, 2026, pp. 1–4

CoCoFold2: scalable latent refinement of diffusion-based protein structure predictions from limited-particle cryo-EM data

Junwen Liao,[1] Mingxu Hu[2,*] and Chenglong Bao[3,4,5,*]

[1]Qiuzhen College, Tsinghua University, 100084, Beijing, China
[2]Institute of Bio-Architecture and Bio-Interactions, Shenzhen Medical Academy of Research and Translation, 518100, Shenzhen, China
[3]Yau Mathematical Sciences Center, Tsinghua University, 100084, Beijing, China
[4]Yanqi Lake Beijing Institute of Mathematical Sciences and Applications, 101408, Beijing, China
[5]State Key Laboratory of Membrane Biology, School of Life Sciences, Tsinghua University, 100084, Beijing, China

*Corresponding authors. humingxu@smart.org.cn and clbao@mail.tsinghua.edu.cn

Abstract

This document provides supplementary material for CoCoFold2.

Supplementary Note S1: Conditional effect of fixing diffusion stochasticity

We provide a compact variance-based interpretation of the fixed-stochasticity design. This analysis is an illustrative plain-SGD approximation, rather than a convergence result for the AdamW procedure used in CoCoFold2. Let $\vartheta = (\Delta z, w, \sigma)$ collect the trainable latent perturbation and renderer parameters, let ξ denote the diffusion sampling state, and let $\mathcal{B}$ denote a sampled particle minibatch. In the tested resampled-stochasticity control, variation in ξ arises from the initial diffusion noise and random reorientation; the per-step injected noise has zero amplitude because $\gamma_0 = \gamma_{\min} = 0$. The minibatch loss can be written as

$$\ell(\vartheta; \xi, \mathcal{B}) = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \mathcal{L}_i(D_{\theta_0}(z + \Delta z; \xi), w, \sigma) + \mathcal{R}(w, \sigma), \tag{S1}$$

where D_{θ_0} is the frozen Protenix-v1 diffusion decoder, $\mathcal{L}_i$ includes fixed-frame projection and contrast-transfer-function comparison for particle i, and $\mathcal{R}$ denotes the renderer penalties.

Resampling diffusion stochasticity corresponds to optimizing the expected objective $\overline{J}(\vartheta)$, whereas fixing one realization ξ_0 defines a conditional objective $J_{\xi_0}(\vartheta)$:

$$\overline{J}(\vartheta) = \mathbb{E}_{\xi,\mathcal{B}}[\ell(\vartheta; \xi, \mathcal{B})], \qquad J_{\xi_0}(\vartheta) = \mathbb{E}_{\mathcal{B}}[\ell(\vartheta; \xi_0, \mathcal{B})]. \tag{S2}$$

Thus, the fixed and resampled procedures do not optimize the same objective. In particular, CoCoFold2 does not claim that one fixed realization is a more accurate approximation to the expectation over all diffusion samples.

For the resampled procedure, the stochastic gradient is

$$g(\vartheta; \xi, \mathcal{B}) = \nabla_\vartheta \ell(\vartheta; \xi, \mathcal{B}). \tag{S3}$$

Conditioned on ϑ, the law of total covariance separates variation caused by diffusion resampling from variation caused by particle minibatching:

$$\mathrm{Cov}_{\xi,\mathcal{B}}(g) = \mathbb{E}_{\mathcal{B}}[\mathrm{Cov}_{\xi}(g \mid \mathcal{B})] + \mathrm{Cov}_{\mathcal{B}}(\mathbb{E}_{\xi}[g \mid \mathcal{B}])\,. \tag{S4}$$

The first term is the diffusion-induced contribution for a fixed particle batch. Under fixed stochasticity, $\xi_t = \xi_0$ at every sampler call, and therefore

$$\mathrm{Cov}_{\xi_t}[\nabla_\vartheta \ell(\vartheta; \xi_t, \mathcal{B}) \mid \vartheta, \mathcal{B}] = 0. \tag{S5}$$

Particle batches are still shuffled, so $\mathrm{Cov}_{\mathcal{B}}[\nabla_\vartheta \ell(\vartheta; \xi_0, \mathcal{B})]$ is generally nonzero. Fixing ξ_0 therefore removes diffusion-resampling variability from successive updates; it does not make the complete optimization trajectory free of stochastic gradient noise.

To relate this distinction to a standard descent bound, let J denote either $\overline{J}$ or J_{ξ_0}, assume that J is L-smooth, and assume that the corresponding stochastic-gradient estimator is unbiased with conditional covariance Σ_t. For the plain-SGD update $\vartheta_{t+1} = \vartheta_t - \eta g_t$,

$$\mathbb{E}[J(\vartheta_{t+1}) \mid \vartheta_t] \le J(\vartheta_t) - \eta\left(1 - \frac{L\eta}{2}\right) \|\nabla J(\vartheta_t)\|^2 + \frac{L\eta^2}{2} \mathrm{Tr}(\Sigma_t). \tag{S6}$$

For $0 < \eta < 2/L$, this bound guarantees an expected decrease when the descent term exceeds the covariance term. Resampling ξ_t introduces a diffusion-dependent contribution to Σ_t; fixing ξ_0 removes this contribution across optimization steps, although particle-minibatch variance remains.

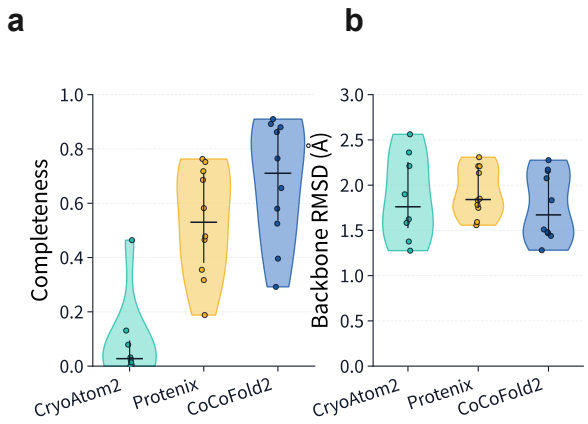

Figure S1 Map-derived latent refinement on ten low-resolution maps. a, Completeness for CryoAtom2, Protenix-v1 and CoCoFold2. **b,** Backbone RMSD for the same targets and methods. Each point denotes one target. Protenix-v1 and CoCoFold2 have $n = 10$; CryoAtom2 has $n = 8$ targets with available values. Values are reported in Supplementary Table S6.

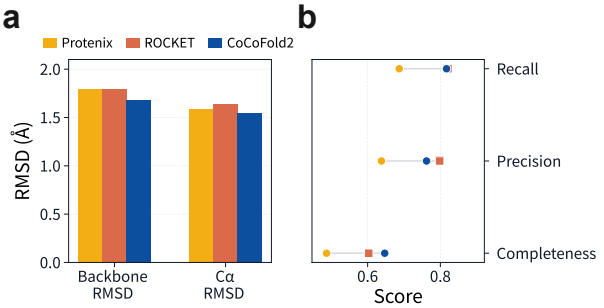

Figure S2 Single-target cryo-ET-like scope test. a, Backbone and Cα RMSD for Protenix-v1, ROCKET and CoCoFold2 on GroEL chain A. **b,** Recall, precision and completeness for the same outputs. The input was a 9.6 Å low-resolution, map-derived observation. Values are reported in Supplementary Table S7.

The analysis assumes smooth local behavior, unbiased plain-SGD gradients and a fixed coordinate-frame branch. The actual implementation uses AdamW, jointly updates renderer parameters, shuffles particles and includes a discrete flip safeguard, so the bound is mechanistic rather than predictive. Empirically, fixed-stochasticity latent refinement gave better backbone RMSD and completeness endpoints than resampling in the three selected stress-test targets (Supplementary Table S4). These observations support fixed stochasticity for the single-structure conditional-refinement setting tested here, without implying that it is preferable for posterior or conformational-ensemble objectives.

Supplementary Note S2: Exploratory observation operators and scope tests

Map-derived projection supervision

For an experimental density map M, synthetic projection observations are formed at specified viewing geometries:

$$Y_i^{\mathrm{map}} = C_i \circ \Pi_{R_i,t_i}[M], \tag{S7}$$

where Π_{R_i,t_i} denotes projection at rotation R_i and translation t_i, C_i denotes the selected contrast-transfer operator and $\circ$ denotes Fourier-domain multiplication. The structure generated from $z + \Delta z$ is rendered at the corresponding geometries and compared with Y_i^{map} using the particle-space loss. The decoder remains frozen, and the update is applied only through the target-specific latent perturbation and renderer parameters.

Across ten low-resolution map targets, mean completeness increased from 0.53 for Protenix-v1 to 0.68 for CoCoFold2, mean F1 score increased from 0.61 to 0.71, and mean backbone RMSD decreased from 1.92 to 1.77 Å (Supplementary Table S6; Supplementary Fig. S1). CryoAtom2 has values available for eight of the ten targets. This experiment shows that map-derived projections can supply a correction signal in the evaluated maps; it does not establish equivalence to raw-particle supervision.

Cryo-ET-like low-resolution density example

The cryo-ET-like experiment used the 9.6 Å GroEL density map (PDB ID 8P4P) and evaluated chain A. The density was converted to map-derived projections using Eq. (S7), and the same frozen-prior latent-refinement formulation was applied. CoCoFold2 obtained a backbone RMSD of 1.68 Å, a Cα RMSD of 1.55 Å, and a completeness of 0.65, compared with 1.80 Å, 1.59 Å and 0.49, respectively, for Protenix-v1 (Supplementary Table S7; Supplementary Fig. S2). ROCKET obtained a backbone RMSD of 1.79 Å, a Cα RMSD of 1.64 Å, and a completeness of 0.60.

X-ray diffraction proof of concept

For Miller index h, the differentiable calculated structure factor is represented as

$$F_{\mathrm{calc}}(h) = \sum_{j=1}^{N} o_j f_j(s_h) \exp\left(-\frac{B_j s_h^2}{4}\right) \exp(2\pi \mathrm{i}\, h \cdot r_j)\,, \tag{S8}$$

where o_j, f_j, B_j and r_j denote the occupancy, atomic scattering factor, isotropic displacement parameter and atomic coordinate, respectively, in the crystallographic frame. The calculated and observed diffraction terms are compared through the implemented crystallographic loss, a negative log-likelihood-gain objective,

$$\mathcal{L}_{\mathrm{Xray}} = -\,\mathrm{LLG}(F_{\mathrm{calc}}, F_{\mathrm{obs}}; \psi)\,, \tag{S9}$$

where ψ collects scaling, uncertainty, solvent and error-model nuisance terms.

The single experiment on the extracellular domain of hepatocyte growth factor activator inhibitor 1 (HAI-1; PDB ID 5H7V) yielded a backbone RMSD of 1.08 Å, a Cα RMSD of 0.98 Å, and a completeness of 0.85 for CoCoFold2, compared with 1.38 Å, 1.28 Å and 0.79, respectively, for Protenix-v1 (Supplementary Table S8; Supplementary Fig. S3). ROCKET yielded a backbone RMSD of 1.07 Å, a Cα RMSD of 0.94 Å, and a completeness of 0.83. Thus, latent refinement improved the starting prediction in this case but did not produce the lowest RMSD.

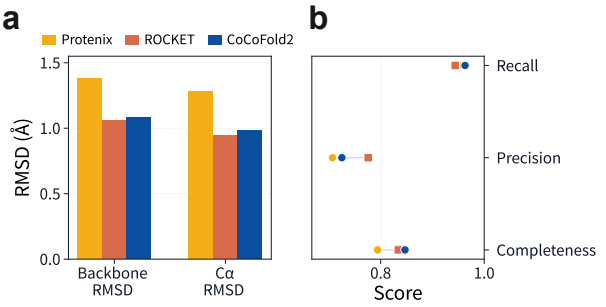

Figure S3 Single-target low-resolution X-ray scope test. a, Backbone and Cα RMSD for Protenix-v1, ROCKET and CoCoFold2 on the HAI-1 extracellular domain. **b,** Recall, precision and completeness for the same outputs. Values are reported in Supplementary Table S8.

Supplementary Note S3: Post hoc latent diagnostics

Scaling a learned target-specific perturbation.

With the decoder, cached representation and diffusion state fixed, a learned perturbation can be scaled after training as

$$x(\alpha) = D_\theta(z + \alpha\Delta z; \xi_0). \tag{S10}$$

Here $\alpha = 0$ reproduces the initial conditional prediction, $\alpha = 1$ applies the fitted CoCoFold2 perturbation, negative values probe the reverse direction, and $\alpha > 1$ extrapolates beyond the learned magnitude. The coordinate-frame transformation, renderer policy and sampling state must remain fixed across α. Deposited reference coordinates are used only to evaluate and visualize the scan retrospectively; they are not used to optimize Δz or to select an operational checkpoint.

In the displayed case, moving from $\alpha = 0$ to $\alpha = 1$ increased completeness from 0.63 to 0.97 and decreased backbone RMSD from 1.47 to 0.61 Å (Supplementary Fig. S4). Both reversal and larger positive extrapolation degraded at least one metric, indicating a finite useful range for this fitted perturbation under the fixed conditional decoder. The scan is a post hoc sensitivity diagnostic, not a physical trajectory, dynamical pathway or energy landscape.

Cross-quality interpolation.

For one target, perturbations learned from approximately 1.3k and 40k particles were compared by

$$\Delta z(\alpha) = (1-\alpha)\Delta z_{\text{low}} + \alpha\Delta z_{\text{high}}, \qquad \alpha \in [0,1]. \tag{S11}$$

Along the displayed path, completeness remained between 0.99 and 1.00, and backbone RMSD changed from 0.61 to 0.54 Å toward the endpoint obtained with more particles (Supplementary Fig. S5). In this case, the two fitted perturbations were compatible under the shared conditional decoder.

Supplementary Note S4: Homologous-target latent reuse

We examined one source–target pair to test whether a learned perturbation could be mapped between homologous proteins. The source was tuna apo P-glycoprotein (PDB ID 9NFJ) and the target was human apo P-glycoprotein (PDB ID 8GMG); the reported sequence identity and similarity were 62% and 78%, respectively. The source perturbation was mapped onto the target according to a sequence alignment before decoding under the target's frozen conditional representation and fixed diffusion realization.

For tuna P-glycoprotein, target-specific CoCoFold2 refinement changed backbone RMSD from 1.44 to 0.80 Å and completeness from 0.65 to 0.96. Applying the mapped perturbation to human P-glycoprotein increased completeness from 0.59 to 0.76, whereas backbone RMSD changed from 1.42 to 1.43 Å (Supplementary Fig. S6, Supplementary Table S9).

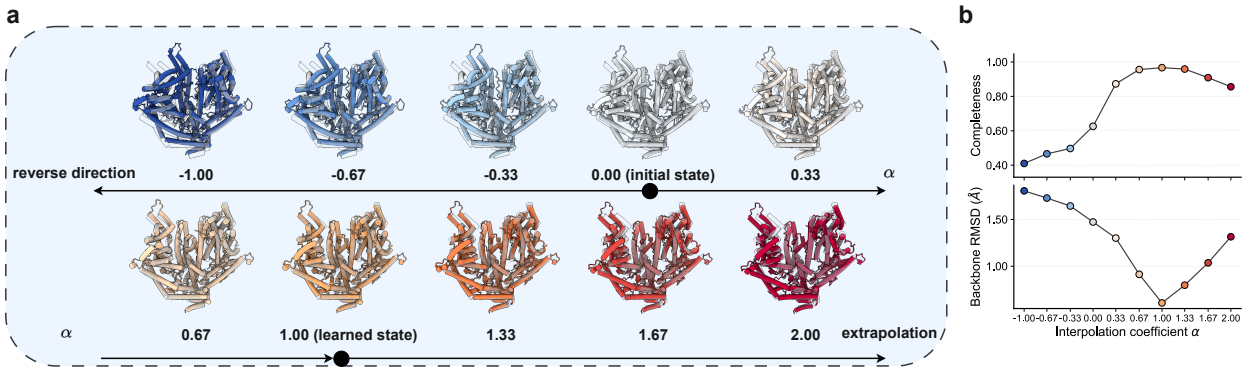

Figure S4 Post hoc scaling of one learned target-specific latent perturbation (PDB ID 6ZBH). a, Structures generated from $\alpha = -1$ to $\alpha = 2$ using Eq. (S10). **b,** Completeness and backbone RMSD across the same coefficient grid. All points share the same cached representation, decoder, diffusion state and coordinate-frame policy. The reference is used only for retrospective evaluation.

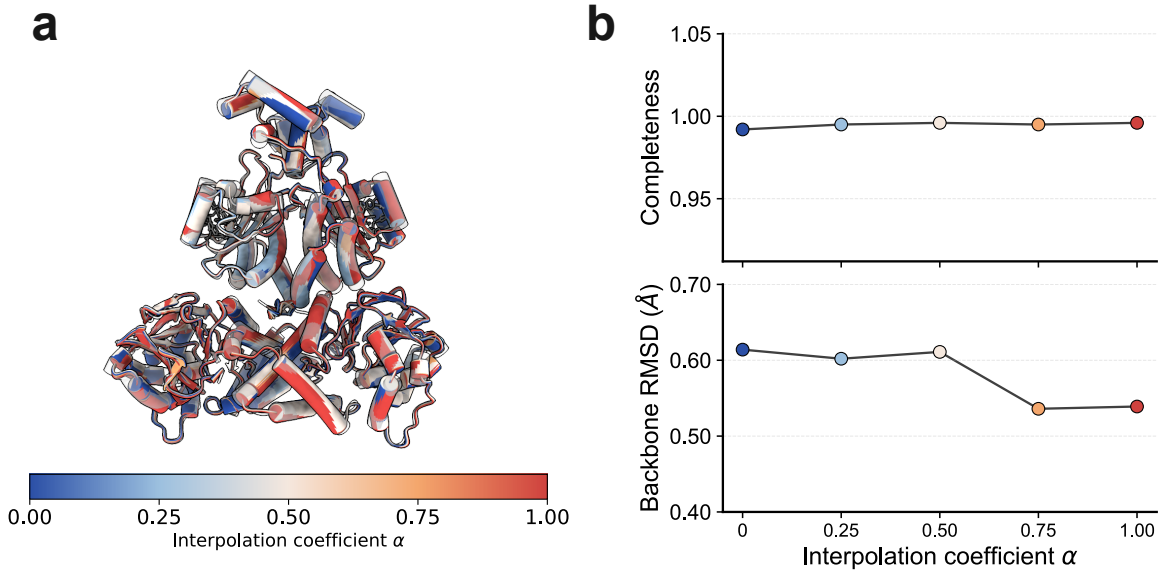

Figure S5 Interpolation between perturbations learned from two particle sets for one target (PDB ID 8DQ0). a, Overlay of structures generated using the interpolated latent perturbations defined in Eq. (S11). Structures are colored by the interpolation coefficient α, as indicated by the color bar. **b,** Completeness and backbone RMSD of the structures shown in **a**, plotted against α.

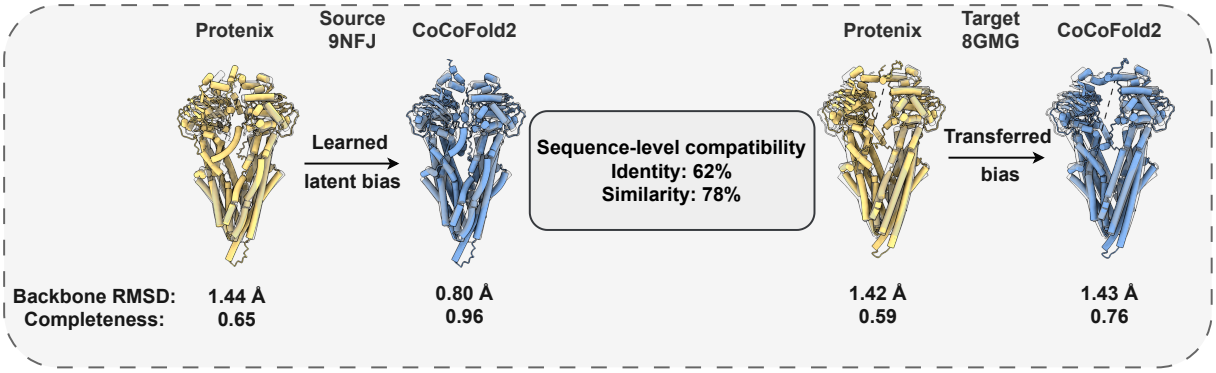

Figure S6 Single homologous-target example of latent reuse. A perturbation learned for tuna apo P-glycoprotein is mapped to human apo P-glycoprotein using the reported sequence alignment. Source and target structures are shown before and after applying their respective perturbations. Both backbone RMSD and completeness are reported because target completeness improves while target backbone RMSD changes from 1.42 to 1.43 Å.