in\mathrm{SE}(3)}\left\|\hat x_{\text{tgt}}^{\text{DECAF}} - \mathtt{sg}\big(\zeta(g)\circ\hat x_0^{\text{EDM}}\big)\right\|^2\right]. \tag{10}$$

Here we take argmin over the entire group $\mathrm{SE}(3)$: $\zeta(g)$ is its matrix representation and is the rigid alignment step performed using the Kabsch algorithm, while $\hat x_0^{\text{EDM}}$ is the prediction computed by the pre-trained EDM teacher. We emphasize that this exact formulation fails under a velocity loss, as subtracting translation would lose a degree of freedom, and thus complicates the distillation setup.

The loss in Eq. 10 naturally supports both off-diagonal training and diagonal training through sampling of noise levels $(\rho,\sigma)\sim\text{Sampler}(\sigma_{\min},\sigma_{\max})$. In particular, when $\rho=\sigma$, we learn to match exactly the score associated with the PF-ODE of the EDM teacher, while in all other off-diagonal cases, we learn to take $(\rho-\sigma)$ jumps along the trajectory of the PF-ODE as in Eq. 7.

Algorithm 1 DECAF γ-sampling

Require: FlowMap X; N steps; $\gamma\in[0,1]$.
$x_{\sigma_N}\sim\mathcal{N}(0,\sigma_N^2 I)$
for $n=N$ **down to** 1 **do**
  $\tilde\sigma_{n-1}\leftarrow\sqrt{1-\gamma^2}\,\sigma_{n-1}$
  $x_{\tilde\sigma_{n-1}}\leftarrow X(x_{\sigma_n},\sigma_n,\tilde\sigma_{n-1})$ ▷ *flowmap*
  $\epsilon\sim\mathcal{N}(0,I)$
  $x_{\sigma_{n-1}}\leftarrow x_{\tilde\sigma_{n-1}}+\gamma\,\sigma_{n-1}\,\epsilon$ ▷ *re-noise*
return x_{σ_0}

Flow map sampling algorithm. To sample from DECAF we design a stochastic γ-sampler, which shares inspiration from the consistency models literature (Kim et al., 2023). We present this γ-sampler in Algorithm 1, which allows us to toggle deterministic sampling ($\gamma=0$) to more stochastic sampling for $\gamma>0$. As we later demonstrate in our experiments §4, the added stochasticity aids overall performance at no added cost and also enjoys seamless integration with our inference strategy in §3.2.

3.2. Inference-Time Search

The challenge of generating physically plausible biomolecular structures has motivated a vast literature on computationally intensive inference-time correction of pre-trained all-atom diffusion models. While DECAF is primarily designed as a few-step all-atom model that reduces inference latency for conventional deployment, it also enables a second computational advantage. Specifically, DECAF also enables substantial gains in *inference-time search* (i.e., reward alignment (Uehara et al., 2025)) for the cofolding problems we consider.

In biomolecular modeling, a common form of inference-time search defines a terminal reward $R:\mathbb{R}^d\to\mathbb{R}$ that measures physical validity, for example, penalizing steric clashes or violations of bond lengths and angles (Wohlwend et al., 2025, Section 4). The goal is to steer generation toward samples with high $R(x_0)$ while preserving structural accuracy. A central difficulty is that R is only defined on clean structures, whereas steering decisions must be made from noisy intermediate states x_σ. Consequently, we require an efficient estimate of the reward expected after denoising x_σ to a clean structure. DECAF provides such an estimate through its learned two-time flow map: given an intermediate state x_σ, we compute a look-ahead (end point) prediction over clean samples

$$\hat x_0 = X(x_\sigma,\sigma,0) = x_\sigma - \sigma\cdot u_{\sigma,0}(x_\sigma), \tag{11}$$

and use $R(\hat x_0)$ as a proxy for the expected reward of x_σ. Similar flow map look-aheads have been used for steering

in the image domain (Sabour et al., 2025a; Holderrieth et al., 2026a; Potaptchik et al., 2026a; Huang et al., 2026). We explore their analogue in all-atom biomolecular generation, which differs in two important ways. First, our base sampler is not a standard ODE or SDE integrator but γ-sampler. Second, rewards are based on physical violations whose supervisory signal is highly non-smooth and uninformative at noisy states x_σ. These considerations motivate refining generations through clean-space look-aheads rather than through gradients in noisy state space.

DECAF-SEARCH. We introduce an inference-time search algorithm DECAF-SEARCH for all-atom flow maps (see Algorithm 2). DECAF-SEARCH adapts standard inference-time steering methods, including Feynman–Kac (FK) (Singhal et al., 2025; Skreta et al., 2025) and diffusion-based MCTS (Jain et al., 2025), to flow maps and the γ-sampling setting. Starting from a population of particles, we repeatedly denoise each particle from x_σ to a clean look-ahead $\hat{x}_0$ (possibly over several flow map steps), evaluate its reward, and optionally improve it by clean-space gradient ascent

$$\hat{x}_0 \leftarrow \hat{x}_0 + \beta \nabla_{\hat{x}_0} R(\hat{x}_0). \tag{12}$$

Then, we renoise the improved structure to a noisy state $x_{\tilde{\sigma}}$ before continuing to the next γ-sampling iteration. Inspired by Holderrieth et al. (2026a, Proposition 5.1.), we also consider a variant where the gradient is an average of several Monte Carlo samples (MC-GRAD in §A). Across particles, compute is preferentially allocated to promising branches using either SMC resampling or an upper-confidence-bound criterion. With finite-temperature resampling, DECAF-SEARCH resembles FK steering with UCB/UCT selection and recovers a simple MCTS-style variant (Jain et al., 2025). DECAF-SEARCH also extends the FK-style steering used in Boltz-1x (Wohlwend et al., 2025): intermediate clean predictions are obtained using the learned DECAF flow map, which can provide more accurate or more efficient look-aheads than a single-step denoiser.

4. Experiments

We investigate the application of DECAF for the task of cofolding protein-ligand interactions. In particular, we distill a pre-trained Boltz-1 model using Eq. 10 and sample using DECAF-SEARCH (Algorithm 2), in contrast to Boltz-1x the physical potential variant (Wohlwend et al., 2025). For fair comparison, DECAF shares architecture, training data, and pretraining with Boltz-1 (see §B for architecture, experimental setup, and hyperparameters). We include additional ablations in §C.

Benchmarks. The primary benchmark we use for analysis is Runs N' Poses (RnP) (Škrinjar et al., 2025). We limit the experiments to a stricter set of 702 structures with a cutoff date of `2023-06-01`. We also conduct additional analysis on PoseBusters (Buttenschoen et al., 2024) with cutoff date of `2021-10-01`, yielding 282 structures that can be handled by Boltz-1 on a single A100-80GB GPU.

Metrics. Following the standard practice, we report the following metrics: (i) *RMSD < 2 Å* defined as the percentage of the test structures for which root mean square distance between ground-truth and generated ligand poses is under 2 Å; (ii) *PB-Valid* defined as the percentage of test structures that are physically valid according to the PoseBusters library; (iii) *lDDT-PLI* defined as local distance difference test (Mariani et al., 2013) on the short-range protein–ligand contacts within a 6 Å protein-ligand pocket, where side-chain atoms typically outnumber backbone atoms. We also define *Success Rate (%)* as a percentage of the test structures that satisfy RMSD < 2 Å and PB-Valid criteria.

Our experiments seek to answer three key questions that test the empirical caliber of DECAF:

(Q1.) Low NFE regime. Does DECAF outperform Boltz-1 with a limited inference budget (§4.1)?

(Q2.) Analysis of compute-optimal frontier. What is the Pareto frontier of DECAF-SEARCH against Boltz-1x for inference time scaling across *any* inference compute budget (§4.2)?

(Q3.) Performance analysis. What are the relative quantitative and qualitative differences in generation quality at the sample-level of DECAF in comparison to its teacher Boltz-1 (§4.3)?

4.1. Performance at low NFEs (Q1.)

We evaluate DECAF and Boltz-1 on Runs N' Poses for various NFE inference regimes. We report our main results in Table 1 (average over 5 poses) on Runs N' Poses structures drawn entirely from PDB depositions released *after* the `2023-06` cutoff date for training. Specifically, we sample DECAF-SEARCH with SMC resampling over 4 particles—bearing similarity to FK steering (Singhal et al., 2025)—at compute budgets of 10, 20, 25, 40, and 50 NFEs. We also include a high-budget (800 NFEs) comparison between Boltz-1x and DECAF-SEARCH with a variation that is closest to MCTS. As Table 1 demonstrates, Boltz-1x fails catastrophically at generating plausible structures at low steps regime under the default sampling configuration. As an orthogonal contribution, we remedy this by tuning the step scale in the default Boltz-1x to $\eta = 1$ when ≤ 15 diffusion steps, which restores stable SDE sampling in the few-step regime. We refer to this setting as Boltz-1x-tuned.

At *every* low compute budget in Table 1, DECAF outperforms Boltz-1x-tuned on *every* metric. The improvement is significant in the 20–160 NFE range (paired Wilcoxon signed-rank, $p < 0.001$ on nearly all metrics). At 50 NFEs,

Table 1. **DECAF vs Boltz-1x on Runs N' Poses benchmark.** The best recipe at or below the NFE budget is reported per method averaged over 5 poses. **Bold** marks the winner between Boltz-1x and DECAF; stars indicate paired Wilcoxon signed-rank significance vs the other model (two-sided): $^{*}p<0.05$, $^{**}p<0.01$, $^{***}p<0.001$. † Indicates numbers taken from (Genesis Research, 2025)

Method	PB Valid↑	Success Rate↑	lDDT-PLI↑
≤ 10 NFEs (2 steps)			
Boltz-1x (default)	0.0	0.0	0.0
Boltz-1x (tuned)	21.8	17.2	45.8
DECAF-SEARCH	***24.0**	****19.6**	**46.4**
≤ 20 NFEs (5 steps)			
Boltz-1x (tuned)	75.9	50.3	59.8
DECAF-SEARCH	*****83.5**	*****57.4**	*****66.5**
≤ 25 NFEs (6 steps)			
Boltz-1x (tuned)	73.4	51.9	61.8
DECAF-SEARCH	*****80.1**	*****57.0**	*****67.8**
≤ 40 NFEs (10 steps)			
Boltz-1x (tuned)	86.6	55.4	64.2
DECAF-SEARCH	*****91.6**	*****61.8**	*****67.2**
≤ 50 NFEs (12 steps)			
Boltz-1x (tuned)	86.3	56.5	63.7
DECAF-SEARCH	*****92.7**	*****62.6**	****67.1**
≤ 160 NFEs			
Boltz-1x (default, 40 steps)	88.5	58.7	67.6
DECAF-SEARCH (MC-GRAD, 15 steps)	*****92.1**	*****63.7**	*****69.2**
800 NFEs (Full budget)			
Boltz-1x (ref., 200 steps)	**95.9**	64.9	**69.9**
DECAF-SEARCH (MCTS, 10 steps)	95.5	**65.0**	69.5
Frontier cofolding models†			
AlphaFold 3	73.9	60.7	80.9
Pearl	96.1	72.1	81.9

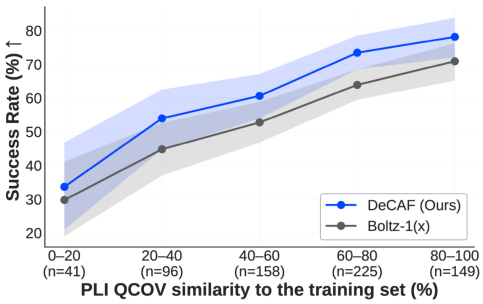

Figure 2. Success rate vs. training-set similarity on the RnP benchmark at 40 NFE.

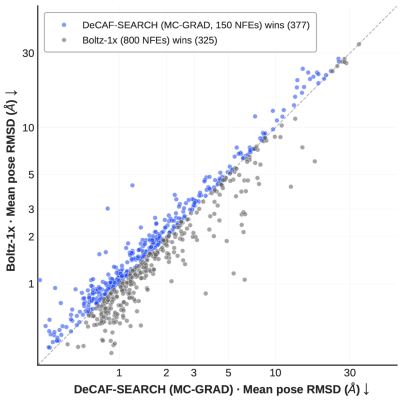

Figure 3. Mean RMSD per structure for DECAF-SEARCH (150 NFEs) (x) vs. Boltz-1x (800 NFEs) (y) on RnP.

DECAF is also competitive with frontier cofolding models that use $\geq$ 800 NFEs—surpassing even AlphaFold 3 on the Success Rate and PB-Valid metrics. At a matched budget of 800 NFEs, DECAF is on par with Boltz-1x on all three metrics.

Generalization. An important criterion for cofolding models is their ability to handle difficult targets, as it is a proxy for model generalization beyond the training set. In Figure 2, we plot the success rate against the pocket-similarity metric *PLI Q-Coverage* (shaded regions show 95% bootstrap confidence intervals (1000 resamples per bin). We observe that DECAF's margin over Boltz-1x holds across every quartile, including the most out-of-distribution bin (lowest PLI Q-Coverage), confirming that the performance gain is not driven by structures that are close to the training set.

Lastly, in Figure 3 we perform a fine-grained analysis of performance at the sample-level. In particular, we conduct a comparison between DECAF-SEARCH (MC-GRAD) at 150 NFEs and Boltz-1x (800 NFEs) for each of the 702 structures in Runs N' Poses. Each point gives one target's mean pose RMSD under DECAF (x) and Boltz-1x (y); points above the $y = x$ diagonal indicate DECAF wins. DECAF matches Boltz-1x on the per-target distribution of mean RMSD at 5.3× less inference compute. The advantage holds in the difficult tail: among the targets where at least one method exceeds 10,Å, DECAF attains the lower RMSD on 71% of them.

4.2. Analysis of the compute-optimal frontier (Q2.)

We next investigate the efficacy of DECAF in comparison to Boltz-1x as a function of increasing inference budget on the PoseBusters benchmark. Through this study, we characterize the peak attainable performance of Boltz-1x and DECAF. As such, we elucidate the precise inference recipe for DECAF-SEARCH that is optimal at each NFE budget. We use PoseBusters because its modest size (∼300 structures) makes the dense recipe-and-NFE sweep tractable. Moreover, its 2021-10-01 cutoff enables the principled investigation of the in-distribution Pareto frontier.

Figure 4 studies the frontier across PB Valid, RMSD<2 Å and Success Rate with a best @5 selection criterion for each NFE threshold. We find significant inference cost reductions for DECAF-SEARCH over Boltz-1x full-scale configuration with 3 particles (600 NFEs) with as few as 30 NFEs—*a 20× inference cost reduction*. Importantly, owing to DECAF's flowmap lookahead, this reduction also leads to better quantitative performance with DECAF-SEARCH's Pareto frontier dominating Boltz-1x across every NFE budget on all metrics. We further find that at different inference budgets, the exact Pareto-optimal recipe for DECAF-

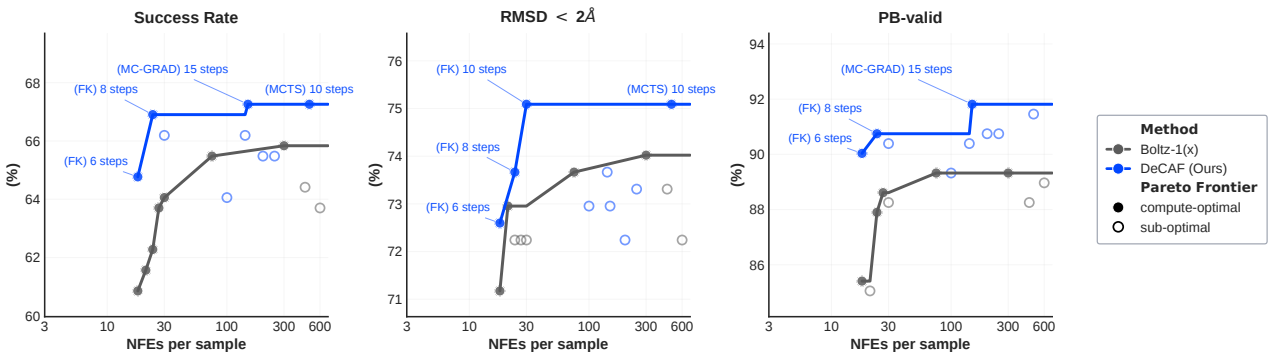

Figure 4. **DECAF is cost-effective as we increase compute.** We compare DECAF-SEARCH and Boltz-1x as we increase NFEs budget. The solid lines are the per-NFE compute optimal frontier for the two methods.

Table 2. **Parameterization ablation.** The x_0-aligned parameterization implements Equation (9); the velocity parameterization predicts $v_\rho(x_\rho)$; consistency distillation (CD) follows (Song et al., 2023). Numbers on PoseBusters, 30 NFEs.

Parametrization	PB-Valid ↑	RMSD<2Å ↑
x_0-aligned	**90.4**	**75.6**
velocity	9.2	49.1
Cons. Distillation	0.0	49.5

SEARCH varies. Specifically, at low NFEs (≤ 30), particle-based SMC akin to FK-steering is optimal. At the moderate NFEs ($100 - 250$), we find our Monte Carlo estimation of the reward gradient (MC-GRAD) to be the most effective. Finally, at large NFE budgets, we find DECAF-SEARCH with MCTS to be the most impactful at NFEs ≥ 142.

Fine grained analysis. As we increase the NFE based on the hyperparams in DECAF-SEARCH, however, we note that the head-to-head with Boltz-1x favors different complexes, i.e., the successful complexes at high NFE are not a superset of the successful complexes at low NFE (Table 7). This suggests that a more complex interplay between sampling accuracy and reward guidance within DECAF. We highlight representative samples from DECAF-SEARCH (FK) at low NFE, which have higher accuracy than any pose sampled from Boltz-1x at any NFE (Figure 5, additional samples in Figure 9).

4.3. Performance Analysis (Q3.)

Posebusters validity. We qualitatively study the pose quality at various NFE budgets. In Figure 5 we visually depict the samples and observe that, in comparison to Boltz-1x, the MCTS version of DECAF-SEARCH (600 NFEs) can improve pose accuracy. Meanwhile, MC-GRAD (150 NFEs) reduces steric clashes in comparison to Boltz-1x, highlighting improved physical reward alignment.

DECAF-SEARCH methods benefit from having both generally high pose quality (i.e., reward optimization) and more accurate ligand placement. While few-NFE inference with DECAF already yields better failure rates in relation to Boltz-1x (c.f. Table 8) we see further refinement in pose quality when scaling NFEs further with the MC-GRAD and MCTS variants. We additionally note that pose quality on PoseBusters is greatly determined by reward design, and leads to common failure modes across DECAF and Boltz-1x that share the same potentials. For instance, we observe that sp2-hybridized bonds are often not planar. In addition, some PoseBusters quality checks have stricter tolerances than are chemically accurate, such as flagging metal ion coordination as a "clash" based solely on the neutral atoms' van der Waals radii (Figure 7). We find majority of these failures can be mitigated through inference search.

Generalization on chemically-relevant targets. We stratify the Runs N' Poses benchmark to evaluate our models on challenging scenarios most relevant to drug discovery tasks. We consider several slices which probe these settings: drug-like ligands only, ligands interacting with ions and cofactors, and ligands at protein-protein interfaces. In these subsets, DECAF-SEARCH outperforms Boltz-1x to a statistically significant degree (Table 3). We curate representative poses in Figure 8.

Parametrization. We ablate DECAF's denoiser-aligned parametrization (Equation (9)) against a velocity-predictor student (Geng et al., 2025b) that directly regresses $v_\rho(x_\rho)$ under the same JVP schedule and data budget. We also experimented with consistency distillation parametrization (Song et al., 2023). Table 2 confirms that only our chosen x_0-aligned parametrization can sample valid poses. The denoiser parametrization reuses EDM preconditioning and enables Kabsch alignment on $\hat{x}_0$—which we find critical for stable training.

Table 3. **Method performance on challenging and chemically-relevant targets.** Best@5 joint RMSD < 2 Å and PB-valid (higher is better) on slices of Runs N' Poses. Drug-likeness uses a QED $\geq$ 0.65 threshold (Bickerton et al., 2012). Asterisks (**) mark statistically-significant improvements over the best Boltz-1x variant (paired Wilcoxon signed-rank, $p < 0.01$).

Slice	N	DECAF-SEARCH		Boltz-1x	
		NFE = 40	NFE = 800	NFE = 40	NFE = 800
Drug-like ligands	283	**$\mathbf{0.809}$	0.806	0.787	0.783
Ion/cofactor coordination	134	0.754	**$\mathbf{0.761}$	0.749	0.760
Ligand at PPI	292	0.777	**$\mathbf{0.784}$	0.779	0.780

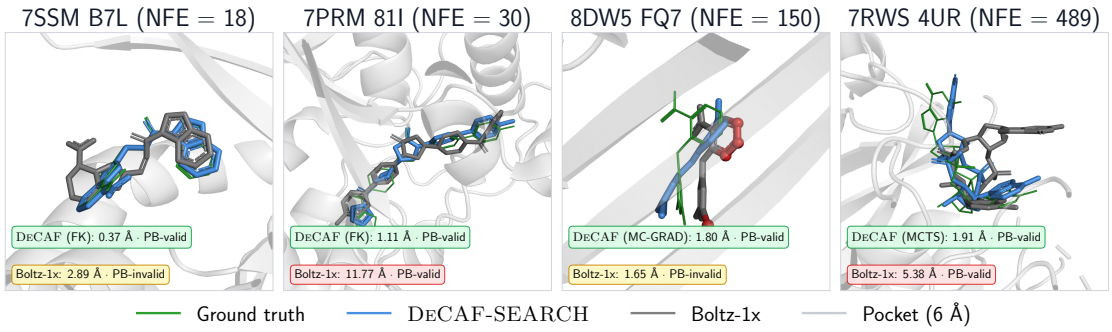

Figure 5. Each panel overlays the **ground-truth** crystal ligand against **DECAF-SEARCH** and **Boltz-1x** samples at the specified NFE. The protein pocket is shown as a light-gray cartoon, and predicted ligand atoms that clash with the protein in red. At the bottom, we report the RMSD and posebusters validity of the pose.

5. Related Work

Protein generation. Early generative models focused on backbone design (Yim et al., 2023; Watson et al., 2023; Bose et al., 2023; Huguet et al., 2024; Geffner et al., 2025b;a). Since AlphaFold 3 (Abramson et al., 2024) established EDM-style denoising (Karras et al., 2022) over all-atom coordinates as the dominant paradigm for cofolding, several open and closed-source systems have followed, including Boltz-1(x) (Wohlwend et al., 2025), ProteniX (Team et al., 2026), Chai-1 (Boitreaud et al., 2024), Pearl (Genesis Research, 2025), Complexa (Didi et al., 2026), and DISCO (Rector-Brooks et al., 2026). These all share a costly inference-time bottleneck of $\mathcal{O}(200)$ NFEs per sample.

Few-step generative models. Several methods compress full generative trajectories, including consistency models (Song et al., 2023; Song & Dhariwal, 2024), Consistency Trajectory Models (Kim et al., 2024), shortcut models (Frans et al., 2025), and the MeanFlow (Geng et al., 2025a; Boffi et al., 2025a). Despite this flourishing literature, the distillation of all-atom cofolding models remains underexplored, with only DCFold as closed-source concurrent work (Zhang et al., 2026).

Inference-time steering. Inference-time techniques for diffusion include classifier guidance (Ho & Salimans, 2021), Universal Guidance (Bansal et al., 2024), and particle-based methods derived from SMC (Del Moral et al., 2006), including Feynman–Kac methods (Singhal et al., 2025; Skreta et al., 2025) and diffusion MCTS (Jain et al., 2025).

In the protein setting, AF3-family models routinely employ physics-informed potentials such as the Boltz-1x stereochemical potentials (Wohlwend et al., 2025) whose cost grows linearly in both denoising steps and particle count.

6. Conclusion

We introduced DECAF, a flow map that distills a pretrained all-atom cofolding diffusion model into a few-step generator. The construction rests on two technical choices: a reparameterization in σ-space that matches the EDM-style adopted by standard cofolding models, and a denoiser parametrization that preserves the $\mathrm{SE}(3)$ rigid alignment. On top of DECAF, we built DECAF-SEARCH, an inference-time search framework that exploits flow map lookahead for higher-fidelity reward alignment. Empirically, DECAF-SEARCH matches the 600-NFE Boltz-1 teacher with a 20$\times$ reduction in function evaluations on PoseBusters and improves over Boltz-1x on Runs N' Poses at every low-NFE setting we considered. Several limitations point to next steps. Our reward signal is inherited from Boltz-1x, so failure modes such as non-planar sp^2 bonds reflect this choice rather than the search procedure. The σ-space denoiser formulation is not specific to cofolding and may extend to nucleic acids, larger assemblies, and multi-chain systems. Finally, the non-monotone relationship between NFE budget and per-target success suggests that adaptive search and joint training of the flow map with task-specific rewards are natural avenues for further gains.

References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630 (8016):493–500, 2024.

Bansal, A., Chu, H.-M., Schwarzschild, A., Sengupta, S., Goldblum, M., Geiping, J., and Goldstein, T. Universal guidance for diffusion models. In *ICLR*, 2024.

Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S., and Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nature Chemistry*, 4(2):90–98, January 2012. ISSN 1755-4349. doi: 10.1038/nchem.1243. URL http://dx.doi.org/10.1038/nchem.1243.

Boffi, N. M., Albergo, M. S., and Vanden-Eijnden, E. Flow map matching with stochastic interpolants: A mathematical framework for consistency models. *Transactions on Machine Learning Research*, 2025a. ISSN 2835-8856.

Boffi, N. M., Albergo, M. S., and Vanden-Eijnden, E. How to build a consistency model: Learning flow maps via self-distillation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025b.

Boitreaud, J., Dent, J., McPartlon, M., Meier, J., Reis, V., Rogozhnikov, A., and Wu, K. Chai-1: Decoding the molecular interactions of life. *bioRxiv*, 2024. doi: 10.1101/2024.10.10.615955. URL https://www.biorxiv.org/content/10.1101/2024.10.10.615955v1.

Bose, A. J., Akhound-Sadegh, T., Huguet, G., Fatras, K., Rector-Brooks, J., Liu, C.-H., Nica, A. C., Korablyov, M., Bronstein, M., and Tong, A. Se (3)-stochastic flow matching for protein backbone generation. *arXiv preprint arXiv:2310.02391*, 2023.

Buttenschoen, M., Morris, G. M., and Deane, C. M. PoseBusters: Ai-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science*, 15(9):3130–3139, 2024. doi: 10.1039/D3SC04185A.

Cao, L., Goreshnik, I., Coventry, B., Case, J. B., Miller, L., Kozodoy, L., Chen, R. E., Carter, L., Walls, A. C., Park, Y.-J., Strauch, E.-M., Stewart, L., Diamond, M. S., Veesler, D., and Baker, D. De novo design of picomolar sars-cov-2 miniprotein inhibitors. *Science*, 370(6515): 426–431, 2020.

Chevalier, A., Silva, D.-A., Rocklin, G. J., Hicks, D. R., Vergara, R., Murapa, P., Bernard, S. M., Zhang, L., Lam, K.-H., Yao, G., et al. Massively parallel de novo protein design for targeted therapeutics. *Nature*, 550(7674):74–79, 2017.

Del Moral, P., Doucet, A., and Jasra, A. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B*, 68(3):411–436, 2006.

Didi, K., Zhang, Z., Zhou, G., Reidenbach, D., Cao, Z., Cha, S., Geffner, T., Dallago, C., Tang, J., Bronstein, M. M., et al. Scaling atomistic protein binder design with generative pretraining and test-time compute. *arXiv preprint arXiv:2603.27950*, 2026.

Ebrahimi, S. B. and Samanta, D. Engineering protein-based therapeutics through structural and chemical design. *Nature Communications*, 14(1):2411, 2023.

Frans, K., Hafner, D., Levine, S., and Abbeel, P. One step diffusion via shortcut models. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Gainza, P., Sverrisson, F., Monti, F., Rodola, E., Boscaini, D., Bronstein, M. M., and Correia, B. E. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods*, 17(2): 184–192, 2020.

Geffner, T., Didi, K., Cao, Z., Reidenbach, D., Zhang, Z., Dallago, C., Kucukbenli, E., Kreis, K., and Vahdat, A. La-proteina: Atomistic protein generation via partially latent flow matching. *arXiv preprint arXiv:2507.09466*, 2025a.

Geffner, T., Didi, K., Zhang, Z., Reidenbach, D., Cao, Z., Yim, J., Geiger, M., Dallago, C., Kucukbenli, E., Vahdat, A., et al. Proteina: Scaling flow-based protein structure generative models. *arXiv preprint arXiv:2503.00710*, 2025b.

Genesis Research, T. Pearl: A foundation model for placing every atom in the right location. *arXiv preprint arXiv:2510.24670*, 2025.

Geng, Z., Deng, M., Bai, X., Kolter, J. Z., and He, K. Mean flows for one-step generative modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025a.

Geng, Z., Lu, Y., Wu, Z., Shechtman, E., Kolter, J. Z., and He, K. Improved mean flows: On the challenges of fastforward generative models. *arXiv preprint arXiv:2512.02012*, 2025b.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

Holderrieth, P., Chen, D., Eyring, L., Shah, I., Anantharaman, G., He, Y., Akata, Z., Jaakkola, T., Boffi, N. M., and Simchowitz, M. Diamond maps: Efficient reward

alignment via stochastic flow maps. *arXiv preprint arXiv:2602.05993*, 2026a.

Holderrieth, P., Singer, U., Jaakkola, T., Chen, R. T. Q., Lipman, Y., and Karrer, B. GLASS flows: Efficient inference for reward alignment of flow and diffusion models. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026b.

Huang, J. Y., Lin, J., Shah, S., Nair, K., and Boffi, N. M. How to guide your flow: Few-step alignment via flow map reward guidance. *arXiv preprint arXiv:2604.27147*, 2026.

Huguet, G., Vuckovic, J., Fatras, K., Thibodeau-Laufer, E., Lemos, P., Islam, R., Liu, C.-H., Rector-Brooks, J., Akhound-Sadegh, T., Bronstein, M., et al. Sequence-augmented se (3)-flow matching for conditional protein generation. *Advances in neural information processing systems*, 37:33007–33036, 2024.

Jain, V., Sareen, K., Pedramfar, M., and Ravanbakhsh, S. Diffusion tree sampling: Scalable inference-time alignment of diffusion models. *arXiv preprint arXiv:2506.20701*, 2025.

Karras, T., Aittala, M., Aila, T., and Laine, S. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, 2022.

Kim, D., Lai, C.-H., Liao, W.-H., Murata, N., Takida, Y., Uesaka, T., He, Y., Mitsufuji, Y., and Ermon, S. Consistency trajectory models: Learning probability flow ode trajectory of diffusion. *arXiv preprint arXiv:2310.02279*, 2023.

Kim, D., Lai, C.-H., Liao, W.-H., Murata, N., Takida, Y., Uesaka, T., He, Y., Mitsufuji, Y., and Ermon, S. Consistency trajectory models: Learning probability flow ODE trajectory of diffusion. In *International Conference on Learning Representations (ICLR)*, 2024.

Lee, C., Yoo, J., Agarwal, M., Shah, S., Huang, J., Raghunathan, A., Hong, S., Boffi, N. M., and Kim, J. Flow map language models: One-step language modeling via continuous denoising. *arXiv preprint arXiv:2602.16813*, 2026.

Lu, Y., Lu, S., Sun, Q., Zhao, H., Jiang, Z., Wang, X., Li, T., Geng, Z., and He, K. One-step latent-free image generation with pixel mean flows. *arXiv preprint arXiv:2601.22158*, 2026.

Mariani, V., Biasini, M., Barbato, A., and Schwede, T. lddt: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29(21):2722–2728, 2013.

Murray, K. A., Hu, C. J., Griner, S. L., Pan, H., Bowler, J. T., Abskharon, R., Rosenberg, G. M., Cheng, X., Seidler, P. M., and Eisenberg, D. S. De novo designed protein inhibitors of amyloid aggregation and seeding. *Proceedings of the National Academy of Sciences*, 119 (34):e2206240119, 2022.

Nie, W., Berner, J., Liu, C., and Vahdat, A. Nvidia fastgen: Fast generation from diffusion models, 2026. URL https://github.com/NVlabs/FastGen.

Passaro, S., Corso, G., Wohlwend, J., Reveiz, M., Thaler, S., Somnath, V. R., Getz, N., Portnoi, T., Roy, J., Stark, H., et al. Boltz-2: Towards accurate and efficient binding affinity prediction. *BioRxiv*, 2025.

Potaptchik, P., Saravanan, A., Mammadov, A., Prat, A., Albergo, M. S., and Teh, Y. W. Meta flow maps enable scalable reward alignment. *arXiv preprint arXiv:2601.14430*, 2026a.

Potaptchik, P., Yim, J., Saravanan, A., Holderrieth, P., Vanden-Eijnden, E., and Albergo, M. S. Discrete flow maps. *arXiv preprint arXiv:2604.09784*, 2026b.

Rector-Brooks, J., Lambert, T., Skreta, M., Roth, D., Long, Y., Li, Z.-Q., Zhang, X., Cretu, M., Li, F.-Z., Ganapathy, T., et al. General multimodal protein design enables dna-encoding of chemistry. *arXiv preprint arXiv:2604.05181*, 2026.

Roos, D., Davis, O., Eijkelboom, F., Bronstein, M., Welling, M., Ceylan, İ. İ., Ambrogioni, L., and van de Meent, J.-W. Categorical flow maps. *arXiv preprint arXiv:2602.12233*, 2026.

Sabour, A., Albergo, M. S., Domingo-Enrich, C., Boffi, N. M., Fidler, S., Kreis, K., and Vanden-Eijnden, E. Test-time scaling of diffusions with flow maps. *arXiv preprint arXiv:2511.22688*, 2025a.

Sabour, A., Fidler, S., and Kreis, K. Align your flow: Scaling continuous-time flow map distillation, 2025b. URL https://arxiv.org/abs/2506.14603.

Silva, D.-A., Yu, S., Ulge, U. Y., Spangler, J. B., Jude, K. M., Labão-Almeida, C., Ali, L. R., Quijano-Rubio, A., Ruterbusch, M., Leung, I., et al. De novo design of potent and selective mimics of il-2 and il-15. *Nature*, 565(7738): 186–191, 2019.

Singhal, R., Horvitz, Z., Teehan, R., Ren, M., Yu, Z., McKeown, K., and Ranganath, R. A general framework for inference-time scaling and steering of diffusion models. *arXiv preprint arXiv:2501.06848*, 2025.

Skreta, M., Akhound-Sadegh, T., Ohanesian, V., Bondesan, R., Aspuru-Guzik, A., Doucet, A., Brekelmans, R., Tong,

A., and Neklyudov, K. Feynman–Kac correctors in diffusion: Annealing, guidance, and product of experts. In *International Conference on Machine Learning*, 2025.

Škrinjar, P., Eberhardt, J., Durairaj, J., and Schwede, T. Have protein-ligand co-folding methods moved beyond memorisation? *BioRxiv*, pp. 2025–02, 2025.

Song, Y. and Dhariwal, P. Improved techniques for training consistency models. In *International Conference on Learning Representations (ICLR)*, 2024.

Song, Y., Dhariwal, P., Chen, M., and Sutskever, I. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Strauch, E.-M., Bernard, S. M., La, D., Bohn, A. J., Lee, P. S., Anderson, C. E., Nieusma, T., Holstein, C. A., Garcia, N. K., Hooper, K. A., et al. Computational design of trimeric influenza-neutralizing proteins targeting the hemagglutinin receptor binding site. *Nature biotechnology*, 35(7):667–671, 2017.

Team, P., Zhang, Y., Gong, C., Zhang, H., Ma, W., Liu, Z., Chen, X., Guan, J., Wang, L., Yang, Y., et al. Protenix-v1: Toward high-accuracy open-source biomolecular structure prediction. *bioRxiv*, pp. 2026–02, 2026.

Tweedie, M. C. Statistical properties of inverse gaussian distributions. ii. *The Annals of Mathematical Statistics*, pp. 696–705, 1957.

Uehara, M., Zhao, Y., Wang, C., Li, X., Regev, A., Levine, S., and Biancalani, T. Inference-time alignment in diffusion models with reward-guided generation: Tutorial and review. *arXiv preprint arXiv:2501.09685*, 2025.

Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976): 1089–1100, 2023.

Wohlwend, J., Corso, G., Passaro, S., Getz, N., Reveiz, M., Leidal, K., Swiderski, W., Atkinson, L., Portnoi, T., Chinn, I., et al. Boltz-1 democratizing biomolecular interaction modeling. *BioRxiv*, pp. 2024–11, 2025.

Yim, J., Campbell, A., Foong, A. Y., Gastegger, M., Jiménez-Luna, J., Lewis, S., Satorras, V. G., Veeling, B. S., Barzilay, R., Jaakkola, T., et al. Fast protein backbone generation with se (3) flow matching. *arXiv preprint arXiv:2310.05297*, 2023.

Zhang, Z., Feng, Y., Song, Y., Qiu, K., Zhou, H., and Ma, W.-Y. DCFold: Efficient protein structure generation with single forward pass. In *International Conference on Learning Representations (ICLR)*, 2026.

Appendix

A. DECAF-SEARCH

Algorithm 2 Inference-time search with DECAF-SEARCH

Require: FlowMap X; noise schedule $\sigma_N > \sigma_{N-1} > \cdots > \sigma_0 = 0$; particles P; resampling interval L; steering weight λ; number of search iterations S; potential R; $\gamma \in [0, 1]$; rollout horizon K (with $K = 1$ for FK).

Ensure: Sample $\hat{x}_0$.

1: Draw $\{x_{\sigma_N}^{(p)}\}_{p=1}^{P} \sim \mathcal{N}(0, \sigma_N^2 I)$ ▷ *initialize P particles*

2: $\mathcal{X} \leftarrow \{x_{\sigma_N}^{(p)}\}_{p=1}^{P}, \quad n \leftarrow 0$ ▷ *global search tree, current noise index*

3: **for** $s = 0, 1, \ldots, S-1$ **do**

4: $\quad \mathcal{T} \leftarrow \emptyset$ ▷ *rollout trajectories*

5: $\quad n_{\text{end}} \leftarrow \min(N,\ n + K)$

6: $\quad$ **for** $m = n, \ldots, n_{\text{end}} - 1$ **do** ▷ *denoise rollout (all particles in parallel)*

7: $\quad\quad \hat{x}_0^{(p)} \leftarrow X(x_{\sigma_m}^{(p)}, \sigma_m, 0), \quad \forall p$

8: $\quad\quad$ *Optional:* gradient steps on $\hat{x}_0^{(p)}$ w.r.t. R ▷ *see Equation (12)*

9: $\quad\quad \tilde{\sigma}_{m+1} \leftarrow \sqrt{1-\gamma^2}\,\sigma_{m+1}$ ▷ *γ renoise*

10: $\quad\quad \tilde{x}^{(p)} \leftarrow \hat{x}_0^{(p)} + \tilde{\sigma}_{m+1} \dfrac{x_{\sigma_m}^{(p)} - \hat{x}_0^{(p)}}{\sigma_m}, \quad \forall p$

11: $\quad\quad \epsilon^{(p)} \sim \mathcal{N}(0, I), \quad \forall p$

12: $\quad\quad x_{\sigma_{m+1}}^{(p)} \leftarrow \tilde{x}^{(p)} + \gamma\,\sigma_{m+1}\,\epsilon^{(p)}, \quad \forall p$

13: $\quad\quad \mathcal{T} \leftarrow \mathcal{T} \cup \{(x_{\sigma_{m+1}}^{(p)},\ m+1)\}_{p=1}^{P}$

14: $\quad$ Compute particle scores $R^{(p)}(\hat{x}_0^{(p)}), \forall p$

15: $\quad$ **if** SEARCH = MCTS **then**

16: $\quad\quad \mathcal{X} \leftarrow \mathcal{X} \cup \mathcal{T}$

17: $\quad\quad$ Update weights of all $(x, m) \in \mathcal{T}$ ▷ *backup*

18: $\quad\quad$ Draw $(x_{\sigma_i}^{(p)}, i) \sim \mathcal{X}$ according to UCT, $\forall p$

19: $\quad$ **else if** SEARCH = FK **then** ▷ *$K = 1$, so $\mathcal{T}$ holds one entry per particle*

20: $\quad\quad (x_{\sigma_i}^{(p)}, i) \leftarrow$ the entry of $\mathcal{T}$ for particle $p, \forall p$

21: $\quad\quad$ **if** $s \bmod L = 0$ **then**

22: $\quad\quad\quad$ Resample $\{x_{\sigma_i}^{(p)}\}_{p=1}^{P}$ with weights $\propto R^{(p)}$

23: $\quad n \leftarrow i + 1$ ▷ *advance to next noise level*

24: **return** $x_0^{(p^\star)}$ with $p^\star = \arg\max_p \{R(x_0^{(p)}) : (x_0^{(p)}, 0) \in \mathcal{X}\}$

DECAF-SEARCH (MC-Grad). As a slight extension of Algorithm 2, we realized that the gradient estimate in $\nabla_{x_{\sigma_0}} R(x_{\sigma_0})$ can be a noisy estimate of the optimal guidance direction. We hypothesize that a Monte Carlo average of several gradients might be more favorable, as observed previously (Holderrieth et al., 2026a). In order to do so, we set

$$x_0 \leftarrow x_0 + \frac{\beta}{L} \sum_{l=1}^{L} w_l \nabla_{x_0^l} R(x_0^l) \tag{13}$$

where weights w_l are the softmax of importance logits derived from the local reward and the renoise prior, $\beta = (\sigma/\sigma_{\text{data}}^2)$, and x_0^l is obtained by renoising x_σ back to $x_{\tilde{\sigma}}$ and then performing the rollouts step in line 7 (i.e. this induces one extra loop that we omit here for readability).

B. Experimental Setup

B.1. Architecture and Training

We adopt Boltz-1 original architecture and codebase (Wohlwend et al., 2025) except for the implementation of dual-time conditioning. The denoiser parametrization $u(x_\rho, \rho, \sigma)$ requires fusing two noise levels into the score network's conditioning

Table 4. Shared inference-parameter settings (constant across NFE budgets and methods).

Parameter	Value
EDM noise schedule	Karras et al. (ρ=7)
σ_{min}, σ_{max}, σ_{data}	0.0004, 160, 16
Stochastic-step noise scale (`noise_scale`)	0.901
Per-step gradient updates	20
MCTS selection rule	UCT
UCT exploration constant (c)	1.0
Progressive widening (k, α)	2.0, 0.5

Table 5. Per-method settings on Runs N' Poses and PoseBusters. `Steps`: number of sampling steps. `Params`: per-method hyperparameters that scale the per-pose NFE. All other sampler hyperparameters are fixed across NFE budgets.

Method	Steps	Params	NFE per pose
Runs N' Poses (4 FK particles)			
Boltz-1x / DECAF (FK)	2	4 particles	8
Boltz-1x / DECAF (FK)	5	4 particles	20
Boltz-1x / DECAF (FK)	6	4 particles	24
Boltz-1x / DECAF (FK)	10	4 particles	40
Boltz-1x / DECAF (FK)	12	4 particles	48
DECAF-SEARCH (MC-GRAD)	15	10 samples	150
Boltz-1x	40	4 particles	160
Boltz-1x *(ref.)*	200	4 particles	800
DECAF-SEARCH (MCTS)	10	4 children, 50 simulations	800
PoseBusters (3 FK particles)			
Boltz-1x	$[6, \ldots, 200]$	3 particles	$[18, \ldots, 600]$
Boltz-1x *(ref.)*	200	3 particles	600
DECAF (FK)	6	3 particles	18
DECAF (FK)	8	3 particles	24
DECAF (FK)	10	3 particles	30
DECAF (MC-GRAD)	10	10 samples	100
DECAF (MCTS)	10	5 sims, 4 children	142
DECAF (MC-GRAD)	15	10 samples	150
DECAF (MC-GRAD)	20	10 samples	200
DECAF (MC-GRAD)	25	10 samples	250
DECAF (MCTS)	10	15 sims, 4 children	489

stream. We adopt a dual-time conditioning module from FastGen (Nie et al., 2026) that departs from the single-time conditioning of the EDM teacher. We train DECAF as a distillation head with our $x0$-aligned loss (Equation (10)), while the trunk and EDM modules are frozen and initialized from the Boltz-1 open-source checkpoint. Training uses RCSB PDB structures released before 2021-09-30, filtered to resolution ≤ 9.0 Å. We train for 100 epochs with 51,200 samples per epoch on 64 H200 GPUs, using a per-GPU batch size of 2 (effective batch size 128) and diffusion multiplicity 16. The optimizer is Muon with momentum 0.95 and Nesterov updates, with an AdamW fallback ($\beta_1 = 0.9$, $\beta_2 = 0.95$) for 1D parameters at learning rate ratio 0.1; weight decay is 0.01 throughout. The learning rate follows an AlphaFold 3-style schedule with linear warmup over 1,000 steps to a peak of $1.8e^{-3}$. The σ-weighting follows the Karras EDM schedule ($\sigma_{\text{min}} = 4e^{-4}$, $\sigma_{\text{max}} = 160$, $\rho = 7$, $\sigma_{\text{data}} = 16$), with 10% diagonal samples ($\sigma_r = \sigma_t$) and 90% off-diagonal. We further re-scale loss $\mathcal{L}$ in Equation (10) by $\frac{1}{\sqrt{(\mathcal{L}+1e^{-6})}}$

B.2. Hyperparameters

We report the inference hyperparameters in tables 4 and 5. Values in Table 4 are constant across all NFE budgets and across both methods unless otherwise noted. Values in Table 5 are the only knobs that change with the compute budget.

The key sampler tuning we apply to Boltz-1x in the few-step regime is setting the step scale $\sigma_{\text{scale}} = 1.0$ (vs. the default 1.638 used at 200 steps). With the default 1.638, the SDE diverges under aggressive step-size schedules, producing 0.0% on every metric at 2 steps (Table 1, "Boltz-1x (default)"). Setting noise scale $\eta = 1.0$ recovers stable sampling and is what we

Table 6. Sweep of γ for Alg. 2 on PoseBusters. 10 sampling steps, DECAF-SEARCH (FK), $P = 3$.

γ	RMSD<2 ↑	PB Valid ↑	Success Rate ↑
0.3	75.3	89.2	65.2
0.5	**75.6**	90.3	**65.9**
0.7	72.8	90.3	63.8
1.0	70.3	**90.7**	61.3

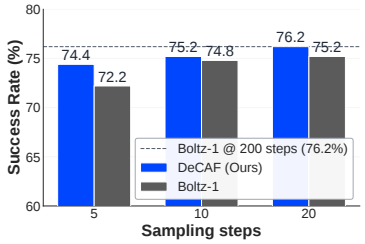

Figure 6. RMSD$<$ 2 Å best@5 on PoseBusters. DECAF (FlowMap sampler Alg. 1, γ=0) vs Boltz-1 (ODE) at matched sampling steps; dashed line marks Boltz-1 at 200 steps.

Table 7. **Success rate breakdown for DECAF-SEARCH as NFEs grow.** We report the fraction of Runs N' Poses that succeed along the DECAF-SEARCH Pareto front, highlighting that while increasing NFEs does improve accuracy, the low-NFE variants do offer complementary successes to the most expensive MCTS-based sampling method.

Outcome	count	%
All three succeed	479	68.23%
Only DECAF FK (NFE=40) + DECAF MCTS (NFE=800) succeed	24	3.42%
Only DECAF FK (NFE=20) + DECAF MCTS (NFE=800) succeed	7	1.00%
Only DECAF FK (NFE=20) + DECAF FK (NFE=40) succeed	7	1.00%
Only DECAF MCTS (NFE=800) succeeds	25	3.56%
Only DECAF FK (NFE=40) succeeds	7	1.00%
Only DECAF FK (NFE=20) succeeds	7	1.00%
None succeeds	146	20.80%
Total	702	100.00%

report as "Boltz-1x (tuned)" throughout.

C. Additional Ablations

C.1. Sampler ablations

Stochasticity γ sweep. Table 6 sweeps the CTM stochasticity parameter $\gamma \in \{0.3, 0.5, 0.7, 1.0\}$ for Algorithm 2 on PoseBusters, holding the rest of the sampler fixed (10 denoising steps, DECAF-SEARCH (FK), $P = 3$ FK particles). We adopt $\gamma = 0.5$, which we hypothesize balances the determinism of $\gamma \to 0$ against the variance of sampling at γ=1. Empirically, Success Rate peaks at γ=0.5 (65.9%), with a 4.7-pt spread between the best (γ=0.5) and worst (γ=1.0) settings, while PB-validity is essentially saturated across all γ ($\Delta \leq 1.5$ pp).

DECAF vs Boltz-1 without inference scaling. Figure 6 isolates the sampler axis: both methods run unguided ODE-style integration so the comparison reflects the underlying sampler quality in isolation rather than any inference-time augmentation. DECAF's FlowMap sampler (taking velocity predictions directly from the trained flow map at each step) outperforms Boltz-1's standard ODE solver at every sampling-step count — at 5 / 10 / 20 steps respectively by +2.2, +0.4, and +1.0 pp on RMSD$<$ 2 Å. Most notably, DECAF at 20 steps (76.2%) *matches* Boltz-1 at 200 steps (76.2%, dashed reference), a 10× reduction in inference compute.

C.2. Qualitative analysis

PoseBusters quality checks. Table 8 collates statistics of the different PoseBusters sub-check failures for DECAF-SEARCH and Boltz-1x. As noted in the main text Section 4.3, we see high failure rates due to lack of sp2-hybridized bond flatness across all models, likely due to suboptimal reward design. Other checks are failed much less frequently and generally improve with increasing NFE (consider e.g. the bond angles failure mode). In Figure 7 we highlight some notable PoseBusters failures, including examples of the pervasive sp2 planarity issue (middle row). We also flag some false positives of the PoseBusters checks, where accurate poses which capture ionic coordination are flagged as having incorrect distance and volume overlap.

Table 8. **Per-complex failure rates by PoseBusters sub-check** (best@5 over PoseBusters metrics). DECAF-SEARCH exceeds Boltz-1x at 600 NFE across all NFE levels, and we see generally increasing pose quality with increasing NFE. Lower is better for all entries in this table.

Sub-check	**DECAF-SEARCH**				**Boltz-1x**
	NFE = 20	NFE = 40	NFE = 75	NFE = 489	NFE = 600
double bond flatness	6.34%	8.10%	7.27%	7.29%	7.83%
volume overlap with protein	2.82%	2.11%	0.69%	2.43%	3.20%
bond angles	3.87%	2.11%	0.35%	0.00%	0.00%
bond lengths	1.76%	1.41%	0.00%	0.00%	0.00%
volume overlap with organic cofactors	1.06%	1.76%	0.00%	1.07%	1.07%
minimum distance to protein	0.35%	0.00%	0.35%	0.00%	0.36%
minimum distance to organic cofactors	0.00%	0.00%	0.35%	0.00%	0.36%
minimum distance to waters	0.00%	0.00%	0.35%	0.00%	0.36%
internal steric clash	0.70%	0.00%	0.00%	0.00%	0.00%
PB-valid pass rate	89.79%	90.14%	**92.04%**	<u>90.62%</u>	88.97%

Practically-relevant slices of Runs N' Poses. As we note in Section 4.3, it is critical to validate the performance of DECAF on systems that are relevant for users of cofolding models. To emulate this setting, we slice Runs N' Poses to a subset of structures that highlight common settings in small-molecule drug discovery and report results in Table 3. We note that DECAF-SEARCH has strong performance at both low and high NFE, and matches or outperforms Boltz-1x at similar NFEs. We also highlight select structures from each category in Figure 8.

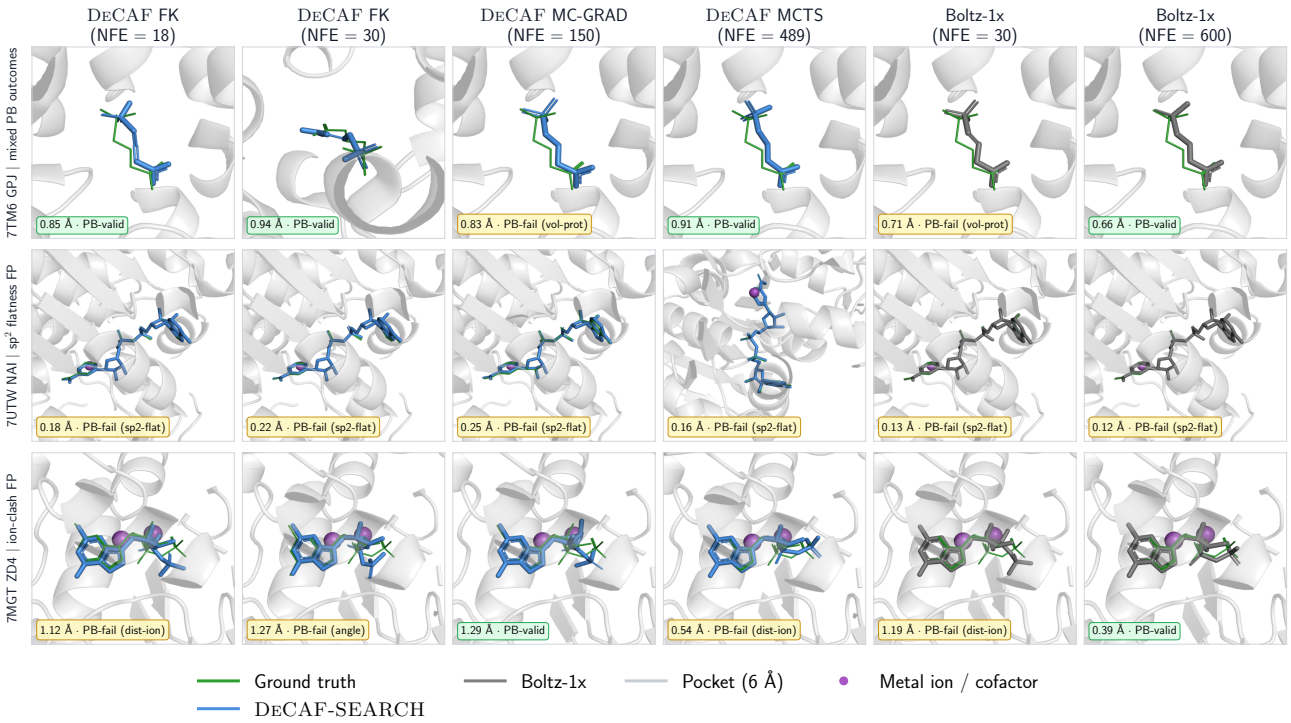

Figure 7. **Example failure modes due to PoseBusters checks** DECAF predictions blue, Boltz-1 gray. Pocket cartoon (gray) is the residues within 6 Å of the crystal ligand (green).

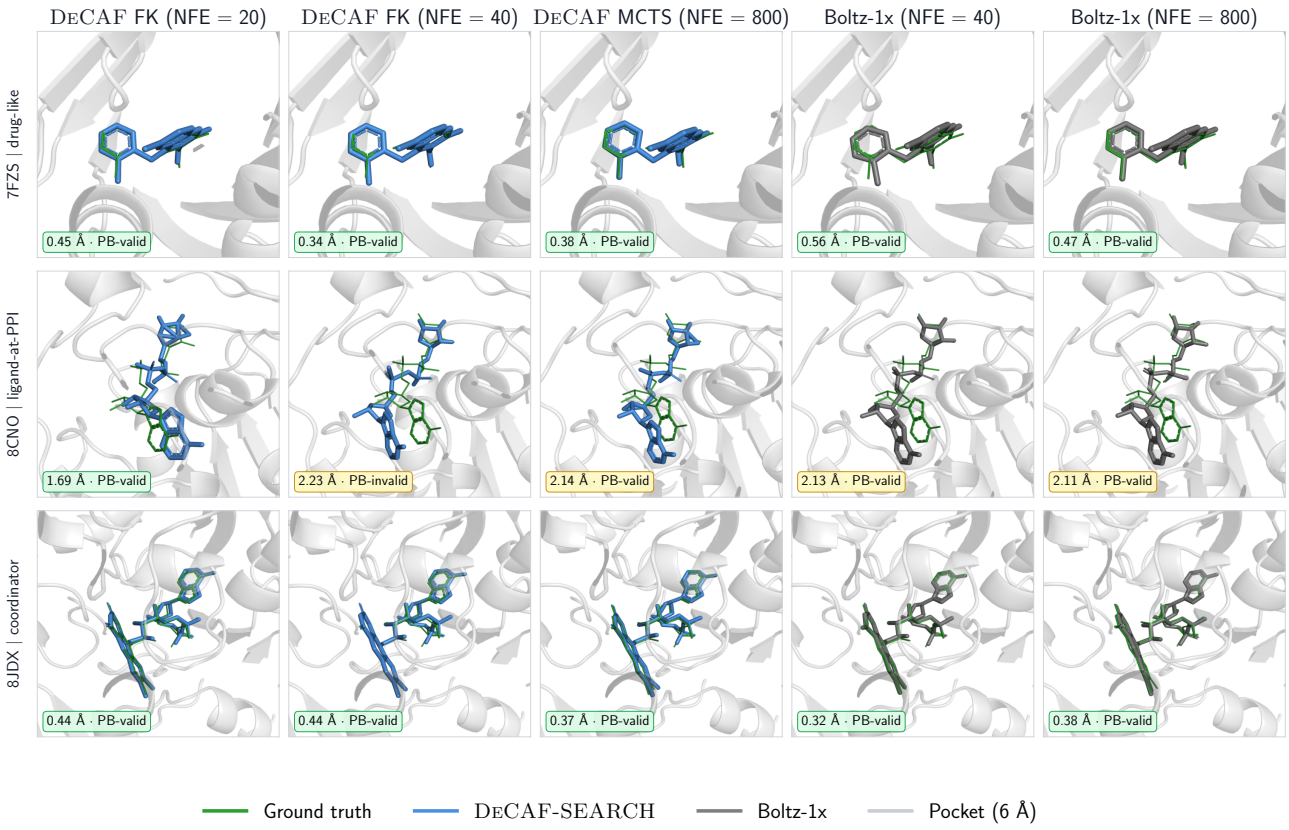

Figure 8. **Qualitative grid for chemically-relevant subsets of Runs N' Poses.** DECAF predictions blue, Boltz-1 gray. Pocket cartoon (gray) is the residues within 6 Å of the crystal ligand (green).

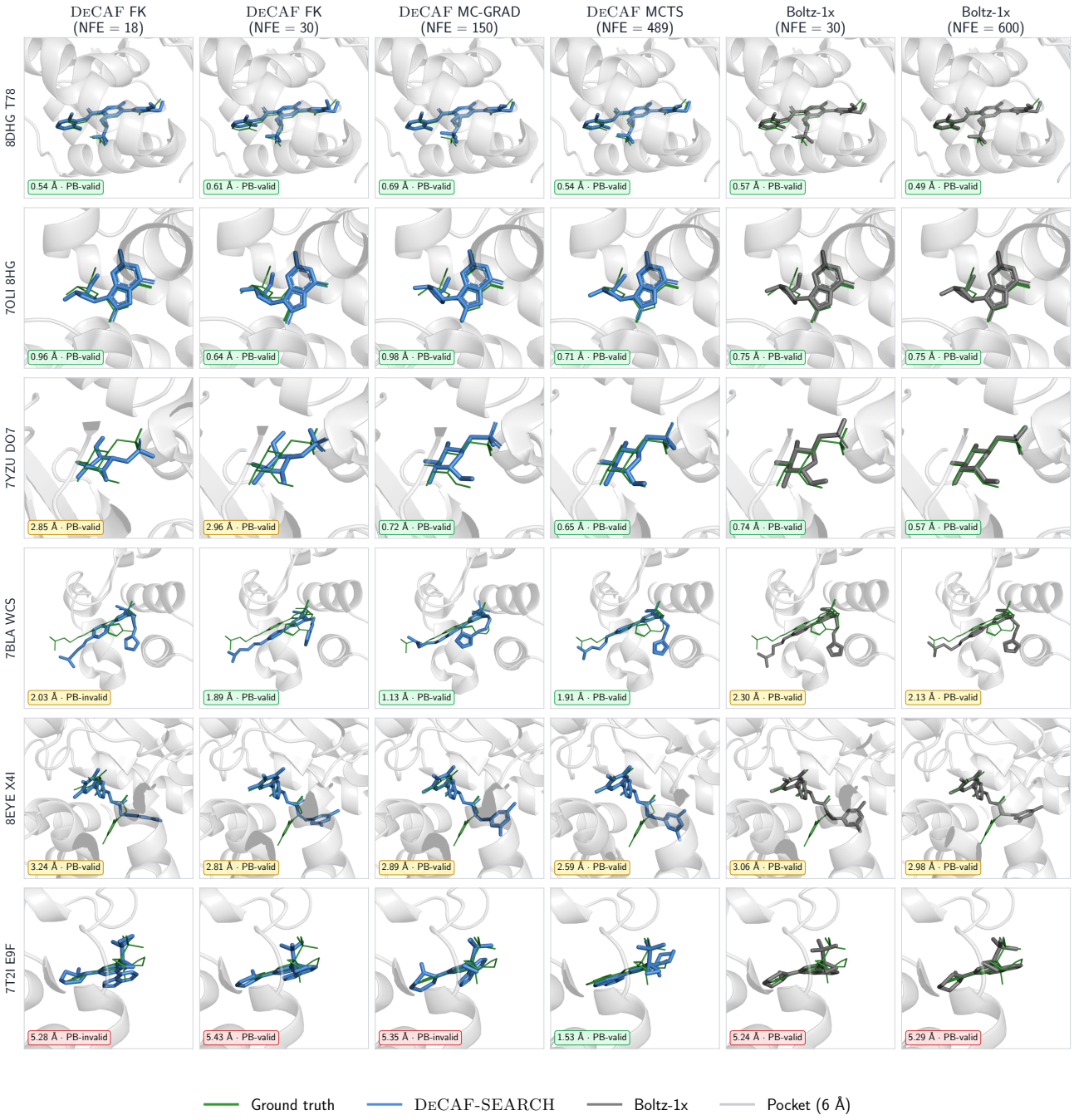

Figure 9. **Multi-method qualitative grid for PoseBusters complexes.** DECAF predictions blue, Boltz-1 gray. Pocket cartoon (gray) is the residues within 6 Å of the crystal ligand (green).