GEOMETRY-CONSISTENT FLOW DISTILLATION FOR UNCERTAINTY-AWARE PROTEIN STRUCTURE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion and flow-matching models have achieved remarkable performance in protein structure generation and ensemble modeling, yet their practical use is often computationally constrained at scale by expensive multi-step sampling and costly iterative refinement. Recent work has explored applying distillation to protein generative models. However, directly transplanting the methods in image generation to 3D proteins overlooks fundamental geometric and physical constraints, most notably SE(3) gauge freedom and frame-dependent supervision, leading to single-step collapse with poor accuracy and structural quality. To address this challenge, we develop a geometry-consistent distillation (GCD) framework that makes one-step protein generation stable and well-posed. Our formulation captures local backbone correlations by explicitly coupling neighboring residues and further integrates an SE(3)-invariant corruption interface and a residue-local, frame-aligned distillation objective. For flow-matching teachers, we further derive the corresponding harmonic-prior score–velocity identity, connecting the teacher's vector-field parameterization to score-based distillation. Empirically, naive Euclidean distillation collapses in the one-step setting, whereas GCD yields stable and competitive one-step students and transfers across multiple distillation paradigms. These results show that geometry-consistent supervision enables aggressive acceleration of protein generative models while retaining strong structural fidelity and substantial ensemble diversity.

1 INTRODUCTION

Computational protein design has become a driving engine for modern biological research and discovery, enabling efficient design and engineering of proteins with tailored structures and functions, spanning enzymes, binders, assemblies, and conformational switches (Baker, 2019). Over the last decade, the field has moved from physics-based pipelines to hybrid paradigms that combine learned priors with structure- and sequence-level constraints. These paradigms also expand the search space and support more diverse design objectives (Pan & Kortemme, 2021; Albanese et al., 2025). In modern workflows, structure generation and sequence design are tightly coupled. Given a candidate backbone, neural inverse-folding models such as ProteinMPNN (Dauparas et al., 2022), together with context-aware extensions, can rapidly produce sequences compatible with both the geometry and the binding context (Dauparas et al., 2025). This coupling motivates generative models that can efficiently propose high-quality backbone candidates at scale, since downstream screening and optimization typically demand thousands to millions of samples.

A key catalyst has been the emergence of highly accurate structure predictors, most prominently AlphaFold (Jumper et al., 2021; Abramson et al., 2024), which jointly reasons over a per-residue representation and a pair representation encoding residue–residue relationships. These representations are coupled to a structure module built around invariant point attention, enabling SE(3)-equivariant geometric reasoning in residue frames while modeling pairwise geometric constraints (Jumper et al., 2021; Bouatta et al., 2021). These inductive biases have, in turn, influenced a cascade of generative protein models that operate directly in 3D, where correctness is judged not only by coordinate proximity but by rigid-motion–invariant, local geometric fidelity required for designability.

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

Recent protein generative modeling is dominated by diffusion and, increasingly, flow matching, enabling high-quality and controllable backbone and complex generation at scale (Watson et al., 2023; Ingraham et al., 2023; Yim et al., 2023; Jing et al., 2024; Geffner et al., 2025). Despite this progress, both families typically rely on iterative sampling, requiring many expensive network evaluations and limiting throughput in large-scale design pipelines. In vision, this speed bottleneck has been alleviated by distilling iterative diffusion/flow teachers into one- or few-step generators (Song et al., 2023; Yin et al., 2024; Zhou et al., 2024). For protein 3D generation, however, this transfer is not direct. Protein conformations are defined only up to global rigid motions; equivalently, they reside in the quotient space $\mathbb{R}^{3N}/SE(3)$. Objectives expressed in absolute Cartesian coordinates can therefore assign different losses to structures that are physically identical up to rotation or translation. Therefore, distillation for protein 3D generation must reformulate both how corrupted intermediate states are constructed and how teacher–student supervision signals are defined modulo global rigid motions.

Our main contributions are:

- We propose a geometry-consistent distillation (GCD) framework for protein 3D generation that addresses the collapse observed for one-step distillation in the presence of SE(3) gauge freedom. Rather than treating geometry as a final correction to an otherwise Euclidean distillation objective, GCD incorporates geometry directly into the distillation interface. Specifically, generator samples are corrupted through a harmonic-prior-based, gauge-fixed interpolation path, while teacher–student supervision is defined through residue-local, frame-aligned signals.
- For flow-matching teachers, we further derive the corresponding score-based distillation formulation under the harmonic prior, including the score–velocity identity and its relation to the practical Moore–Penrose implementation.
- Empirically, naive Euclidean one-step distillation collapses in this setting, whereas GCD produces stable one-step students and extends naturally to other distribution-matching distillation and self-consistency-style distillation objectives.
- One-step GCD offers the largest inference speedup while retaining strong structural fidelity and substantial ensemble diversity. Additional sampling steps provide a controllable fidelity–coverage trade-off.

2 Background and related work

Protein structure prediction. AlphaFold (Jumper et al., 2021) pioneered the modern end-to-end data-driven paradigm for protein structure prediction. Given an amino-acid sequence, it uses evolutionary context from a multiple sequence alignment (MSA) (and optionally template structures) to predict a single high-accuracy 3D structure, trained with a regression-style geometry loss, such as the FAPE loss, on PDB structures. Subsequent models such as ESMFold (Lin et al., 2023) replace the MSA pipeline with embeddings from a protein language model (PLM) and typically omit template input, while largely retaining the AlphaFold-style architecture and training framework, yielding a fast and practical sequence-to-structure predictor.

Flow matching. Flow matching is a simulation-free approach for training continuous-time generative models by regressing a time-dependent vector field to the conditional flow induced by a prescribed probability path between a prior distribution p_0 and the data distribution p_1 (Lipman et al., 2022). Concretely, one specifies a conditional probability path $p_t(x \mid x_1)$ (or equivalently an interpolant construction) for $t \in [0,1]$, and learns a vector field $v_\theta(x,t)$ such that the Ordinary Differential Equation(ODE) $\frac{dx}{dt} = v_\theta(x,t)$ transports samples from p_0 to p_1.

A simple interpolant uses linear mixing between $x_0 \sim p_0$ and $x_1 \sim p_1$:$x_t = (1-t)x_0 + tx_1$. For this interpolant, the induced conditional flow (velocity) along the path can be written as $u_t(x \mid x_1) = \frac{x_1 - x}{1-t}$ (i.e., $x_0 = \frac{x - tx_1}{1-t}$ for $x = x_t$ and $\frac{dx_t}{dt} = x_1 - x_0$). Flow matching trains v_θ by minimizing a regression objective of the form $\mathbb{E}_{t\sim\mathrm{Unif}[0,1],\, x_1\sim p_1,\, x\sim p_t(\cdot|x_1)}\left[\|v_\theta(x,t) - u_t(x \mid x_1)\|^2\right]$. Moreover, for linear interpolants one can reparameterize the vector field by predicting a denoised target $\hat{x}_1(x,t)$ and setting $v_\theta(x,t) = \frac{\hat{x}_1(x,t) - x}{1-t}$, which yields a denoising-style objective closely related to diffusion training (Albergo et al., 2023).

AlphaFlow. AlphaFlow (Jing et al., 2024) leverages this perspective by extending AlphaFold/ESMFold to accept a noisy structure input (analogous to noisy images in diffusion), using an

input embedding module similar to AlphaFold's template embedding stack, and fine-tuning pretrained weights under a flow-matching objective. The derived AlphaFlow/ESMFlow enable iterative sampling of diverse 3D conformations, achieving strong precision–diversity tradeoffs on PDB ensembles and capturing flexibility and distributional properties when further trained on all-atom molecular dynamics ensembles.

Diffusion distillation. Diffusion models define a forward noising process that gradually perturbs clean data x_0 into x_t over a noise level (or time t). A common Gaussian parameterization writes $x_t = \alpha_t x_0 + \sigma_t \epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$ and α_t, σ_t are predetermined schedules. For this forward kernel, the conditional score satisfies $\nabla_{x_t} \log q(x_t \mid x_0) = \frac{\alpha_t x_0 - x_t}{\sigma_t^2}$, and the (marginal) data score can be written as

$$\nabla_{x_t} \log q_t(x_t) = \mathbb{E}\left[\frac{\alpha_t x_0 - x_t}{\sigma_t^2} \,\middle|\, x_t \right]. \tag{1}$$

Given a pretrained diffusion teacher that predicts a denoised estimate $f_\phi(x_t, t) \approx \mathbb{E}[x_0 \mid x_t]$, one obtains a teacher score estimator $s_\phi(x_t, t) = \frac{\alpha_t f_\phi(x_t, t) - x_t}{\sigma_t^2}$. Zhou et al. (2024) introduced Score identity Distillation (SiD) to distill such a teacher into a fast (e.g., one-step) generator by matching score-related identities of the generator distribution to those implied by the teacher across noise levels. Because the generator's score is not directly available, SiD introduces an auxiliary score network(projector) $s_\psi(x_t, t)$ trained on samples from the generator (after applying the same forward noising) to approximate the generator scores. Distillation is then performed by minimizing a score-matching objective (equivalently, a Fisher-divergence surrogate) of the form

$$\mathbb{E}_t \, \mathbb{E}_{x \sim p_t^\theta} \left[\| s_\psi(x, t) - s_\phi(x, t) \|^2 \right], \tag{2}$$

where p_t^θ denotes the distribution of noised generator samples at noise level t. This formulation enables transferring the teacher's generative behavior to a much faster student without running the full multi-step reverse process at inference time.

2.1 Related work

Recent protein generative modeling is dominated by diffusion and, increasingly, flow-based formulations. Diffusion models have enabled controllable backbone and complex generation, with representative systems including RFdiffusion (Watson et al., 2023), Chroma (Ingraham et al., 2023), and frame-based SE(3) diffusion models such as FrameDiff (Yim et al., 2023); however, high-fidelity sampling typically requires many iterative denoising steps, which limits throughput in large-scale design loops. In parallel, flow matching offers a simulation-free way to train continuous-time generative models (Lipman et al., 2022) and has led to strong protein ensemble and backbone generators, including AlphaFlow/ESMFlow (Jing et al., 2024) as well as large-scale flow-based models such as FoldFlow-2 and Proteína (Geffner et al., 2025). While flow-based samplers can be competitive at relatively small step counts, they still rely on multi-step numerical integration and do not, by themselves, reduce inference to one or a few network evaluations. In image generation, this speed bottleneck has been addressed by distilling iterative diffusion/flow teachers into fast generators, including consistency-style methods (Song et al., 2023) and distributional or score-based distillation objectives such as DMD and related formulations (Yin et al., 2024; Luo et al., 2023; Zhou et al., 2024). Protein distillation is more challenging: AlphaFlow explores supervised single-pass distillation by regressing a student to teacher rollouts (Jing et al., 2024), which can be training-expensive and may underconstrain the student distribution when only final outputs are supervised, and recent studies report that naive one-step adaptations can degrade designability and require additional stabilization such as inference-time noise modulation (Xie et al., 2025). Together, these lines of work suggest that effective step reduction for protein generators is tightly coupled to how SE(3) symmetry and residue-level geometric constraints are imposed during distillation.

3 Methods

We aim to distill an iterative flow-matching protein generator into a fast one-step or few-step student for protein 3D generation. We propose a geometry-consistent distillation (GCD) framework that addresses the collapse observed for naive Euclidean one-step distillation in the presence of SE(3) gauge freedom by incorporating geometry directly into the distillation interface, rather than

treating geometry as a final correction to an otherwise Euclidean distillation objective. Section 3.1 introduces the harmonic-prior corruption path, including the gauge-fixed interpolation construction, the score–velocity identity under the harmonic prior, and its relation to the practical Moore–Penrose implementation. Section 3.2 defines the residue-local, frame-aligned projector–generator distillation objectives. Section 3.3 describes the one-step and few-step implementations used for fast sampling and diversity recovery.

Let (A, M) denote the amino-acid sequence and its mask. We denote the frozen flow-matching teacher by f_ϕ, the student generator by G_θ, and the auxiliary projector by f_ψ. Given (A, M), the generator outputs a structure $S_g = G_\theta(A, M)$, and we write $x_g = [S_g]_\beta \in \mathbb{R}^{N\times 3}$ for its pseudo-β coordinates, using C_α for glycine. We use uppercase S for structural outputs, including residue frames and atomic coordinates, and lowercase x for pseudo-β coordinate representations used to define the corruption path.

3.1 Harmonic-prior corruption path

Harmonic polymer prior. We represent conformations by pseudo-β coordinates $x \in \mathbb{R}^{N\times 3}$, using C_α for glycine. To define the corruption path for generator samples, we use a harmonic polymer prior

$$q_0(x_0) \;\propto\; \exp\left(-\frac{\lambda}{2}\sum_{i=1}^{N-1}\|x_{0,i} - x_{0,i+1}\|^2\right), \tag{3}$$

which corresponds to the chain-graph Laplacian precision $\Lambda = \lambda(L \otimes I_3)$, where L is the path-graph Laplacian over residues; see eq. (14). Since Λ has a global-translation nullspace, the score identities are derived using the regularized precision $\Lambda_\varepsilon = \Lambda + \varepsilon I_{3N}$. For a clean generator sample x_g, the linear interpolation $x_t = (1-t)x_0 + tx_g$ with $x_0 \sim \mathcal{N}(0, \Lambda_\varepsilon^{-1})$ induces the conditional corruption kernel

$$x_t \mid x_g \sim \mathcal{N}\big(tx_g,\ (1-t)^2\Lambda_\varepsilon^{-1}\big)\,. \tag{4}$$

Compared with isotropic Gaussian corruption, this kernel introduces chain-correlated perturbations and determines the Laplacian metric that appears in the score–velocity relation below.

Gauge-fixed path construction. Given a generator sample x_g and a base sample $x_0 \sim q_0$, we remove the global gauge via frame alignment minimizing Root-Mean-Square-Deviation (RMSD) and define the interpolation path by sampling $t \sim \mathcal{U}(0, 1)$

$$x_t \;=\; (1-t)\,\tilde{x}_0 \;+\; t\,x_g,\;\; \tilde{x}_0 = \mathrm{RMSDAlign}(x_0, x_g). \tag{5}$$

Remark. Because the chain-Laplacian precision has a translation nullspace, the regularized covariance Λ_ε^{-1} is used in the derivation. In implementation, we work on the centered subspace and sample the harmonic prior using the Moore–Penrose pseudoinverse $\Lambda^\dagger$, obtained by zeroing the global-translation eigenmodes. Equivalently, $\Pi_{\mathcal{H}}(\Lambda + \varepsilon I)^{-1}\Pi_{\mathcal{H}} \;\longrightarrow\; \Lambda^\dagger, \qquad \mathcal{H} = \mathrm{Null}(\Lambda)^\perp$, as $\varepsilon \to 0$, where $\Pi_{\mathcal{H}}$ denotes projection onto the centered subspace. Details are provided in Appendix A.2.

Score–velocity identity under the harmonic prior. Flow matching learns a time-dependent vector field along a prescribed interpolation path, whereas score-based distillation is formulated in terms of marginal scores. To distill a flow-matching teacher with such an objective, we derive the score–velocity identity based on the harmonic-prior. Under the regularized path $x_t = (1-t)x_0 + tx_1$, $x_0 \sim \mathcal{N}(0, \Lambda_\varepsilon^{-1})$, the marginal score satisfies

$$\nabla_x \log p_t(x) = -\frac{1}{(1-t)^2}\Lambda_\varepsilon\left(x - t\,\mathbb{E}[x_1 \mid x_t = x]\right). \tag{6}$$

Equivalently, defining the marginal velocity $u_t(x) = \frac{\mathbb{E}[x_1 | x_t = x] - x}{1-t}$, we have

$$\nabla_x \log p_t(x) = \frac{1}{1-t}\Lambda_\varepsilon\left(tu_t(x) - x\right). \tag{7}$$

For a structural denoiser f, let $x_f(x_t, t) = [\hat{S}_f(A, M, x_t, t)]_\beta$ denote the pseudo-β coordinates extracted from its denoised structural prediction. Replacing the conditional mean in (6) with $x_f(x_t, t)$

gives the harmonic-prior induced score

$$s_f(x_t, t) = -\frac{1}{(1-t)^2}\Lambda_\epsilon\big(x_t - t x_f(x_t, t)\big). \tag{8}$$

The proof is given in Appendix A.2.

3.2 GEOMETRY-CONSISTENT FLOW DISTILLATION

Bilevel projector–generator scheme. GCD trains the distilled generator with a bilevel projector–generator scheme. Given a generator sample $S_g = G_\theta(A, M)$, we construct the corrupted state x_t using the gauge-fixed harmonic path in section 3.1. The projector f_ψ is trained with a flow-matching objective on the generator-induced path, using S_g as the clean endpoint for each corrupted state x_t. Based on the harmonic-prior score–velocity identity, we update G_θ by minimizing a score-matching distillation loss between the frozen teacher score and the generator score represented by the projector.

To remove the global Cartesian gauge from structural comparisons, we express structures in residue-local coordinates. For a structure S, let $\mathcal{R}_i(S)$ denote the representation of S in the local coordinate frame of residue i. For two structures S_1 and S_2, we define their frame-aligned residual as $\widetilde{S}_1 - \widetilde{S}_2 := \mathrm{vec}\left(\{\mathcal{R}_i(S_1) - \mathcal{R}_j(S_2)\}_{1\le i,j\le N}\right)$, where $\mathrm{vec}(\cdot)$ stacks the residue-local coordinates over all residue-frame pairs (i, j).

Projector loss. During the projector update, the generator output is treated as the clean endpoint of the student path. We train the projector with the local-frame flow matching objective

$$\mathcal{L}_\psi = \mathbb{E}_{t,x_t}\left[\left\|\tilde{S}_\psi - \mathrm{sg}(\tilde{S}_g)\right\|_2^2\right], \tag{9}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient.

Generator loss. During the generator update, the teacher is frozen and the projector is fixed. The harmonic-prior SiD derivation gives the following frame-aligned score-distillation objective:

$$\mathcal{L}_\theta = \mathbb{E}_{t,x_t}\left[\omega(t)\frac{t^2}{(1-t)^4}\left((1-\alpha)\left\|\Lambda_\varepsilon\left(\tilde{x}_\phi - \tilde{x}_\psi\right)\right\|_2^2 + \left(\tilde{x}_\phi - \tilde{x}_\psi\right)^\top \Lambda_\varepsilon^2\left(\tilde{x}_\psi - \tilde{x}_g\right)\right)\right]. \tag{10}$$

The complete derivation of generator loss is provided in Appendix A.3.

3.3 ONE-STEP AND FEW-STEP IMPLEMENTATIONS

One-step distillation. In this setting, G_θ generates the final structure in a single sampling step, yielding the largest inference acceleration. The GCD framework makes one-step distillation well-posed under SE(3) gauge freedom and prevents the collapse observed with naive Euclidean distillation. This implementation is therefore most suitable for latency-sensitive or high-throughput structural sampling.

Few-step distillation. The one-step student already retains substantial conformational diversity. We extend the same GCD interface to few-step generators to provide a controllable fidelity–coverage trade-off. Additional refinement steps change the operating point of the sampler: two-step sampling often improves Precision/Recall, whereas four-step sampling produces broader conformational ensembles and improves several ensemble-level observables.

The overall training procedure is summarized in Algorithm 1.

4 EXPERIMENTS

We evaluate our geometry-consistent distillation framework by training one-step and few-step generators distilled from pretrained AlphaFlow teachers (Jing et al., 2024). Following AlphaFlow, we consider two conditional generation regimes: PDB ensemble generation and molecular-dynamics (MD) ensemble generation. For each implementation setting, we draw 50 samples per test target and

Table 1: PDB-ensemble distillation performance and sampling-time evaluation results. **Left:** Median evaluation metrics in percentages are reported based on 100 test target proteins. **Bold** and underlined fonts denote the best and second-best performance in each family. **Right:** Sampling time per protein. The AlphaFlow family is benchmarked on one NVIDIA **A100** GPU, while the ESMFlow family is benchmarked on one **H200** GPU. Mean/median/min/max time per protein are reported on the PDB test dataset.

(a)

Method	Precision	Recall	Diversity
SiD (1 step)	0.49	0.44	0.00
GCD-SiD (1 step)	81.10	<u>80.47</u>	18.31
GCD-SiD (2 step)	82.16	**81.66**	17.39
GCD-SiD (4 step)	79.53	80.00	**21.92**
GCD-DMD (1 step)	<u>83.21</u>	79.59	2.35
GCD-SC (1 step)	**83.49**	79.57	2.59
AlphaFlow (10 step)	80.19	79.78	<u>18.47</u>
GCD-SiD (1 step)	77.50	76.82	18.90
GCD-SiD (2 step)	**79.20**	**78.88**	19.01
GCD-SiD (4 step)	77.10	<u>77.51</u>	**24.41**
ESMFlow (10 step)	<u>77.86</u>	77.41	<u>21.92</u>

(b)

Method	**Mean**	**Median**	**Min**	**Max**	**Speedup**
AlphaFlow family (A100)					
GCD-SiD (1-step)	5:32	4:10	2:19	15:21	**4.67**×
GCD-SiD (2-step)	7:14	6:04	2:14	18:04	3.21×
GCD-SiD (4-step)	12:39	10:30	4:24	35:20	1.86×
Teacher (AlphaFlow)	24:26	19:29	8:42	1:07:15	1.00×
ESMFlow family (H200)					
GCD-SiD (1-step)	1:11	0:54	0:20	4:11	**9.52**×
GCD-SiD (2-step)	2:42	1:46	0:39	17:53	4.85×
GCD-SiD (4-step)	4:58	3:30	1:13	21:10	2.45×
Teacher (ESMFlow)	11:54	8:34	3:00	51:18	1.00×

report aggregated metrics using the exact same evaluation pipeline as AlphaFlow for fair comparison. We first adopt SiD (Zhou et al., 2024) as the primary distillation baseline, and then further combine the proposed GCD framework with multiple representative distillation objectives to assess its generality. Specifically, we instantiate GCD with *DMD* (Yin et al., 2024), a distribution-matching distillation method based on a KL-divergence objective, and self-consistency (*SC*) distillation (Song et al., 2023), where the teacher provides targets that enforce step-to-step consistency of the student. This yields a unified ablation suite for evaluating whether the geometry-consistent distillation interface enables stable and effective instantiations of different distillation formulations in protein 3D generation. For clarity, we denote geometry-consistent variants with the prefix "GCD" (e.g., GCD-SiD, GCD-DMD, and GCD-SC).

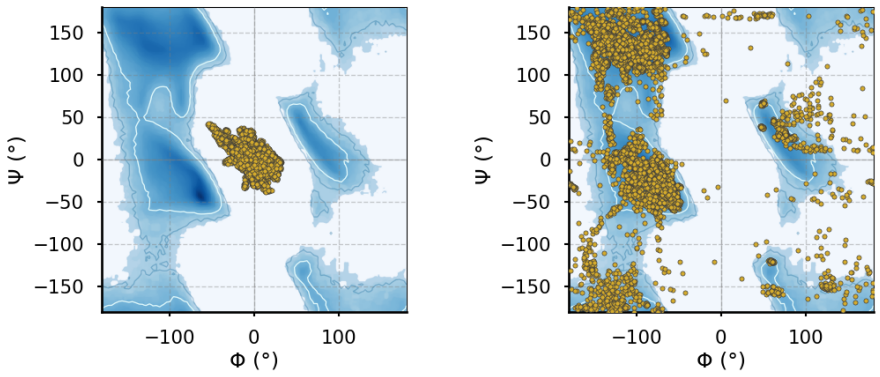
4.1 PDB ENSEMBLE REGIME

We first evaluate whether distilled generators can model proteins with heterogeneous conformations deposited in the Protein Data Bank (PDB) (Berman et al., 2000; wwPDB consortium, 2019). Following AlphaFlow, we use the PDB-ensembles for a test set of 100 targets with experimentally observed conformational variability, including multi-chain complexes. For each target protein, we sample 50 folding structures to form a generated ensemble.

We report three ensemble metrics computed via the local Distance Difference Test, $\mathrm{lDDT}_{C\alpha}$, as in AlphaFold (Jumper et al., 2021) evaluating matching between predicted structures and experimentally observed conformations: **Precision** (avg. $\mathrm{lDDT}_{C\alpha}$ from each prediction to its closest crystal structure), **Recall** (avg. $\mathrm{lDDT}_{C\alpha}$ from each crystal structure to its closest prediction), and **Diversity** (avg. pairwise dissimilarity $1 - \mathrm{lDDT}_{C\alpha}$ among predictions). Precision and Recall respectively measure fidelity and coverage with respect to the crystal ensemble, while Diversity quantifies the spread of the generated ensemble.

Table 1a summarizes the distillation results on the PDB-ensemble benchmark. Overall, the results support three conclusions: (i) directly transferring image-domain distillation methods to protein 3D generation leads to catastrophic failure; (ii) once geometric consistency is enforced, even a one-step GCD-SiD student retains a high-quality and diverse conformational ensemble; and (iii) the number of sampling steps provides a controllable fidelity–coverage trade-off, with two-step sampling generally favoring Precision/Recall and four-step sampling producing broader conformational diversity.

Under the AlphaFlow teacher, SiD (1 step) achieves only 0.49%/0.44% in Precision/Recall with $\sim$0.00 Diversity, indicating that the model nearly degenerates to meaningless outputs or a severely collapsed distribution and thus fails to form a valid conformational ensemble. This failure is closely tied to the geometric nature of protein 3D supervision: if the distillation target is defined in a global coordinate system without explicitly addressing SE(3) rigid-motion freedom and local geometric alignment, the optimization signal is contaminated by coordinate arbitrariness, which destabilizes training and ultimately causes collapse.



(a) SiD (b) GCD-SiD

Figure 1: **Backbone dihedral (Ramachandran) analysis.** Ramachandran plots (ϕ–ψ) aggregated over 50 generated conformations for a representative test target. The naive SiD student exhibits an abnormally concentrated dihedral distribution, whereas GCD-SiD spreads angles across empirically favored regions, indicating improved local backbone geometry.

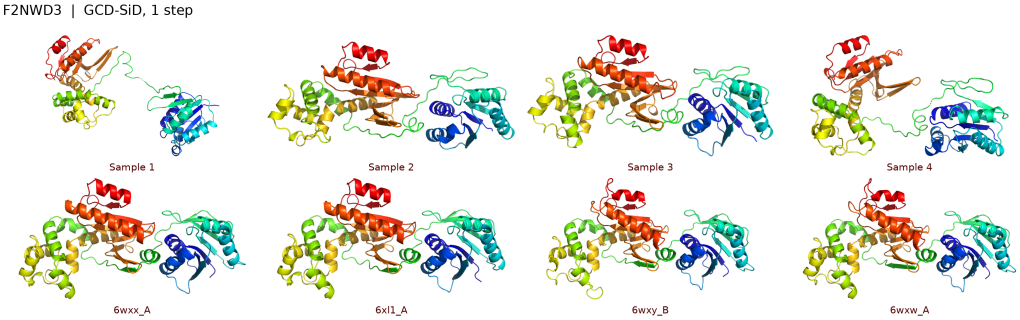


Figure 2: **Distilled conformational sampling example.** Card1 nuclease from *Treponema succinifaciens* (UniProt ID: **F2NWD3**). Four samples from GCD-SiD (top) and four PDB conformers (bottom) are visualized. The samples cover major conformational variations across conformers; additional examples are provided in the appendix.

This collapse is also evident in backbone dihedral statistics. Figure 1 compares Ramachandran plots for a representative test target, aggregating ϕ–ψ angles over the 50 generated conformations. The naive SiD student exhibits an abnormally concentrated cluster of angles in a narrow region, deviating from the canonical α/β basins and indicating severe geometric degeneracy. In contrast, GCD-SiD yields a distribution that is spread across the empirically favored regions, qualitatively consistent with physically plausible backbone geometry. These plots provide a complementary, structure-aware diagnosis that the failure mode is not merely reduced diversity in Cartesian space but a breakdown of local stereochemical validity.

In sharp contrast, GCD-SiD results show that single-step generation need not sacrifice conformational diversity. Under the AlphaFlow teacher, GCD-SiD (1 step) achieves 81.10/80.47/18.31% Precision/Recall/Diversity, compared with 80.19/79.78/18.47% for the 10-step teacher. Thus, a single network evaluation preserves essentially the full teacher performance while providing the largest sampling acceleration. Under the ESMFlow teacher, the one-step student similarly retains 77.50/76.82/18.90%, remaining close to the 77.86/77.41/21.92% teacher despite the substantial reduction in sampling steps. These results show that geometry-consistent distillation does not merely produce a high-precision mode predictor, but can retain substantial conformational coverage even in the single-step regime.

Increasing the number of sampling steps reveals a consistent fidelity–coverage trade-off across both model families. Two-step sampling gives the strongest Precision/Recall, reaching 82.16/81.66% for AlphaFlow and 79.20/78.88% for ESMFlow, while maintaining diversity close to the corresponding teachers. In contrast, four-step sampling broadens the generated ensembles to 21.92% and 24.41%

Table 2: **Evaluation on MD ensembles.** For each method, we compare the predicted ensemble with the reference MD ensemble according to multiple metrics (see the main text for definitions). When applicable, we report the median across the 82 test ensembles. ↑ indicates higher is better; ↓ indicates lower is better.

	AlphaFlow MD	GCD-SiD 1-step	GCD-SiD 2-step	GCD-SiD 4-step	ESMFlow MD	GCD-SiD 1-step	GCD-SiD 2-step	GCD-SiD 4-step
Predicting flexibility								
Pairwise RMSD ($= 2.878$)	2.760	2.188	2.865	3.739	3.280	3.662	4.101	4.823
Pairwise RMSD r ↑	0.420	0.380	0.386	0.388	0.180	0.264	0.247	0.239
All-atom RMSF ($= 1.705$)	1.620	1.124	1.582	2.075	2.140	2.067	2.546	3.002
Global RMSF r	0.490	0.427	0.453	0.470	0.300	0.310	0.307	0.316
Per-target RMSF r	0.840	0.694	0.752	0.787	0.740	0.597	0.643	0.693
Distributional accuracy								
Root mean W_2-dist. ↓	2.600	3.498	3.383	3.351	3.670	4.475	4.437	4.398
↪ Translation contrib. ↓	2.220	2.770	2.613	2.646	3.260	3.542	3.473	3.536
↪ Variance contrib. ↓	1.300	1.987	1.863	1.891	1.760	2.446	2.279	2.554
MD PCA W_2-dist. ↓	1.460	1.637	1.586	1.560	1.620	1.671	1.755	1.677
Joint PCA W_2-dist. ↓	2.250	3.036	3.060	3.032	3.200	4.304	4.239	4.166
% PC-sim > 0.5 ↑	36.590	25.610	28.049	29.268	30.490	21.951	26.829	25.610
Ensemble observables								
Weak contacts J ↑	0.610	0.410	0.522	0.565	0.530	0.379	0.467	0.512
Transient contacts J ↑	0.390	0.271	0.348	0.378	0.320	0.230	0.282	0.312
Exposed residue J ↑	0.480	0.506	0.571	0.625	0.450	0.444	0.518	0.556
Exposed MI matrix ρ ↑	0.200	0.096	0.200	0.255	0.160	0.111	0.167	0.231

Diversity, respectively, exceeding the diversity of both 10-step teachers, with only modest changes in Precision/Recall. Importantly, diversity is not simply monotonic between one and two steps: the one-step models already capture substantial conformational spread. Rather, additional refinement steps shift the operating point from fast teacher-like sampling toward progressively broader ensemble coverage.

The table also includes GCD-DMD and GCD-SC as additional controls, demonstrating that the geometry-consistent formulation can be incorporated into multiple distillation paradigms.

4.2 Sampling efficiency

As shown in Table 1b (Fig. 3 for visualization in the Appendix), we report per-protein sampling time on the PDB test set together with median-based speedups. As expected, fewer student steps lead to faster inference. The one-step GCD-SiD student achieves the largest acceleration, with $4.67\times$ speedup for the AlphaFlow family on A100 and $9.52\times$ for the ESMFlow family on H200. Two- and four-step variants provide intermediate operating points, trading some speed for different fidelity–coverage profiles, with four-step sampling producing the broadest ensembles. Because runtimes vary substantially across targets due to sequence length and conditioning complexity, we report the mean/median/min/max and compute speedups from median runtimes under the same hardware setting.

4.3 Molecular dynamics ensemble regime

To assess distillation performance considering structural diversity covering protein folding dynamics, we evaluate on molecular-dynamics (MD) ensembles from the ATLAS dataset (Vander Meersche et al., 2024). ATLAS provides all-atom, explicit-solvent MD trajectories for 1,390 non-membrane proteins, with three replicate simulations per selected protein sequence (100 ns; 10,000 frames each). We use AlphaFlow-MD as the teacher model. For each test target protein, we generate an ensemble of 50 samples conditioned on its sequence and compare it to the corresponding reference MD ensemble. In the main evaluation, we pool replicate trajectories to form the reference ensemble.

We report MD-ensemble metrics grouped into three categories: (i) *flexibility*, measuring agreement in intra-ensemble structural variability, e.g., global pairwise RMSD and local residue-level Root-Mean-Square-Fluctuation (RMSF); (ii) *distributional accuracy*, quantified by Wasserstein-style distances between MD and generated ensembles, including root-mean and PCA-projected Wasserstein-2 distances W_2, as well as the similarity of dominant motion dynamics via computing the cosine similarity between top principal components (PC-sim); (iii) *ensemble observables*, evaluating agreement on MD-derived signals such as weak/transient contacts and cryptic solvent-exposure statistics.

As reported in Table 2, GCD provides a controllable efficiency–coverage trade-off in the MD-ensemble regime. Notably, the updated results show that even single-step GCD recovers substantial conformational variability, rather than collapsing to a narrowly concentrated ensemble. Increasing the number of sampling steps further expands the generated ensemble and generally improves the recovery of ensemble-level observables. However, these gains are not uniformly monotonic across all distributional metrics, indicating that additional steps primarily modulate ensemble spread and coverage rather than simply improving fidelity in every respect.

The effect is especially clear from the flexibility metrics. For the AlphaFlow-based student, pairwise RMSD increases from 2.188 at one step to 2.865 at two steps, nearly matching the reference MD value of 2.878, while the corresponding all-atom RMSF increases from 1.124 to 1.582, approaching the MD reference of 1.705. Increasing to four steps further broadens the ensemble (pairwise RMSD 3.739; RMSF 2.075), slightly overshooting the reference fluctuation magnitude. The ESMFlow-based student exhibits a similar but stronger broadening trend, with pairwise RMSD increasing from 3.662 to 4.823 from one to four steps. Thus, the number of sampling steps acts as an effective control over conformational spread, with the optimal degree of refinement depending on the underlying teacher family.

Correlation-based flexibility metrics show a complementary pattern. Per-target RMSF correlation generally improves as more steps are used, indicating better recovery of the relative distribution of flexible and rigid regions across individual proteins. In contrast, global RMSF and pairwise-RMSD correlations change more modestly, suggesting that increasing ensemble spread does not automatically translate into uniformly improved global calibration.

Distributional metrics exhibit a more mixed dependence on the number of steps. Root-mean W_2 distance improves moderately with additional steps for both model families, while its translation and variance components, as well as PCA-projected W_2 metrics, are not strictly monotonic. For example, the AlphaFlow-based student achieves its lowest translation and variance contributions at two steps, whereas four steps produce a broader ensemble without a corresponding improvement in all distributional distances. These results indicate that few-step sampling improves coverage of the conformational landscape, but excessive broadening can introduce a trade-off between ensemble diversity and precise distributional calibration.

In contrast, the ensemble observables show a substantially more consistent benefit from additional sampling steps. For both AlphaFlow- and ESMFlow-based students, the Jaccard similarities of weak and transient contacts, exposed-residue agreement, and exposure mutual-information correlation generally increase from one to four steps. These observables depend on coordinated fluctuations and relatively infrequent conformational events, and their improvement suggests that additional refinement steps help recover biologically relevant ensemble behavior even when global distributional metrics have already saturated or become non-monotonic. Overall, GCD supports a tunable sampling regime: one-step generation provides the greatest acceleration while retaining meaningful MD-like variability, whereas a small number of additional steps increases conformational coverage and improves higher-order ensemble observables.

5 Conclusions

We presented **Geometry-Consistent Distillation (GCD)**, a framework for distilling iterative flow-matching protein structure generators into fast one-step and few-step students. GCD makes the distillation interface itself geometry-consistent by combining a harmonic-prior-based, gauge-fixed corruption path with residue-local, frame-aligned teacher–student supervision, and derives the corresponding harmonic-prior score–velocity formulation for flow-matching teachers. Across PDB and MD ensemble benchmarks, GCD stabilizes one-step distillation where naive Euclidean objectives collapse. One-step students provide the largest inference acceleration while retaining strong structural fidelity and substantial ensemble variability. Additional sampling steps tune the fidelity–coverage operating point and can improve higher-order ensemble observables, although distributional metrics are not uniformly monotonic with step count.

In addition, while one-step students retain substantial diversity on the PDB benchmark, reproducing MD-scale distributional statistics and higher-order observables remains more challenging. The bilevel training setup also requires simultaneously maintaining a frozen teacher, generator, and projector, increasing training memory and engineering overhead.

AI USE STATEMENT

We used generative AI tools to improve the readability and presentation of the manuscript and to assist with LaTeX formatting. All AI-assisted edits were reviewed by the authors. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

This work is based entirely on computational modeling and publicly available structural and molecular-dynamics data. It does not involve human subjects, personal or sensitive data, or human-subject experimentation. Our experiments focus on protein structure and conformational ensemble generation rather than functional protein design or wet-lab validation. As with other protein generative modeling methods, broader downstream use may raise dual-use considerations if extended to functional biomolecule design.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed algorithm, training procedure, and evaluation methodology in the main paper and appendices. The appendices additionally document the implementation and experimental settings necessary to reproduce the reported results.

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*, pp. 1–3, 2024.

Katherine I Albanese, Sophie Barbe, Shunsuke Tagami, Derek N Woolfson, and Thomas Schiex. Computational protein design. *Nature Reviews Methods Primers*, 5(1):13, 2025.

Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

David Baker. What has de novo protein design taught us about protein folding and biophysics? *Protein Science*, 28(4):678–683, 2019.

Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The Protein Data Bank. *Nucleic acids research*, 28(1): 235–242, 2000.

Nazim Bouatta, Peter Sorger, and Mohammed AlQuraishi. Protein structure prediction by AlphaFold2: are attention and symmetries all you need? *Biological Crystallography*, 77(8):982–991, 2021.

Sandeep Chakraborty, Ravindra Venkatramani, Basuthkar J Rao, Bjarni Asgeirsson, and Abhaya M Dandekar. Protein structure quality assessment based on the distance profiles of consecutive backbone cα atoms. *F1000Research*, 2:211, 2013.

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.

Justas Dauparas, Gyu Rie Lee, Robert Pecoraro, Linna An, Ivan Anishchenko, Cameron Glasscock, and David Baker. Atomic context-conditioned protein sequence design using LigandMPNN. *Nature Methods*, pp. 1–7, 2025.

Tomas Geffner, Kieran Didi, Zuobai Zhang, Danny Reidenbach, Zhonglin Cao, Jason Yim, Mario Geiger, Christian Dallago, Emine Kucukbenli, Arash Vahdat, et al. Proteina: Scaling flow-based protein structure generative models. *arXiv preprint arXiv:2503.00710*, 2025.

John B Ingraham, Max Baranov, Zak Costello, Karl W Barber, Wujie Wang, Ahmed Ismail, Vincent Frappier, Dana M Lord, Christopher Ng-Thow-Hing, Erik R Van Vlack, et al. Illuminating protein space with a programmable generative model. *Nature*, 623(7989):1070–1078, 2023.

Bowen Jing, Ezra Erives, Peter Pao-Huang, Gabriele Corso, Bonnie Berger, and Tommi Jaakkola. Eigenfold: Generative protein structure prediction with diffusion models. *arXiv preprint arXiv:2304.02198*, 2023.

Bowen Jing, Bonnie Berger, and Tommi Jaakkola. AlphaFold meets flow matching for generating protein ensembles. *arXiv preprint arXiv:2402.04845*, 2024.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *nature*, 596(7873):583–589, 2021.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Weijian Luo, Tianyang Hu, Shifeng Zhang, Jiacheng Sun, Zhenguo Li, and Zhihua Zhang. Diff-instruct: A universal approach for transferring knowledge from pre-trained diffusion models. *Advances in Neural Information Processing Systems*, 36:76525–76546, 2023.

Xingjie Pan and Tanja Kortemme. Recent advances in de novo protein design: Principles, methods, and applications. *Journal of Biological Chemistry*, 296, 2021.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. 2023.

Yann Vander Meersche, Gabriel Cretin, Aria Gheeraert, Jean-Christophe Gelly, and Tatiana Galochkina. ATLAS: protein flexibility description from atomistic molecular dynamics simulations. *Nucleic acids research*, 52(D1):D384–D392, 2024.

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.

wwPDB consortium. Protein Data Bank: the single global archive for 3d macromolecular structure data. *Nucleic acids research*, 47(D1):D520–D528, 2019.

Liyang Xie, Haoran Zhang, Zhendong Wang, Wesley Tansey, and Mingyuan Zhou. Distilled protein backbone generation. *arXiv preprint arXiv:2510.03095*, 2025.

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.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6613–6623, 2024.

Mingyuan Zhou, Huangjie Zheng, Zhendong Wang, Mingzhang Yin, and Hai Huang. Score identity distillation: Exponentially fast distillation of pretrained diffusion models for one-step generation. In *Forty-first International Conference on Machine Learning*, 2024.

A PROOF

Setup. Let $x = (x_1, \dots, x_N) \in \mathbb{R}^{N\times 3}$ denote N residue coordinates (vectorized as $x \in \mathbb{R}^{3N}$), and let $q(x_1)$ be the data distribution. For $t \in [0, 1)$ we consider the linear interpolation

$$x_t \;=\; (1-t)x_0 + tx_1, \qquad x_1 \sim q(\cdot), \quad x_0 \sim p_0(\cdot). \tag{11}$$

This induces the conditional path density $p_t(\cdot \mid x_1)$ and the marginal path density

$$p_t(x) \;=\; \int p_t(x \mid x_1)\, q(x_1)\, dx_1. \tag{12}$$

A.1 REGULARIZED HARMONIC POLYMER BASE

We start from the harmonic polymer energy on a chain:

$$E(x) \;:=\; \frac{\lambda}{2} \sum_{i=1}^{N-1} \|x_i - x_{i+1}\|^2, \qquad \lambda > 0. \tag{13}$$

Let $L \in \mathbb{R}^{N\times N}$ be the chain graph Laplacian and define the Laplacian-structured precision

$$\Lambda \;:=\; \lambda\,(L \otimes I_3) \in \mathbb{R}^{3N\times 3N}. \tag{14}$$

where I_3 is the 3×3 identity matrix, $\otimes$ the Kronecker product, and $L \in \mathbb{R}^{N\times N}$ is the chain-graph (path) Laplacian matrix based on the protein sequence:

$$L_{ij} = \begin{cases} 1, & i = j \in \{1, N\}, \\ 2, & i = j,\ 1 < i < N, \\ -1, & |i-j| = 1, \\ 0, & |i-j| > 1. \end{cases} \tag{15}$$

and $\lambda > 0$ controls the stiffness/precision scale of the harmonic polymer prior, inducing correlated covariance Λ^{-1} across residues and spatial dimensions. Using the standard Laplacian quadratic-form identity,

$$\sum_{i=1}^{N-1} \|x_i - x_{i+1}\|^2 \;=\; x^\top (L \otimes I_3)\, x, \tag{16}$$

we can rewrite (13) as $E(x) = \frac{1}{2} x^\top \Lambda x$.

Epsilon regularization. Since $\Lambda \succeq 0$ has a nontrivial nullspace (global translations), we regularize it by adding a small isotropic term:

$$\Lambda_\varepsilon \;:=\; \Lambda + \varepsilon I_{3N}, \qquad \varepsilon > 0. \tag{17}$$

Then $\Lambda_\varepsilon \succ 0$ is full-rank, hence invertible. We define the (proper) Gaussian base

$$p_{0,\varepsilon}(x) \;:=\; Z_\varepsilon^{-1} \exp\!\left(-\frac{1}{2} x^\top \Lambda_\varepsilon x\right) \;=\; \mathcal{N}\big(0,\ \Lambda_\varepsilon^{-1}\big)\,. \tag{18}$$

A.1.1 PRACTICAL MOORE–PENROSE IMPLEMENTATION

Let $J := \lambda L \in \mathbb{R}^{N\times N}$, so that

$$\Lambda = J \otimes I_3.$$

Let

$$J = P\,\mathrm{diag}(0, \nu_2, \dots, \nu_N)\, P^\top, \qquad \nu_i > 0,$$

and define the centered subspace

$$\mathcal{H} := \mathrm{Null}(\Lambda)^\perp, \qquad \Pi_{\mathcal{H}} = \left(I_N - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\right) \otimes I_3.$$

Then Λ is positive definite on $\mathcal{H}$.

Lemma 1 (Practical pseudoinverse sampler)**.** *The Moore–Penrose pseudoinverse of Λ is the ambient-coordinate representation of the inverse of Λ on $\mathcal{H}$:*

$$\Lambda^\dagger = \Pi_{\mathcal{H}}(\Lambda|_{\mathcal{H}})^{-1}\Pi_{\mathcal{H}}.$$

Moreover,

$$\Pi_{\mathcal{H}}(\Lambda + \epsilon I_{3N})^{-1}\Pi_{\mathcal{H}} \underset{\epsilon\downarrow 0}{\longrightarrow} \Lambda^\dagger.$$

Finally, the practical sampler obtained by zeroing the translational null mode of J (equivalently, the three translational null modes of Λ) is exactly Gaussian with covariance $\Lambda^\dagger$.

Proof. Since $\Lambda = J \otimes I_3$,

$$\Lambda = (P \otimes I_3)\,\mathrm{diag}(0,0,0,\nu_2,\nu_2,\nu_2,\ldots,\nu_N,\nu_N,\nu_N)\,(P \otimes I_3)^\top.$$

Hence

$$\Pi_{\mathcal{H}} = (P \otimes I_3)\,\mathrm{diag}(0,0,0,1,\ldots,1)\,(P \otimes I_3)^\top,$$

so on $\mathcal{H}$, the operator Λ is represented by

$$\mathrm{diag}(\nu_2,\nu_2,\nu_2,\ldots,\nu_N,\nu_N,\nu_N),$$

and therefore

$$\Pi_{\mathcal{H}}(\Lambda|_{\mathcal{H}})^{-1}\Pi_{\mathcal{H}} = (P \otimes I_3)\,\mathrm{diag}(0,0,0,\nu_2^{-1},\nu_2^{-1},\nu_2^{-1},\ldots,\nu_N^{-1},\nu_N^{-1},\nu_N^{-1})\,(P \otimes I_3)^\top = \Lambda^\dagger.$$

Likewise,

$$\Pi_{\mathcal{H}}(\Lambda + \epsilon I_{3N})^{-1}\Pi_{\mathcal{H}} = (P \otimes I_3)\,\mathrm{diag}(0,0,0,(\nu_2+\epsilon)^{-1},(\nu_2+\epsilon)^{-1},(\nu_2+\epsilon)^{-1},\ldots)\,(P \otimes I_3)^\top,$$

which converges entrywise to $\Lambda^\dagger$ as $\epsilon \downarrow 0$.

For the implementation, let

$$D^\dagger := \mathrm{diag}(0,\nu_2^{-1},\ldots,\nu_N^{-1}), \qquad X = P(D^\dagger)^{1/2}Z, \qquad Z_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1),$$

with $X, Z \in \mathbb{R}^{N\times 3}$. Then each Cartesian axis satisfies

$$\mathrm{Cov}(X_{:,k}) = PD^\dagger P^\top = J^\dagger,$$

hence

$$\mathrm{Cov}(\mathrm{vec}(X)) = J^\dagger \otimes I_3 = \Lambda^\dagger.$$

Thus the practical sampler is exactly the centered Gaussian with covariance $\Lambda^\dagger$. □

We therefore interpret the practical harmonic base as

$$p_0^{\mathcal{H}} = \mathcal{N}_{\mathcal{H}}\big(0,(\Lambda|_{\mathcal{H}})^{-1}\big),$$

whose covariance is represented in ambient coordinates by $\Lambda^\dagger$.

A.2 Main identity

Proposition 1 (Marginal score identity under a regularized harmonic prior)**.** *Under* (11)–(12) *and the regularized harmonic prior* (18)*, the marginal score admits the conditional-mean form*

$$\nabla_x \log p_t(x) \;=\; -\frac{1}{(1-t)^2}\,\Lambda_\varepsilon\Big(x - t\,\mathbb{E}[x_1 \mid x_t = x]\Big). \tag{19}$$

Equivalently, defining

$$u_t(x) \;:=\; \frac{\mathbb{E}[x_1 \mid x_t = x] - x}{1-t}, \tag{20}$$

we have the velocity form

$$\nabla_x \log p_t(x) \;=\; \frac{1}{1-t}\,\Lambda_\varepsilon\big(tu_t(x) - x\big). \tag{21}$$

Proof. We begin with (12) and differentiate:

$$\begin{aligned}\nabla_x \log p_t(x) &= \frac{1}{p_t(x)}\,\nabla_x p_t(x) = \frac{1}{p_t(x)}\,\nabla_x \int p_t(x \mid x_1)\, q(x_1)\, dx_1 \\ &= \frac{1}{p_t(x)} \int \nabla_x p_t(x \mid x_1)\, q(x_1)\, dx_1 = \frac{1}{p_t(x)} \int p_t(x \mid x_1)\, q(x_1)\, \nabla_x \log p_t(x \mid x_1)\, dx_1 \\ &= \int \underbrace{\frac{p_t(x \mid x_1)\, q(x_1)}{p_t(x)}}_{= q(x_1 \mid x_t = x)} \nabla_x \log p_t(x \mid x_1)\, dx_1 = \int q(x_1 \mid x_t = x)\, \nabla_x \log p_t(x \mid x_1)\, dx_1 \qquad (22) \\ &= \mathbb{E}_{x_1 \sim q(\cdot \mid x_t = x)}[\nabla_x \log p_t(x \mid x_1)]\,. \qquad (23)\end{aligned}$$

It remains to compute $\nabla_x \log p_t(x \mid x_1)$ under the regularized base. Fix x_1. By (11),

$$x_t \;=\; t x_1 + (1-t) x_0. \qquad (24)$$

Since $x_0 \sim \mathcal{N}(0, \Lambda_\varepsilon^{-1})$ by (18), the affine map (24) implies the conditional path density is Gaussian:

$$p_t(x \mid x_1) \;=\; \mathcal{N}\big(t x_1,\; (1-t)^2 \Lambda_\varepsilon^{-1}\big)\,. \qquad (25)$$

Therefore, using the Gaussian score formula $\nabla_x \log \mathcal{N}(x; m, C) = -C^{-1}(x - m)$, we obtain

$$\nabla_x \log p_t(x \mid x_1) \;=\; -\big((1-t)^2 \Lambda_\varepsilon^{-1}\big)^{-1}(x - t x_1) \;=\; -\frac{1}{(1-t)^2}\Lambda_\varepsilon\,(x - t x_1). \qquad (26)$$

Substituting (26) into (23) yields

$$\nabla_x \log p_t(x) = \mathbb{E}\left[-\frac{1}{(1-t)^2}\Lambda_\varepsilon\,(x - t x_1) \;\middle|\; x_t = x\right] = -\frac{1}{(1-t)^2}\Lambda_\varepsilon\Big(x - t\,\mathbb{E}[x_1 \mid x_t = x]\Big),$$

which proves (19). Finally, from (20) we have $\mathbb{E}[x_1 \mid x_t = x] = x + (1-t)u_t(x)$, and thus

$$x - t\mathbb{E}[x_1 \mid x_t = x] = x - t\big(x + (1-t)u_t(x)\big) = (1-t)\big(x - t u_t(x)\big).$$

Plugging this into (19) gives (21). □

Special case (isotropic Gaussian base). If $p_0(x) = \mathcal{N}(0, I)$, then $\Lambda_\varepsilon = I$ (up to scaling) and Proposition 1 reduces to

$$\nabla_x \log p_t(x) = -\frac{1}{(1-t)^2}\,(x - t\,\mathbb{E}[x_1 \mid x_t = x]) = \frac{t u_t(x) - x}{1-t}.$$

A.3 Generator loss derivation under the regularized harmonic prior

We derive the SiD loss based on the regularized harmonic prior following (Zhou et al., 2024).

Let $x_g \sim p_\theta(\cdot)$ and define the noising kernel

$$x_t \mid x_g \sim q_t(\cdot \mid x_g) := \mathcal{N}\big(t x_g,\; (1-t)^2 \Lambda_\varepsilon^{-1}\big)\,. \qquad (27)$$

Then

$$\nabla_{x_t} \log q_t(x_t \mid x_g) = -\frac{1}{(1-t)^2}\Lambda_\varepsilon(x_t - t x_g) = \frac{1}{(1-t)^2}\Lambda_\varepsilon(t x_g - x_t), \qquad (28)$$

and Proposition 1 gives

$$\nabla_{x_t} \log p_t(x_t) = -\frac{1}{(1-t)^2}\Lambda_\varepsilon\Big(x_t - t\,\mathbb{E}[x_1 \mid x_t]\Big) = \frac{1}{(1-t)^2}\Lambda_\varepsilon\Big(t\,\mathbb{E}[x_1 \mid x_t] - x_t\Big). \qquad (29)$$

Let $S_\phi(\cdot, t)$ and $S_\psi(\cdot, t)$ be teacher and auxiliary score fields, and consider

$$\mathcal{J}_\phi(t) := \mathbb{E}_{x_t \sim p_t}\Big[\|S_\phi(x_t, t) - \nabla_{x_t} \log p_t(x_t)\|_2^2\Big]\,. \qquad (30)$$

We use the bilinear surrogate

$$\mathcal{J}_\phi(t) \;\approx\; \mathbb{E}_{x_t\sim p_t}\Big[(S_\phi(x_t,t)-S_\psi(x_t,t))^\top\Big(S_\phi(x_t,t)-\nabla_{x_t}\log p_t(x_t)\Big)\Big]. \tag{31}$$

Expanding and substituting (29) yields

$$\begin{aligned}&\mathbb{E}_{x_t\sim p_t}\Big[(S_\phi-S_\psi)^\top\Big(S_\phi-\nabla\log p_t\Big)\Big]\\&=\mathbb{E}_{x_t\sim p_t}\big[(S_\phi-S_\psi)^\top S_\phi\big]-\mathbb{E}_{x_t\sim p_t}\left[(S_\phi-S_\psi)^\top\frac{1}{(1-t)^2}\Lambda_\varepsilon\Big(t\,\mathbb{E}[x_1\mid x_t]-x_t\Big)\right].\end{aligned} \tag{32}$$

Under (27) we also have

$$\begin{aligned}&\mathbb{E}_{x_g\sim p_\theta}\mathbb{E}_{x_t\sim q_t(\cdot|x_g)}\big[(S_\phi-S_\psi)^\top\nabla_{x_t}\log q_t(x_t\mid x_g)\big]\\&\qquad=\mathbb{E}_{x_g}\mathbb{E}_{x_t|x_g}\left[(S_\phi-S_\psi)^\top\frac{1}{(1-t)^2}\Lambda_\varepsilon\big(tx_g-x_t\big)\right].\end{aligned} \tag{33}$$

Combining the Λ_ε-terms gives

$$\begin{aligned}&\mathbb{E}_{x_t\sim p_t}\big[(S_\phi-S_\psi)^\top\nabla_{x_t}\log p_t(x_t)\big]-\mathbb{E}_{x_g\sim p_\theta}\mathbb{E}_{x_t\sim q_t(\cdot|x_g)}\big[(S_\phi-S_\psi)^\top\nabla_{x_t}\log q_t(x_t\mid x_g)\big]\\&=\frac{t}{(1-t)^2}\,\mathbb{E}_{x_g\sim p_\theta}\mathbb{E}_{x_t\sim q_t(\cdot|x_g)}\Big[(S_\phi(x_t,t)-S_\psi(x_t,t))^\top\Lambda_\varepsilon\Big(\mathbb{E}[x_1\mid x_t]-x_g\Big)\Big].\end{aligned} \tag{34}$$

Replacing $\mathbb{E}[x_1\mid x_t]\approx f_\phi(x_t,t)$, define

$$\mathcal{L}_\theta^{(2)}(t):=\frac{t}{(1-t)^2}\,\mathbb{E}_{x_g\sim p_\theta}\mathbb{E}_{x_t\sim q_t(\cdot|x_g)}\Big[(S_\phi(x_t,t)-S_\psi(x_t,t))^\top\Lambda_\varepsilon\Big(f_\phi(x_t,t)-x_g\Big)\Big], \tag{35}$$

and

$$\mathcal{L}_\theta^{(1)}(t):=\mathbb{E}_{x_g\sim p_\theta}\mathbb{E}_{x_t\sim q_t(\cdot|x_g)}\big[\|S_\phi(x_t,t)-S_\psi(x_t,t)\|_2^2\big]. \tag{36}$$

The score-based SiD loss is

$$\mathcal{L}_\theta(t):=\mathcal{L}_\theta^{(2)}(t)-\alpha\,\mathcal{L}_\theta^{(1)}(t). \tag{37}$$

Define the score proxy induced by a denoiser $f(\cdot,t)$:

$$S_f(x_t,t):=-\frac{1}{(1-t)^2}\Lambda_\varepsilon\big(x_t-tf(x_t,t)\big), \tag{38}$$

so that

$$S_{f_\phi}(x_t,t)-S_{f_\psi}(x_t,t)=\frac{t}{(1-t)^2}\Lambda_\varepsilon\big(f_\phi(x_t,t)-f_\psi(x_t,t)\big). \tag{39}$$

Substituting (39) into (35)–(36) yields the explicit per-sample loss:

$$\boxed{\begin{aligned}\mathcal{L}_\theta(x_t,t,x_g)=\omega(t)\,\frac{t^2}{(1-t)^4}\Bigg[&(1-\alpha)\,\Big\|\Lambda_\varepsilon\big(f_\phi(x_t,t)-f_\psi(x_t,t)\big)\Big\|_2^2\\&+\big(f_\phi(x_t,t)-f_\psi(x_t,t)\big)^\top\Lambda_\varepsilon^2\big(f_\psi(x_t,t)-x_g\big)\Bigg].\end{aligned}} \tag{40}$$

B Algorithm and Experiment details

Generator-loss weighting. For the generator loss in (10), we use the time-dependent weight

$$\omega(t) = \frac{(1-t)^4}{t^2 \|x_g - x_\phi(x_t, t)\|_{1,\mathrm{sq}}}, \tag{41}$$

The term $\|\cdot\|_{1,\mathrm{sq}}$ denotes the squared L_1-style normalization used in the implementation.

Harmonic-prior stiffness. For the harmonic polymer prior, we set $\lambda = \frac{3}{3.8^2}$. This follows the harmonic diffusion formulation (Jing et al., 2023). The 3.8 value is consistent with protein-structure quality assessment based on distances between consecutive backbone C_α atoms (Chakraborty et al., 2013).

Other hyperparameters. Unless otherwise specified, we adopt the default hyperparameter settings of each distillation algorithm and follow the AlphaFlow evaluation protocol for sampling, metrics, and data processing. For diffusion-based distillation methods, We adapt their diffusion-time parameterization to the continuous flow-time setting of AlphaFlow while otherwise retaining the original hyperparameter configuration.

Algorithm 1 Bilevel geometry-consistent distillation for flow matching

Require: Training data $\{(A, M)\}$; teacher f_ϕ (frozen); generator G_θ; projector f_ψ; coefficient α

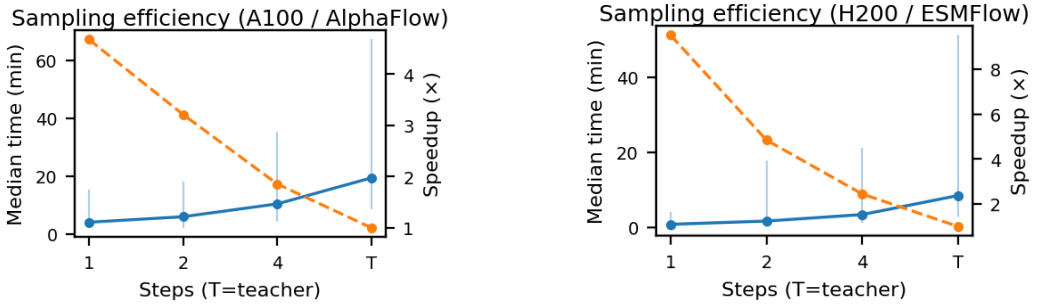
Initialize $\theta \leftarrow \phi,\ \psi \leftarrow \phi$
repeat
  Sample (A, M)
  Sample backbone $S_g \leftarrow G_\theta(A, M)$, $x_g = [S_g]_\beta$
  Sample $x_0 \sim q_0$ and align $\tilde{x}_0 \leftarrow \mathrm{RMSDAlign}(x_0, x_g)$
  Sample $t \sim \mathcal{U}(0,1)$ and set $x_t \leftarrow (1-t)\tilde{x}_0 + t x_g$
  $\hat{S}_\psi \leftarrow f_\psi(A, M, x_t, t)$
  Update ψ by minimizing $\mathcal{L}_\psi = \mathbb{E}_{t,x_t}\left[\left\|\tilde{S}_\psi - \mathrm{sg}(\tilde{S}_g)\right\|_2^2\right]$
  $\hat{S}_\phi \leftarrow f_\phi(A, M, x_t, t)$
  Update θ by minimizing $\mathcal{L}_\theta$ in (10)
until convergence
return G_θ

C Additional results

C.1 PDB ensemble regime

Table 3: PDB-ensemble distillation results. Mean results are reported.

Method	Precision	Recall	Diversity
Euclidean SiD (1 step)	0.78	0.69	0.00
GCD-SiD (1 step)	74.73	74.16	26.98
GCD-SiD (2 step)	76.28	75.66	25.51
GCD-SiD (4 step)	74.52	74.78	29.75
GCD-DMD (1 step)	77.17	73.31	11.32
GCD-SC (1 step)	77.39	73.87	11.60
AlphaFlow (baseline)	77.23	76.03	25.40
GCD-SiD (1 step)	70.10	69.53	29.07
GCD-SiD (2 step)	72.84	72.57	27.64
GCD-SiD (4 step)	71.64	71.93	31.17
ESMFlow	72.53	72.49	28.05



(a) AlphaFlow family (A100).

(b) ESMFlow family (H200).

Figure 3: **Sampling efficiency.** Per-protein sampling time on the PDB test set. Solid blue lines show median sampling time (left y-axis) with min–max ranges; dashed orange lines show median-based speedup relative to the teacher (right y-axis).

We report median metrics in the main text, following AlphaFlow, to summarize performance on a typical target and reduce sensitivity to target-level outliers. Table 3 provides the corresponding means. Although the exact method ranking differs between mean and median aggregation, both views show the same broad step-wise pattern: two-step sampling tends to favor fidelity, whereas four-step sampling yields the broadest ensembles. The larger mean than median Diversity values in several settings further suggest substantial target-level heterogeneity and positive skew in conformational variability.

C.2 MD ENSEMBLE REGIME

Figure 5 shows the complementary CDF (1−CDF) of the unsigned cosine similarity between the first principal component directions (PC1) of the predicted ensemble and the corresponding MD ensemble, computed per target (higher is better). Figure 6 further provides qualitative visualizations on representative targets.

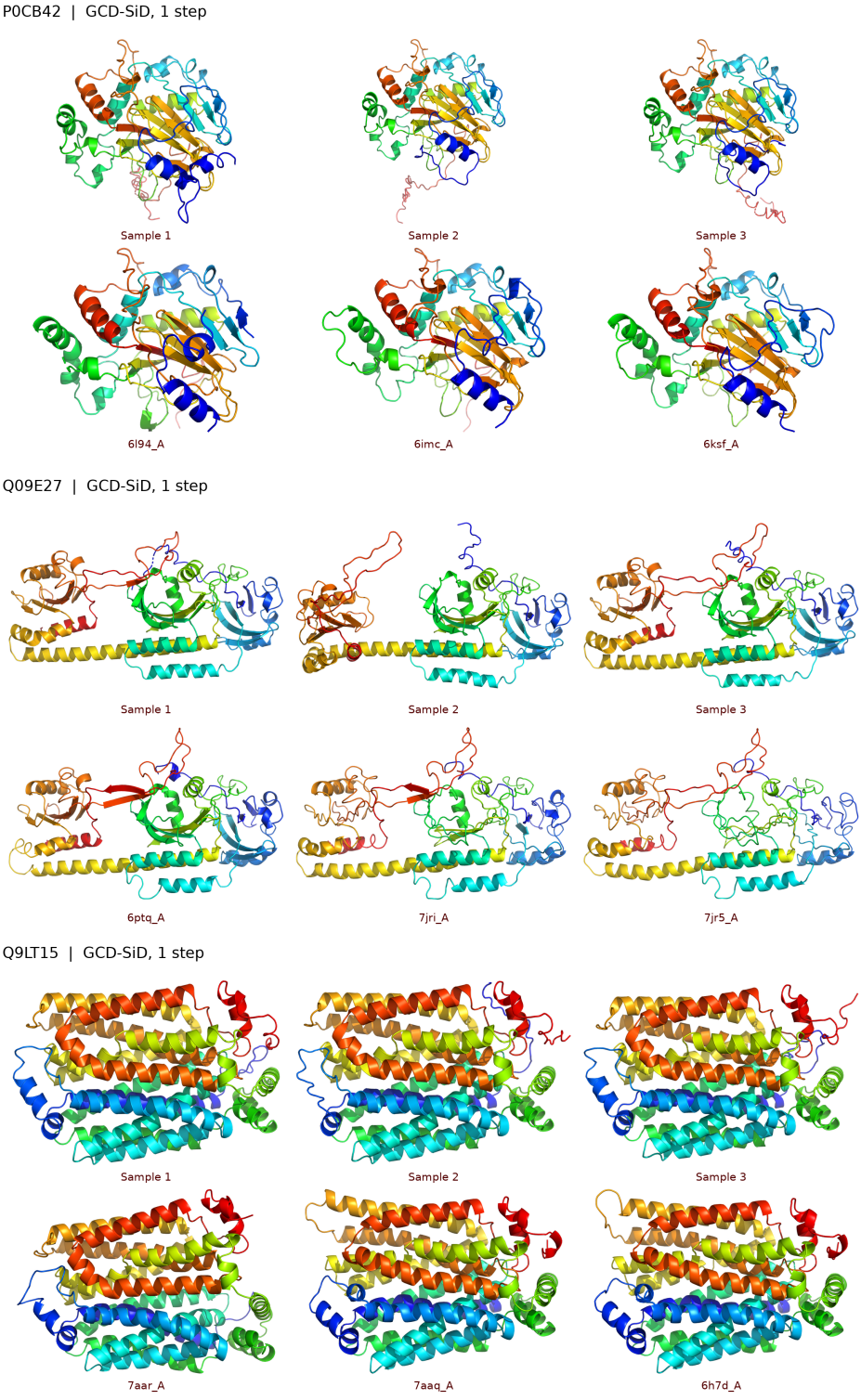


Figure 4: **Distilled conformational sampling example.** Three samples from GCD-SiD (top) and three PDB conformers (bottom) are visualized.

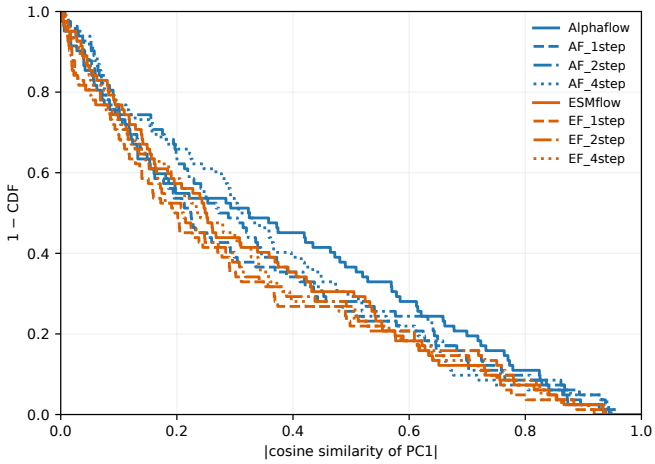


Figure 5: **1−CDF of PCA subspace similarity.** Complementary CDF of the unsigned cosine similarity between the top principal components of the predicted ensemble and the corresponding MD ensemble.

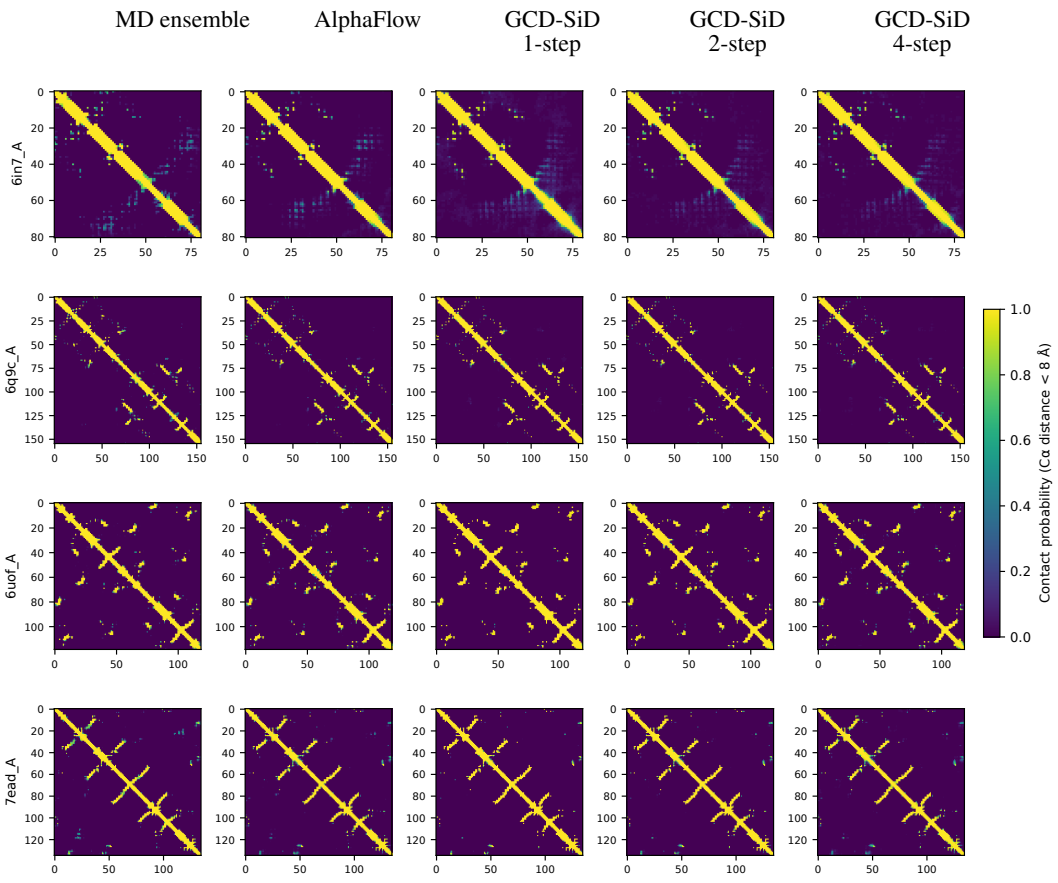


Figure 6: **Residue–residue contact probability maps for representative MD targets.** Columns show the reference MD ensemble, the AlphaFlow-MD teacher, and GCD-SiD with one-step, two-step, and four-step sampling. Rows correspond to representative protein targets. Brighter intensities indicate higher contact probability.

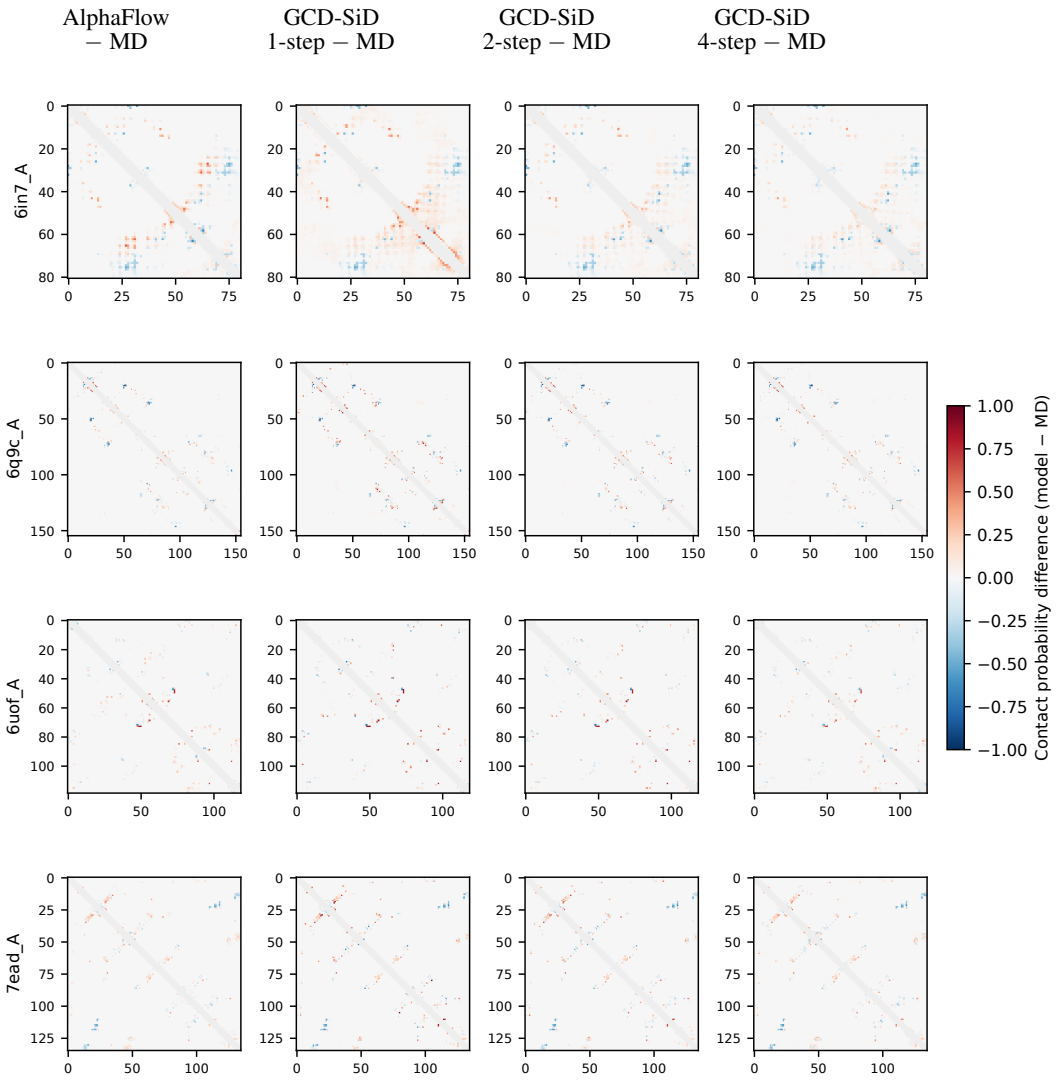


Figure 7: **Contact probability differences relative to MD.** Shown are $P_{\text{model}} - P_{\text{MD}}$ for AlphaFlow-MD and GCD-SiD (1/2/4-step); red/blue denote over-/under-estimation.