GUIDING PROTEIN ENSEMBLE GENERATORS WITH CRYO-EM PARTICLES, WITHOUT MAPS OR POSES

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding a protein's function requires knowing its conformational distribution: what forms it takes and how often. Cryo-electron microscopy provides this distribution in principle, since every particle images a single molecule, but the averaging needed to address the experiment's low signal-to-noise ratio throws away the very heterogeneity we seek. Pretrained generative models of protein ensembles could restore part of what is lost if individual particles steered their sampling. However, this steering would require a likelihood of the observed images given a generated structure, and constructing this term requires knowledge of the particle's pose. We build this likelihood in a low-dimensional code shared by a frozen pretrained generative model for protein conformations and a frozen cryo-EM image encoder, and fit the code with simulated structure–image pairs. Because the code holds only information available in both models' representations, it suppresses imaging parameters such as defocus. The scatter that imaging, pose included, still leaves between images of the same conformation sets the likelihood's covariance. Using this likelihood, we guide generation with only 128 particles and no density map or pose estimates toward rare states and PDB structures. This is possible because the prior fills in high-resolution details, while the particles constrain large-scale rearrangements. The same code also recovers free-energy landscapes by deconvolving imaging scatter, resolving barriers that particle counting cannot. We test the approach on simulated particles of a GroEL subunit, lysenin and DppA.

1 INTRODUCTION

Proteins function by moving between conformations, including rare states that can be essential to their mechanisms (Henzler-Wildman & Kern, 2007; Sekhar & Kay, 2013). Understanding these mechanisms involves identifying the structures a protein adopts, characterising the transitions between them, and estimating their populations. At equilibrium, these populations determine a free energy landscape through their negative logarithm.

Cryo-electron microscopy records individual molecules, each frozen in one conformation and viewed at an unknown orientation (a particle). A dataset of 10^4–10^6 particle images (Toader et al., 2023) therefore samples the conformational ensemble, but extracting that ensemble is difficult. Low-dose images have signal-to-noise ratios of order 10^{-1} (Palovcak et al., 2020), and beneath the shot noise sits structural noise of the same order as the signal, the contribution of vitreous ice and other amorphous material that differs from particle to particle (Baxter et al., 2009). Recovering structure means pooling information across particles or drawing on prior knowledge of protein conformations. Standard reconstruction groups particles and averages them into density maps (Scheres, 2012; Punjani et al., 2017), suppressing variation within each group.

Heterogeneous reconstruction retains more of this variation by assigning each particle a latent coordinate and decoding it into a volume or displacement field (Zhong et al., 2021; Punjani & Fleet, 2023; Gilles & Singer, 2025). Its latent axes need not correspond to physical degrees of freedom (Klindt et al., 2024), and atomic models must subsequently be built into the densities (Jamali et al., 2024). Methods that deform an explicit structure, select among candidate structures, or optimise a fixed number of structures and their populations work directly in coordinates, within the structures or motions each is given (Schwab et al., 2024; Li et al., 2024; Dingeldein et al., 2025; Silva-Sánchez et al., 2026). These structural assumptions can strongly influence the result (Schwab et al., 2024).

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.

Generative models of protein ensembles provide a prior from which new conformations can be sampled (Lewis et al., 2025; Jing et al., 2024). They are usually conditioned on sequence or a starting structure. Cryo-EM measurements have also been used to guide sampling, typically through a reconstructed density map (Levy et al., 2024a; Raghu et al., 2025; Maddipatla et al., 2026a).

Conditioning these models on individual particles would avoid the averaging behind those maps, provided a likelihood can relate the images to the generator's current, partially denoised structure. Training-free guidance provides a way to use such a likelihood (Chung et al., 2023b; Song et al., 2023), but in pixel space that likelihood depends on each particle's unknown pose. Marginalising over pose, a sharply multimodal nuisance, is difficult, and moving to an image embedding does not by itself solve the problem, because cryo-EM image embeddings also vary with orientation and imaging conditions (Edelberg & Lederman, 2023; Herreros et al., 2025; Seitz et al., 2022; Yan et al., 2026).

Imaging uncertainty also affects estimates of conformational populations. Counting particles in a learned coordinate space includes the spread caused by noise and viewing conditions, which can flatten inferred free energy barriers (Evans et al., 2026); because that spread is per-particle, more particles do not remove it. Existing approaches infer landscapes along image manifolds (Frank & Ourmazd, 2016), whose coordinates can depend on viewing direction (Seitz et al., 2022), or over predefined paths and simulation ensembles (Giraldo-Barreto et al., 2021; Włodarski et al., 2024).

Conditioning a structural generator and estimating populations thus rest on the same ingredient: a measure of conformation together with an estimate of its imaging uncertainty. Can particles, without poses or a map, condition a pretrained ensemble generator? And can the same measure of conformation, once its imaging uncertainty is known, recover conformational populations?

We construct this measure as a code shared by two frozen models, the protein ensemble generator BioEmu (Lewis et al., 2025) and the cryo-EM image encoder Cryo-IEF (Yan et al., 2026). On simulated structure–image pairs, we regress an internal representation, computed by BioEmu during denoising, on the image embedding. Because images of one conformation share a structural target, the regression suppresses variation associated with imaging (§3.2). The resulting code can be estimated both from a set of particles and from a partially denoised structure.

Guidance compares these two estimates, accounting for the uncertainty of each (Figure 1b). One estimate is uncertain because the structure is still partly noise; the other because the particles are noisy images taken under different imaging conditions. We measure both sources of scatter on the simulated structure–image pairs before processing the observed particles, and use their combined uncertainty to weight the disagreement. Code directions measured more precisely receive greater weight, and more particles reduce the uncertainty of the image estimate. We call the resulting term *evidence-weighted guidance* (Algorithm S1, Appendix B). It constrains a low-dimensional subspace of the generator's 512-dimensional representation, leaving the remaining directions to the prior.

To estimate populations, we use each candidate conformation's simulated images to measure its spread in the code, then fit candidate weights that account for this spread (§3.4). These weights give a free energy landscape.

Contributions.

(i) We condition a frozen protein ensemble generator on individual cryo-EM particle images, without reconstructing a density map or estimating particle poses (§3.2–3.3). In simulation, 128 particles of a single GroEL conformation steer sampling toward targets in either 0.2% tail of the generator's ensemble and toward 10 PDB structures, improving accuracy for every target (§4.2). Guidance improves accuracy on both lysenin targets; on DppA, whose open state the prior rarely samples, reweighting by the same likelihood halves median core RMSD to both PDB structures (§4.3).

(ii) We use the shared code to infer conformational populations while accounting for imaging uncertainty (§3.4). On imposed GroEL landscapes, this recovers barriers that direct particle counting does not resolve, and on an independent molecular-dynamics ensemble it locates the single basin (§4.4).

2 Related work

Heterogeneous reconstruction. The dominant response to continuous heterogeneity either estimates a per-particle latent and maps it back to a volume or a deformation field (Zhong et al., 2021; Punjani & Fleet, 2021; 2023; Gilles & Singer, 2025), sometimes inferring orientation jointly (Levy et al., 2022; 2025), or embeds the images and reads a landscape off the embedding (Dashti et al., 2014; Frank & Ourmazd, 2016). Latent distance and density are generally not conformational distance and density. The measurement deforms the metric (Lederman & Toader, 2023), and the density is recovered only by linear-subspace methods that deconvolve the embedding (Gilles & Singer, 2025). A latent read from the image rather than from the structure is moreover not pose-invariant (Edelberg & Lederman, 2023; Seitz et al., 2022; 2023).

Structure-space methods. Another family models a reference structure explicitly, as rigid bodies, pseudo-atoms or atoms, and asks how it must move (Jin et al., 2014; Nakane et al., 2018; Schwab et al., 2024; Li et al., 2024; Ducrocq et al., 2025), so the representable motion is a deformation of that reference within a class each method fixes. A further family leaves its candidate structures fixed and scores each particle against a pre-enumerated set (Giraldo-Barreto et al., 2021; Tang et al., 2023; Dingeldein et al., 2025), so a conformation absent from the set has no support. One method does move the structures, against a particle-image likelihood under a force-field prior, but over a set whose size is fixed in advance (Silva-Sánchez et al., 2026).

Conditioning a frozen generator. Generative models of conformational ensembles supply the learned prior these families lack (Lewis et al., 2025; Jing et al., 2024). Conditioning a pretrained structure model on a reconstructed density map (Levy et al., 2024a; Raghu et al., 2025), fine-tuning one on such maps (Shenoy et al., 2026), or optimising a frozen predictor's input or internal embeddings against an experimental likelihood (Fadini et al., 2026; Li et al., 2026) all assume picking, pose estimation and reconstruction have succeeded. Training-free guidance takes a forward map as given, an operator or a predictor (Chung et al., 2023b; Song et al., 2023; Bansal et al., 2024), which a raw particle does not supply, because its unknown orientation makes the true likelihood a marginal over a sharply multimodal nuisance. Non-isotropic likelihood covariances have been built from a known operator (Song et al., 2023), per singular value (Kawar et al., 2022), or from the denoiser's covariance pushed through the operator (Boys et al., 2024). Ours is measured on the observation side, from repeated simulated images of one conformation, and sharpens with the number of particles pooled. Appendix J gives the evidence for this section's claims and the families left out here.

3 Structural inference in a shared code

Sections 3.1 and 3.2 fit one code on simulated structure–image pairs, which §3.3 uses to guide sampling and §3.4 to infer populations. Both pretrained models stay frozen throughout.

3.1 Paired structure and image representations

BioEmu (Lewis et al., 2025) generates conformations by reverse diffusion over Cα positions and per-residue rotations, which make up the state x_t, with the score function $s_\theta(x_t, t)$. We extract $h(x_t, t) \in \mathbb{R}^{512}$ from its first structure-module layer, mean-pooled over residues (Appendix G). Cryo-IEF (Yan et al., 2026) maps a particle image I to an embedding $e(I) \in \mathbb{R}^{768}$. The representation h describes a noisy conformation; the embedding e describes both conformation and imaging.

To relate them, we build a paired dataset from 3,200–3,240 conformations per protein sampled by BioEmu (Figure 1a, Table S8). Each conformation c is simulated as 128 images $I_{c,v}$ with cryoJAX (O'Brien et al., 2026) and forward-noised at the sampler's 50 diffusion times to give $h_{c,t}$. Images and representations are paired through their common conformation, so fitting the code needs no collective variable or conformational annotation.

3.2 A shared conformation code

The code is defined on the image side, by a map that predicts the generator's representation; a second fit then regresses the code back on that representation.

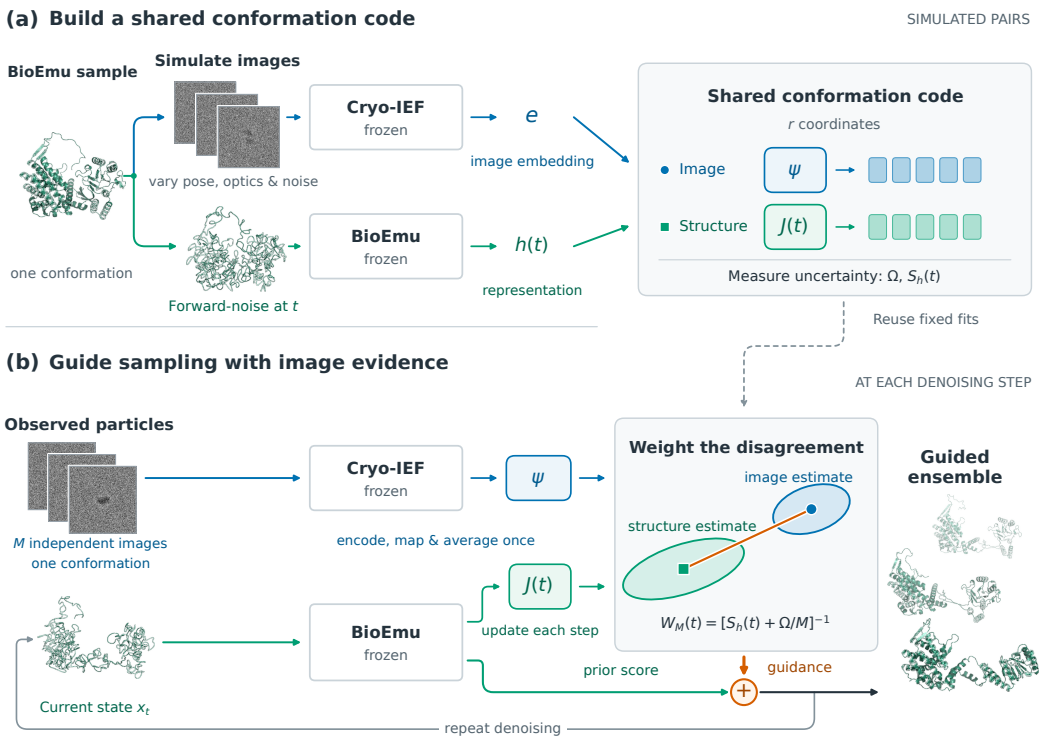

Figure 1: **(a)** The paired dataset, built once for each system. One conformation from the frozen generator is simulated as particle images, with orientation, optics and noise drawn independently for each image, and embedded by the frozen encoder; separately, it is noised to a diffusion time t, and its representation is extracted from the frozen generator's first structure-module layer. Two maps fitted in sequence, ψ on the image side and then $J(t)$ on the structure side, map to the same coordinates, and the scatter of each, Ω and $S_h(t)$, is measured on this dataset. **(b)** One denoising step. The M observed particles are encoded and averaged. The representation of the partially noised structure is recomputed at every step. The two estimates are compared under the precision $W_M(t) = (S_h(t) + \Omega/M)^{-1}$, and their weighted disagreement joins the generator's score.

Let $h_{c,t}$ be the representation of conformation c at noise level t, $\overline{h}(t)$ the dataset mean at that time, and h_c^* the average of $h_{c,t} - \overline{h}(t)$ over the 50 diffusion times. We fit an image map $\phi : \mathbb{R}^{768} \to \mathbb{R}^r$ and a structural decoder $A : \mathbb{R}^r \to \mathbb{R}^{512}$, shared across noise levels, by reduced-rank regression (Anderson, 1951; Izenman, 1975),

$$\min_{\phi, A} \sum_{c,v} \|h_c^* - A\phi(e_{c,v})\|_2^2, \qquad \phi(e) \in \mathbb{R}^r. \tag{1}$$

All images of a conformation share its target h_c^*, so the regression favours image features that predict conformation over those that vary with imaging (§4.2). The code dimension r is selected by the held-out error of Eq. 1 (Appendix D): 5 for GroEL, 7 for lysenin and 6 for DppA.

Equation 1 determines only a subspace, since $\phi \to Q\phi$ with $A \to AQ^{-1}$ leaves the objective unchanged. A fixed affine transformation, set once on the paired dataset, removes that freedom by giving the conformation-mean codes zero mean and identity covariance and making $A^\top A$ diagonal (Appendix E); we write $\psi(e)$ for the code in these coordinates. A conformation's reference code is the average over its $V = 128$ simulated images,

$$g_c = \frac{1}{V} \sum_{v=1}^{V} \psi(e_{c,v}). \tag{2}$$

Guidance needs a map from the generator's representation to the code at each diffusion time, with its own calibrated error. The decoder A has a left inverse, but it does not depend on t, so we fit a

second, reverse map with the codes g_c as targets. With ψ and the g_c held fixed, a ridge regression at each of the 50 times, on one pair $(h_{c,t}, g_c)$ per conformation, maps the representation to the code,

$$\hat{g}_h(x_t, t) = J(t)\big[h(x_t, t) - \overline{h}(t)\big]. \tag{3}$$

Centring makes Eq. 3 intercept-free, and its penalty lets the fit decline to answer by shrinking $J(t)$ toward zero, which returns the dataset's mean code. At high noise, the representation carries almost no information about the code, and at any one time some code directions are predicted far better than others, so a single shared penalty would over-shrink the predictable ones to cover the rest. The penalty is therefore selected on held-out error, per time and per code coordinate (Appendix E).

3.3 Conditioning the ensemble generator

Guidance compares two estimates of one conformation's code (Figure 1b): the mean code $\hat{g}_z = M^{-1}\sum_{m=1}^{M}\psi(e(I_m))$ of the M observed particles, independent images of that conformation, and the code $\hat{g}_h(x_t, t)$ that Eq. 3 reads off the generator's current representation at each denoising step. Writing g for the conformation's code, we model $\hat{g}_z \mid g \sim \mathcal{N}(g, \Omega/M)$ and $g \mid x_t, t \sim \mathcal{N}\big(\hat{g}_h(x_t, t), S_h(t)\big)$, with the particles independent of x_t given g. Marginalising g (Roweis & Ghahramani, 1999) adds the two covariances, and with $\delta_t = \hat{g}_z - \hat{g}_h(x_t, t)$ the log-likelihood of $\hat{g}_z$ given the current state is

$$\ell_t(x_t) = -\frac{1}{2}\,\delta_t^{\top} W_M(t)\,\delta_t + \text{const}, \qquad W_M(t) = \Big(S_h(t) + \frac{\Omega}{M}\Big)^{-1}. \tag{4}$$

The image covariance Ω is the scatter of single-image codes about the mean code of their conformation, and the structural covariance $S_h(t)$ that of the out-of-fold error in predicting a conformation's code from $h_{c,t}$. Both are measured on the paired dataset and pooled over its conformations (Appendix E), so the likelihood assumes that the observed particles were imaged under conditions the dataset covers; Appendix F.1 tests departures from this. More particles shrink Ω/M, but at a given diffusion time the precision $W_M(t)$ cannot exceed $S_h(t)^{-1}$. At the noisiest diffusion times the prediction falls back to the mean code (§3.2) and $S_h(t)$ rises to the spread of codes across conformations. The guidance term is therefore negligible early in sampling and grows as the structure becomes informative, without an imposed schedule.

Differentiating through the frozen generator gives the evidence-weighted guidance term,

$$u_t = \nabla_{x_t}\ell_t(x_t) = \left(\frac{\partial h(x_t, t)}{\partial x_t}\right)^{\top} J(t)^{\top} W_M(t)\,\delta_t. \tag{5}$$

The likelihood depends on the 512-dimensional representation only through the r combinations $J(t)h$, so the particles constrain at most r of its directions and the generator determines the rest. Each guided score evaluation costs one extra backward pass through BioEmu; no images are simulated during sampling.

In the position channel, the sampler uses the score

$$s_{\text{used}}(x_t, t) = s_\theta(x_t, t) + \kappa_x P u_t, \tag{6}$$

where κ_x sets the guidance strength and P smooths u_t along the chain, favouring collective displacements (Appendix B). Adaptive projected guidance (Sadat et al., 2025) stabilises the rotation updates. Completed samples receive a geometry repair, a restrained refinement followed by projection onto bond-length and clash bounds (Ryckaert et al., 1977; Appendix G). Table S10 lists every setting and how it was fixed. For DppA, we employ sequential Monte Carlo with the same likelihood to counteract prior bias (Appendix C).

3.4 Inferring conformational populations

The guidance likelihood assumed all M particles show one conformation, and asked what that conformation is. Here we drop that assumption and ask the complementary question: given particles of mixed conformations, how often is each one visited? The difficulty is that imaging displaces each particle's code from that of its conformation, so a spread of codes is partly population and partly

measurement, and naive particle counting would attribute all of it to population and drastically flatten the landscape.

We retain the individual image codes $y_n = \psi(e(I_n))$, $n = 1, \ldots, N$, and model their distribution as a mixture over candidate conformations. Let $K_i(y)$ be the code distribution produced by images of candidate i. The observed distribution is modelled as

$$p(y \mid w) = \sum_i w_i K_i(y), \qquad w_i \geq 0, \qquad \sum_i w_i = 1. \tag{7}$$

The weights are the unknown candidate populations. Accounting for the spread of each K_i separates population variation from the variation introduced by imaging.

We estimate K_i as a kernel density estimate on the codes of 128 simulated images of candidate i, one of 300 drawn from the generator's ensemble. Writing $y_{i,v}$ for those codes and $\overline{y}_i$ for their mean,

$$K_i(y) = \frac{1}{V} \sum_{v=1}^{V} T_{b_i}(y - a_{i,v}), \qquad a_{i,v} = \overline{y}_i + s\big(y_{i,v} - \overline{y}_i\big). \tag{8}$$

Where guidance pooled imaging scatter into Ω, this estimate keeps the shape of each candidate's image cloud. Each image contributes one kernel centred at $a_{i,v}$, with coordinate-wise widths b_i measured from that candidate's image-code scatter. The kernel T_{b_i} is a product of Student-t densities. The single scale s expands or contracts the cloud about its mean. The observed particles come from conformations outside the candidate library, so their image scatter need not match that of the candidates, and one fitted scalar absorbs the difference (Appendix I).

The landscape can be read out along any set of coordinates derivable from the candidates, either the code itself or a collective variable of the system. For GroEL the landscape is reported along a structural coordinate ξ_i, the apical–equatorial centroid distance of each candidate (Appendix H). This coordinate is computed from the candidate structures and never enters the shared-code regressions or the guidance likelihood. It does enter the population fit, whose log-weights must vary smoothly along ξ. The weights are $w(\vartheta) = \mathrm{softmax}(B\vartheta)$, with B a harmonic basis evaluated at the candidates' values of ξ, and we fit

$$\widehat{\vartheta} = \arg\min_{\vartheta} \left\{ -\frac{1}{N} \sum_{n=1}^{N} \log \left[\sum_i w_i(\vartheta) K_i(y_n) \right] + \frac{\rho}{2} \vartheta^\top R \vartheta \right\}. \tag{9}$$

The matrix R penalises second differences of $B\vartheta$ along candidates ordered by ξ_i, which stabilises the fit when candidate image distributions overlap.

The basis has a resolution H, the width along ξ that one particle's code cannot resolve. It is measured from the candidates' image-code scatter (Appendix I). The penalty strength ρ is selected on simulated benchmarks with known landscapes; the selection procedure and its limitations are reported in Appendix I.

The fitted weights define a free energy profile along ξ,

$$\frac{F_\Delta(\xi)}{k_\mathrm{B}T} = -\log \left[\sum_i \widehat{w}_i \mathcal{N}\big(\xi; \xi_i, \Delta^2\big) \right] + \mathrm{const}, \tag{10}$$

where Δ is the reporting width set from the spacing of the candidates along ξ.

4 Results

4.1 Systems and evaluation

GroEL is the main system for both guidance and population inference. As a chaperone it assists protein folding through an ATP-driven cycle involving large apical-domain movements associated with substrate binding and encapsulation (Clare et al., 2012; Saibil et al., 2013). We study one 547-residue subunit, using its extended and compact conformations to probe this domain motion.

All particles are simulated at a whole-box signal-to-noise ratio of 0.005, half CryoBench's standard (Jeon et al., 2024; Appendix G); unless stated otherwise, their structures come from the BioEmu

reference ensembles. For guidance, no target conformation (§4.2–4.3) is in the paired data used to fit either map; Appendix A and Table S10 detail the settings.

We assess guidance by its structural accuracy and the spread of the generated ensemble, measured over each protein's stable core, the residues assigned to helix or strand in at least half of its reference ensemble. For analysis we pick one collective variable ξ per system. For GroEL, this is its domain motion, the apical–equatorial centroid distance of §3.4; for lysenin and DppA, ξ is the radius of gyration of the core. These coordinates play no part in guidance. Accuracy is measured separately by Cα RMSD after superposition on the core, with the same residues used to compute the deviation (Appendix H).

4.2 Steering GroEL toward withheld conformations and PDB structures

Before sampling, we check what the fitted code retains. On conformations excluded from the image-map fit, its prediction of GroEL's ξ matches the best 5-dimensional linear summary of the original embedding for that quantity ($R^2 = 0.473$ against 0.475), whereas for defocus the code reaches $R^2 = 0.008$ against 0.697 for the best 5-dimensional summary (Table S6).

We choose targets in the tails of the GroEL reference ensemble: an extended conformation at the 99.9th percentile of ξ and a compact conformation in the lowest 0.2%. Both are rarely sampled by the unguided generator. Each target is observed through 128 independently simulated images. Although we guide with 128 images per target, the ξ predicted from their mean code already settles within its fitting error after about 13 images for the compact target and 36 for the extended one (Appendix D.2).

Guidance moves both ensembles toward their targets, narrows them along ξ and improves core RMSD. The median of ξ moves from 4.42 to 7.00 nm toward the extended target at 7.14 nm, and to 3.90 nm toward the compact target at 3.66 nm (Table S3). The ensembles narrow along ξ by factors of 2.8 and 3.7, respectively. Median core RMSD improves from 1.81 to 1.22 nm for the extended target and from 0.61 to 0.44 nm for the compact target, using 96 guided samples per target and 300 unguided samples (Figure 3a).

Figure 2 repeats the procedure with 10 GroEL structures from the PDB as targets; median core RMSD to the PDB structure improves on all 10.

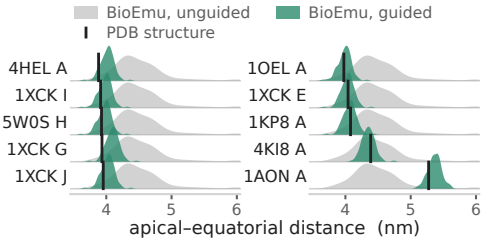

Figure 2: Guided and unguided samples along ξ, with each GroEL chain of Table S12 as the target.

4.3 Guidance moves lysenin; reweighting reaches DppA's open state

Lysenin extends part of its soluble structure into a β-hairpin during pore formation (Bokori-Brown et al., 2016). Its isolated 309-residue chain is too small and variable for conventional single-particle cryo-EM; we adjust the simulator to render it (Table S9) and use it as a stress test for flexible proteins. DppA, a periplasmic dipeptide-binding protein, closes two lobes about a hinge, and its targets are the open and closed PDB chains (Nickitenko et al., 1995; Dunten & Mowbray, 1995).

On lysenin, guidance moves the ensemble toward both targets along ξ (Figure 3b; Table S11), and median core RMSD improves from 2.31 to 1.63 nm (extended) and from 1.59 to 1.27 nm (compact). The samples still vary outside the stable core, where a 7-dimensional code leaves the flexible segments to the prior.

Only 9 of DppA's 281 unguided samples reach its open state (criterion in App. F.1), and guidance does not raise that rate, as the prior dominates. The likelihood still favours the open state, so we instead reweight BioEmu's own trajectories by it with sequential Monte Carlo (Singhal et al., 2025): each population of 24 unguided trajectories is resampled by the likelihood at intermediate noise levels and yields one sample. With this median core RMSD falls from 0.53 to 0.26 nm (open) and from 0.20 to 0.10 nm (closed; Figure 3c, Appendix C).

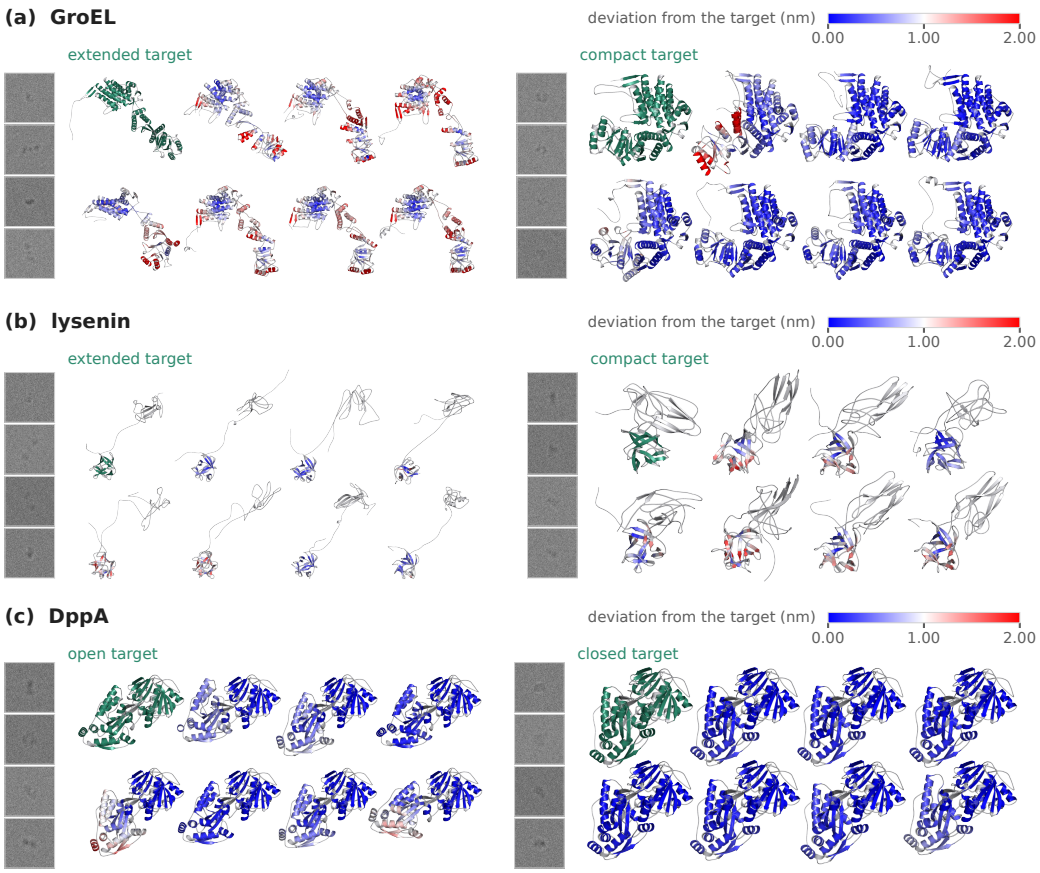

Figure 3: Samples on **(a)** GroEL and **(b)** lysenin under guidance and on **(c)** DppA, sampled via sequential Monte Carlo, at a whole-box signal-to-noise ratio of 0.005. Per protein, the extended (DppA: open, 1DPE:A) target on the left and the compact (closed, 1DPP:A) one on the right. Each panel shows 4 of the 128 particles, the target (green) and 7 samples, coloured by each residue's Cα distance to the target after superposition on the stable core. Only core residues with one secondary structure in at least 98% of the reference ensemble are coloured. Each protein has its own guidance target selection: Appendix H.

4.4 Recovering conformational populations

From GroEL's paired dataset, we reserve 300 conformations as candidates and simulate the observed particles from the remaining 2,900. By importance-resampling along ξ, we impose four landscapes on them: separated states, a rare state containing 12% of the population, a deep barrier, and two states at the minimum separation (Figure 4b–e). For each landscape, we draw 4,000 conformations and represent each by one of its 128 simulated images, chosen at random (Appendix I).

Figure 4 compares three landscapes per test: the true landscape (black), particle counting (orange) and the mixture fit (blue). All three use the reporting width Δ of §3.4. For this candidate library $H = 0.412$ nm, so a barrier is reliably observed only between minima at least 0.82 nm apart.

Particle counting resolves no barrier in any of the four cases. The mixture fit recovers both basins in the first three (Figure 4b–d) on 8 of 9 replicate particle sets, with barrier heights within $0.60\,k_{\mathrm{B}}T$ of the true landscape. At the minimum separation (Figure 4e), the found barrier is barely visible, exposing the practical resolution limit (Appendix I.5).

Figure 4a applies the same estimator, candidates and settings to particles simulated from an independent molecular-dynamics ensemble of the subunit (3,006 frames, Vander Meersche et al., 2024), with s fitted to these particles. Its minimum is recovered at 4.61 nm against 4.62 nm, with an RMS deviation of $0.10\,k_{\mathrm{B}}T$ over the range the ensemble occupies (Appendix I.5).

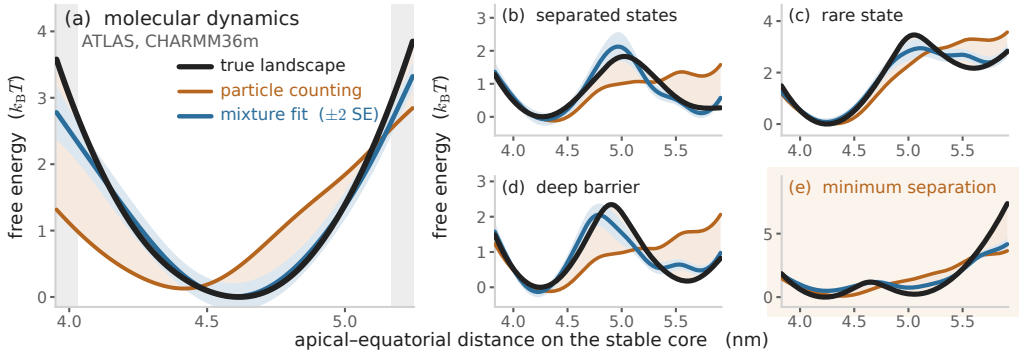

Figure 4: Free energy landscapes from 4,000 particles of mixed conformations, with resolution limit $H = 0.412$ nm: true landscape (black), particle counting (orange) and the mixture fit (blue, ± 2 standard errors from particle sampling alone, Appendix I), all at the reporting width Δ of Eq. 10. The shaded area between the true and counted curves is what imaging spread removes. **(a)** An independent molecular-dynamics ensemble of the subunit (Vander Meersche et al., 2024; Huang et al., 2017), read with the generator's non-MD candidates; grey marks the range it does not visit. **(b–e)** Landscapes imposed on the conformations outside the candidate library: separated states, a rare state, a deep barrier, and two states at the minimum separation.

5 Discussion and Conclusion

We have conditioned a frozen generative model of protein ensembles on individual cryo-EM particles through a low-dimensional code that the generator and a frozen image encoder share, fitted on simulated structure–image pairs. On GroEL, 128 particles of a withheld conformation move the ensemble toward it and narrow it along ξ, and median core RMSD improves with experimental GroEL structures as targets as well as two other systems. The same code, with the image scatter measured for each candidate, turns particles of mixed conformations into populations and recovers barriers that particle counting does not.

So few particles suffice because the likelihood constrains only a handful of directions in the generator's representation and leaves the rest to the prior. Those directions carry the domain-scale motion that separates the conformations, and secondary structure and local packing come from the generator unchanged. Reconstruction from particles alone needs orders of magnitude more of them because the images have to determine every coordinate, whereas here they determine which region of the prior's ensemble the molecule occupies.

All particles are simulated, the reference ensembles are the generator's samples, and the imposed landscapes draw their candidates and particles from the paired dataset the code was fitted on. The withheld conformations and imposed populations show, respectively, that inference is not tied to the conformations or the populations the code was fitted on, but they cannot show that guidance reaches conformations the generator does not already sample. Whether the code and its covariances transfer from simulated to experimental imaging, and how the method behaves when the prior is wrong, are open questions that require experimental particles and independently determined structures.

The guidance likelihood assumes that the particles show one conformation. Standard pipelines already produce such groups, since 3D classification and heterogeneous reconstruction sort a dataset into conformational classes, and guidance would run on each class after that step. Because it needs so few particles, it can be applied to classes too small to reconstruct on their own, which are the rare states whose structures are hardest to obtain by other means.

AI use statement

In this work, we used generative AI tools to implement methods: much of the software that implements the guidance operator, runs the experiments and analyses their output was written with AI assistance. AI tools also contributed to parts of the research methodology design, to the design of the individual experiments, to proposing and refining the hypotheses those experiments test, to parts

of the data analysis (especially writing visualisation code), and to the interpretation of results. We have not used generative AI tools to develop the core conceptual framework, to formulate mathematical claims, to supply critical ingredients for proving them, to assist in the writing of proofs, or to clean data sets. Generating synthetic data sets, assisting translation, and transcribing recordings are not applicable to this work in the sense of the policy: the simulated particle images come from the image simulation of §3.1 and Appendix G, and the conformations from the frozen pretrained generator that is itself the object of study, neither of which is a generative AI writing tool. Additionally, we used generative AI tools for brainstorming, for creating and editing the figures, for drafting and editing parts of the paper text, for editing code, for identifying related literature and for formatting references. We have reviewed all AI-assisted work. The authors wrote the text partly manually and partly with AI assistance, but all text was human-read and, in part, rewritten. We take responsibility for the final content of this work, including text and claims produced with the aid of generative AI.

Ethics statement

This work involves no human subjects, no personal data and no new data collection. Every particle image used here is simulated; the structural inputs are public structures from the PDB and public molecular-dynamics trajectories, used under their own terms. We see no direct pathway to harm from the method, which infers protein conformations and their populations from images of the molecule being studied.

Reproducibility statement

The measurement model, the guided sampler and the population estimator are specified in §3 and expanded in the appendices, one appendix per component a reader would have to rebuild: App. B and Algorithm S1 give the sampler and the order its operators are applied in, App. C and Algorithm S2 the sequential Monte Carlo reweighting used on DppA, App. E the shared code and its two covariances, and App. I the population fit. App. G lists the inputs, the data splits that keep every reported target out of every fit, and, in Table S10, every numerical setting together with how it was fixed; App. A states the study inventory, and App. H the structural coordinates. Source code for the method, the experiment drivers and the producers of the body's figures and numbers is submitted as supplementary material, together with each system's fitted maps and covariances, the PDB target chains and the inputs from which Figures 1, 2 and 4b–e are redrawn; the two frozen pretrained models the guidance operator reads are identified there by repository and commit, and the side-chain packer used to build all-atom structures is identified by citation. App. G states which measured artefacts and producers the archive does not contain. Most of these omissions are the result of the submission size limit.

References

Eya Abid and Slavica Jonić. Learning mappings from Cryo-EM images to atomic coordinates via latent representations. bioRxiv 2026.02.16.706167, 2026. Posted 18 February 2026.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations (ICLR) Workshop Track*, 2017.

T. W. Anderson. Estimating linear restrictions on regression coefficients for multivariate normal distributions. *The Annals of Mathematical Statistics*, 22(3):327–351, 1951. doi: 10.1214/aoms/1177729580.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024.

Miro A. Astore, Geoffrey Woollard, David Silva-Sánchez, Wenda Zhou, Mykhailo Kopylov, Khanh Dao Duc, Roy R. Lederman, Yilai Li, Yi Zhou, Jing Yuan, Fei Ye, Quanquan Gu, Rémi Vuillemot, Slavica Jonić, Lan Dang, Steven J. Ludtke, Hannah Bridges, Serena Liu, Michael McLean, Valentin Peretroukhin, Johannes Schwab, Eduardo R. Cruz-Chú, Peter Schwander, Marc Aurèle Gilles, Amit Singer, David Herreros, Jose-Maria Carazo, Carlos O. S. Sorzano, J. Ryan Feathers, Ellen D. Zhong, Nikolaus Grigorieff, Pilar Cossio, and Sonya M. Hanson.

The inaugural Flatiron Institute Cryo-EM conformational heterogeneity challenge. bioRxiv 2025.07.18.665582, 2025. Version 2 posted 6 November 2025; not peer reviewed.

Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *International Conference on Learning Representations (ICLR)*, 2024.

Cecilia Bartolucci, Doriano Lamba, Saulius Grazulis, Elena Manakova, and Hermann Heumann. Crystal structure of wild-type chaperonin GroEL. *Journal of Molecular Biology*, 354(4):940–951, 2005. doi: 10.1016/j.jmb.2005.09.096.

Heinz H. Bauschke and Jonathan M. Borwein. On projection algorithms for solving convex feasibility problems. *SIAM Review*, 38(3):367–426, 1996. doi: 10.1137/S0036144593251710.

William T. Baxter, Robert A. Grassucci, Haixiao Gao, and Joachim Frank. Determination of signal-to-noise ratios and spectral SNRs in cryo-EM low-dose imaging of molecules. *Journal of Structural Biology*, 166(2):126–132, 2009. doi: 10.1016/j.jsb.2009.02.012.

Monika Bokori-Brown, Thomas G. Martin, Claire E. Naylor, Ajit K. Basak, Richard W. Titball, and Christos G. Savva. Cryo-EM structure of lysenin pore elucidates membrane insertion by an aerolysin family protein. *Nature Communications*, 7:11293, 2016. doi: 10.1038/ncomms11293.

Benjamin Boys, Jakiw Pidstrigach, Mark Girolami, Sebastian Reich, Alan Mosca, and O. Deniz Akyildiz. Tweedie moment projected diffusions for inverse problems. *Transactions on Machine Learning Research*, 2024.

Kerstin Braig, Paul D. Adams, and Axel T. Brünger. Conformational variability in the refined structure of the chaperonin GroEL at 2.8 Å resolution. *Nature Structural Biology*, 2(12):1083–1094, 1995. doi: 10.1038/nsb1295-1083.

Charu Chaudhry, Arthur L. Horwich, Axel T. Brünger, and Paul D. Adams. Exploring the structural dynamics of the *E. coli* chaperonin GroEL using translation-libration-screw crystallographic refinement of intermediate states. *Journal of Molecular Biology*, 342(1):229–245, 2004. doi: 10.1016/j.jmb.2004.07.015.

Muyuan Chen and Steven J. Ludtke. Deep learning-based mixed-dimensional Gaussian mixture model for characterizing variability in cryo-EM. *Nature Methods*, 18(8):930–936, 2021. doi: 10.1038/s41592-021-01220-5.

Hyungjin Chung, Jeongsol Kim, Sehui Kim, and Jong Chul Ye. Parallel diffusion models of operator and image for blind inverse problems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6059–6069, 2023a. doi: 10.1109/CVPR52729.2023.00587.

Hyungjin Chung, Jeongsol Kim, Michael T. McCann, Marc L. Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations (ICLR)*, 2023b.

Daniel K. Clare, Daven Vasishtan, Scott Stagg, Joel Quispe, George W. Farr, Maya Topf, Arthur L. Horwich, and Helen R. Saibil. ATP-triggered conformational changes delineate substrate-binding and -folding mechanics of the GroEL chaperonin. *Cell*, 149(1):113–123, 2012. doi: 10.1016/j.cell.2012.02.047.

Luigi De Colibus, Andreas F.-P. Sonnen, Keith J. Morris, C. Alistair Siebert, Pasquale Abrusci, Judith Plitzko, Vesna Hodnik, Martin Leippe, Elena Volpi, Gregor Anderluh, and Robert J. C. Gilbert. Structures of lysenin reveal a shared evolutionary origin for pore-forming proteins and its mode of sphingomyelin recognition. *Structure*, 20(9):1498–1507, 2012. doi: 10.1016/j.str.2012.06.011.

Haizhao Dai, Qihe Chen, Lingqi Li, Yingjun Shen, Zhenyang Xu, Minzhang Li, Yufan Xie, Jingjing Zheng, Zhijie Liu, Liping Sun, Yuan Pei, Jiakai Zhang, and Jingyi Yu. Cryo-EM as latent structural landscape microscopy. bioRxiv 2026.04.10.717737, 2026. Version 2 posted 17 April 2026.

Ali Dashti, Peter Schwander, Robert Langlois, Russell Fung, Wen Li, Ahmad Hosseinizadeh, Hstau Y. Liao, Jesper Pallesen, Gyanesh Sharma, Vera A. Stupina, Anne E. Simon, Jonathan D. Dinman, Joachim Frank, and Abbas Ourmazd. Trajectories of the ribosome as a Brownian nanomachine. *Proceedings of the National Academy of Sciences*, 111(49):17492–17497, 2014. doi: 10.1073/pnas.1419276111.

Fernando de Helguero. Sui massimi delle curve dimorfiche. *Biometrika*, 3(1):84–98, 1904. doi: 10.1093/biomet/3.1.84.

Marc Delarue and Philippe Dumas. On the use of low-frequency normal modes to enforce collective movements in refining macromolecular structural models. *Proceedings of the National Academy of Sciences*, 101(18):6957–6962, 2004. doi: 10.1073/pnas.0400301101.

Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, pp. 8780–8794, 2021.

Lars Dingeldein, David Silva-Sánchez, Luke Evans, Edoardo D'Imprima, Nikolaus Grigorieff, Roberto Covino, and Pilar Cossio. Amortized template matching of molecular conformations from cryoelectron microscopy images using simulation-based inference. *Proceedings of the National Academy of Sciences*, 122(23):e2420158122, 2025. doi: 10.1073/pnas.2420158122.

Gabriel Ducrocq, Lukas Grunewald, Sebastian Westenhoff, and Fredrik Lindsten. cryoSPHERE: Single-Particle HEterogeneous REconstruction from cryo EM. In *International Conference on Learning Representations (ICLR)*, 2025.

Pete Dunten and Sherry L. Mowbray. Crystal structure of the dipeptide binding protein from *Escherichia coli* involved in active transport and chemotaxis. *Protein Science*, 4(11):2327–2334, 1995. doi: 10.1002/pro.5560041110.

Daniel G. Edelberg and Roy R. Lederman. Using VAEs to learn latent variables: Observations on applications in cryo-EM. arXiv preprint arXiv:2303.07487, 2023.

Dave Epstein, Allan Jabri, Ben Poole, Alexei A. Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023.

Luke Evans, Lars Dingeldein, Roberto Covino, Marc Aurèle Gilles, Erik Thiede, and Pilar Cossio. Counting particles in cryo-electron microscopy may result in incorrect population estimates. *Communications Biology*, 9(1):421, 2026. doi: 10.1038/s42003-026-09859-6. Earlier version: bioRxiv 2025.03.27.644168.

Alisia Fadini, Minhuan Li, Airlie J. McCoy, Suresh Banjara, Hiroki Okumura, Eve Napier, Pietro Fontana, Amir R. Khan, Luca Jovine, Thomas C. Terwilliger, Randy J. Read, Doeke R. Hekstra, and Mohammed AlQuraishi. AlphaFold as a prior: experimental structure determination conditioned on a pretrained neural network. *Nature Methods*, 23(4):785–795, 2026. doi: 10.1038/s41592-026-03047-4. ROCKET. Preprint: bioRxiv 2025.02.18.638828.

Xue Fei, Dong Yang, Nicole LaRonde-LeBlanc, and George H. Lorimer. Crystal structure of a GroEL-ADP complex in the relaxed allosteric state at 2.7 Å resolution. *Proceedings of the National Academy of Sciences*, 110(32):E2958–E2966, 2013. doi: 10.1073/pnas.1311996110.

Joachim Frank and Abbas Ourmazd. Continuous changes in structure mapped by manifold embedding of single-particle data in cryo-EM. *Methods*, 100:61–67, 2016. doi: 10.1016/j.ymeth.2016.02.007.

Marc Aurèle Gilles and Amit Singer. Cryo-EM heterogeneity analysis using regularized covariance estimation and kernel regression. *Proceedings of the National Academy of Sciences*, 122(9):e2419140122, 2025. doi: 10.1073/pnas.2419140122. Preprint: bioRxiv 2023.10.28.564422.

Julian Giraldo-Barreto, Sebastian Ortiz, Erik H. Thiede, Karen Palacio-Rodriguez, Bob Carpenter, Alex H. Barnett, and Pilar Cossio. A Bayesian approach to extracting free-energy profiles from cryo-electron microscopy experiments. *Scientific Reports*, 11(1):13657, 2021. doi: 10.1038/s41598-021-92621-1.

Alexandros Graikos, Srikar Yellapragada, and Dimitris Samaras. Conditional generation from unconditional diffusion models using denoiser representations. In *34th British Machine Vision Conference (BMVC)*. BMVA Press, 2023. The BMVC index page lists this paper as "Conditional Generation from *Pre-Trained* Diffusion Models using Denoiser Representations"; the camera-ready and arXiv 2306.01900 both read "Unconditional", which is the title used here.

Katherine Henzler-Wildman and Dorothee Kern. Dynamic personalities of proteins. *Nature*, 450 (7172):964–972, 2007. doi: 10.1038/nature06522.

David Herreros, Carlos Perez Mata, Chari Noddings, Deli Irene, James Krieger, David A. Agard, Ming-Daw Tsai, Carlos Oscar Sanchez Sorzano, and Jose Maria Carazo. Real-space heterogeneous reconstruction, refinement, and disentanglement of CryoEM conformational states with HetSIREN. *Nature Communications*, 16(1):3751, 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-59135-0. URL https://doi.org/10.1038/s41467-025-59135-0.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. arXiv:2207.12598, 2022. A short version appeared at the NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications.

Jing Huang, Sarah Rauscher, Grzegorz Nawrocki, Ting Ran, Michael Feig, Bert L. de Groot, Helmut Grubmüller, and Alexander D. MacKerell. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nature Methods*, 14(1):71–73, 2017. doi: 10.1038/nmeth.4067.

Alan Julian Izenman. Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5(2):248–264, 1975. doi: 10.1016/0047-259X(75)90042-1.

Kiarash Jamali, Lukas Käll, Rui Zhang, Alan Brown, Dari Kimanius, and Sjors H. W. Scheres. Automated model building and protein identification in cryo-EM maps. *Nature*, 628(8007):450–457, 2024. doi: 10.1038/s41586-024-07215-4.

Minkyu Jeon, Rishwanth Raghu, Miro Astore, Geoffrey Woollard, Ryan Feathers, Alkin Kaz, Sonya M. Hanson, Pilar Cossio, and Ellen D. Zhong. CryoBench: Diverse and challenging datasets for the heterogeneity problem in cryo-EM. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024) Track on Datasets and Benchmarks*, 2024.

Qiyu Jin, Carlos Oscar S. Sorzano, José Miguel de la Rosa-Trevín, José Román Bilbao-Castro, Rafael Núñez-Ramírez, Oscar Llorca, Florence Tama, and Slavica Jonić. Iterative elastic 3D-to-2D alignment method using normal modes for studying structural dynamics of large macromolecular complexes. *Structure*, 22(3):496–506, 2014. doi: 10.1016/j.str.2014.01.004.

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

Maarten Joosten, Joel Greer, James Parkhurst, Tom Burnley, and Arjen J. Jakobi. Roodmus: a toolkit for benchmarking heterogeneous electron cryo-microscopy reconstructions. *IUCrJ*, 11(6): 951–965, 2024. doi: 10.1107/S2052252524009321.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2.

Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32(5):922–923, 1976. doi: 10.1107/S0567739476001873.

Wolfgang Kabsch. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 34(5):827–828, 1978. doi: 10.1107/S0567739478001680.

Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983. doi: 10.1002/bip.360221211.

Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, pp. 23593–23606, 2022.

David A. Klindt, Aapo Hyvärinen, Axel Levy, Nina Miolane, and Frédéric Poitevin. Towards interpretable Cryo-EM: disentangling latent spaces of molecular conformations. *Frontiers in Molecular Biosciences*, 11:1393564, 2024. doi: 10.3389/fmolb.2024.1393564.

Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. In *International Conference on Learning Representations (ICLR)*, 2023.

Roy R. Lederman and Bogdan Toader. On manifold learning in Plato's cave: Remarks on manifold learning and physical phenomena. In *2023 International Conference on Sampling Theory and Applications (SampTA)*, pp. 1–7. IEEE, 2023. doi: 10.1109/SampTA59647.2023.10301403.

Bruce W. Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. Programming refusal with conditional activation steering. In *International Conference on Learning Representations (ICLR)*, 2025.

Axel Levy, Gordon Wetzstein, Julien N. P. Martel, Frédéric Poitevin, and Ellen D. Zhong. Amortized inference for heterogeneous reconstruction in cryo-EM. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, 2022.

Axel Levy, Eric R. Chan, Sara Fridovich-Keil, Frédéric Poitevin, Ellen D. Zhong, and Gordon Wetzstein. Solving inverse problems in protein space using diffusion-based priors. arXiv preprint arXiv:2406.04239, 2024a. ADP-3D.

Axel Levy, Rishwanth Raghu, David Shustin, Adele Rui-Yang Peng, Huan Li, Oliver Biggs Clarke, Gordon Wetzstein, and Ellen D. Zhong. Mixture of neural fields for heterogeneous reconstruction in cryo-EM. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024b.

Axel Levy, Rishwanth Raghu, J. Ryan Feathers, Michal Grzadkowski, Frédéric Poitevin, Jake D. Johnston, Francesca Vallese, Oliver Biggs Clarke, Gordon Wetzstein, and Ellen D. Zhong. CryoDRGN-AI: neural ab initio reconstruction of challenging cryo-EM and cryo-ET datasets. *Nature Methods*, 22(7):1486–1494, 2025. doi: 10.1038/s41592-025-02720-4. Preprint: bioRxiv 2024.05.30.596729.

Sarah Lewis, Tim Hempel, José Jiménez-Luna, Michael Gastegger, Yu Xie, Andrew Y. K. Foong, Victor García Satorras, Osama Abdin, Bastiaan S. Veeling, Iryna Zaporozhets, et al. Scalable emulation of protein equilibrium ensembles with generative deep learning. *Science*, 389(6761): eadv9817, 2025. doi: 10.1126/science.adv9817. Preprint: bioRxiv 2024.12.05.626885.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023.

Minhuan Li, Jiequn Han, Pilar Cossio, and Luhuan Wu. Robust inference-time steering of protein diffusion models via embedding optimization. arXiv:2602.05285 (v2, May 2026), 2026.

Yilai Li, Yi Zhou, Jing Yuan, Fei Ye, and Quanquan Gu. CryoSTAR: leveraging structural priors and constraints for cryo-EM heterogeneous reconstruction. *Nature Methods*, 21(12):2318–2326, 2024. doi: 10.1038/s41592-024-02486-1.

Junwen Liao, Dihan Zheng, Hui Zhang, Linfeng Zhang, Mingxu Hu, and Chenglong Bao. Fine-tuning AlphaFold with limited cryo-EM observations. *Communications Chemistry*, 9(1):95, 2026. doi: 10.1038/s42004-026-01899-7.

Yikai Liu, Zongxin Yu, Richard J. Lindsay, Guang Lin, Ming Chen, Abhilash Sahoo, and Sonya M. Hanson. ExEnDiff: An experiment-guided diffusion model for protein conformational ensemble generation. *PRX Life*, 3(2):023013, 2025. doi: 10.1103/PRXLife.3.023013.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR 2019)*, 2019.

Jiarui Lu, Bozitao Zhong, Zuobai Zhang, and Jian Tang. Str2Str: A score-based framework for zero-shot protein conformation sampling. In *International Conference on Learning Representations (ICLR)*, 2024.

Grace Luo, Trevor Darrell, Oliver Wang, Dan B Goldman, and Aleksander Holynski. Readout guidance: Learning control from diffusion features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8217–8227, 2024. doi: 10.1109/CVPR52733.2024.00785.

Zhenwei Luo, Fengyun Ni, Qinghua Wang, and Jianpeng Ma. OPUS-DSD: deep structural disentanglement for cryo-EM single-particle analysis. *Nature Methods*, 20(11):1729–1738, 2023. doi: 10.1038/s41592-023-02031-6.

Advaith Maddipatla, Nadav Sellam Bojan, Meital Bojan, Volodymyr Masalitin, Sanketh Vedula, Paul Schanda, Ailie Marx, and Alex M. Bronstein. Experiment-guided AlphaFold3 resolves measurement-consistent protein ensembles. *Nature Biotechnology*, 2026a. doi: 10.1038/s41587-026-03166-5. Published online 29 June 2026; journal version of bioRxiv 2025.10.11.681796.

Sai Advaith Maddipatla, Anar Rzayev, Marco Pegoraro, Martin Pacesa, Paul Schanda, Ailie Marx, Sanketh Vedula, and Alexander Bronstein. Inference-time optimization for experiment-grounded protein ensemble generation. In *International Conference on Machine Learning (ICML)*, 2026b.

Henry H. Mattingly, Luke Evans, and Pilar Cossio. Measurement-limited learning of conformational heterogeneity in cryo-electron microscopy. arXiv preprint arXiv:2606.14449, 2026.

Naoki Murata, Koichi Saito, Chieh-Hsin Lai, Yuhta Takida, Toshimitsu Uesaka, Yuki Mitsufuji, and Stefano Ermon. GibbsDDRM: A partially collapsed Gibbs sampler for solving blind inverse problems with denoising diffusion restoration. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 25501–25522. PMLR, 2023.

Takanori Nakane, Dari Kimanius, Erik Lindahl, and Sjors H. W. Scheres. Characterisation of molecular motions in cryo-EM single-particle data by multi-body refinement in RELION. *eLife*, 7: e36861, 2018. doi: 10.7554/eLife.36861.

Alexei V. Nickitenko, Sergei Trakhanov, and Florante A. Quiocho. 2 Å resolution structure of DppA, a periplasmic dipeptide transport/chemosensory receptor. *Biochemistry*, 34(51):16585–16595, 1995. doi: 10.1021/bi00051a006.

Michael J. O'Brien, David Silva-Sánchez, Geoffrey Woollard, Kwanghwi Je, Sonya M. Hanson, Daniel J. Needleman, Pilar Cossio, Erik Henning Thiede, and Miro A. Astore. CryoJAX: a cryo-electron microscopy image-simulation library in JAX. *Acta Crystallographica Section D: Structural Biology*, 82(3):155–167, 2026. doi: 10.1107/S2059798326000550.

Eugene Palovcak, Daniel Asarnow, Melody G. Campbell, Zanlin Yu, and Yifan Cheng. Enhancing the signal-to-noise ratio and generating contrast for cryo-EM images with convolutional neural networks. *IUCrJ*, 7(6):1142–1150, 2020. doi: 10.1107/S2052252520013184.

Yong-Hyun Park, Mingi Kwon, Jaewoong Choi, Junghyo Jo, and Youngjung Uh. Understanding the latent space of diffusion models through the lens of Riemannian geometry. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023.

Xinyu Peng, Ziyang Zheng, Wenrui Dai, Nuoqian Xiao, Chenglin Li, Junni Zou, and Hongkai Xiong. Improving diffusion models for inverse problems using optimal posterior covariance. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 40347–40370. PMLR, 2024.

Marjetka Podobnik, Peter Savory, Nejc Rojko, Matic Kisovec, Neil Wood, Richard Hambley, Jonathan Pugh, E. Jayne Wallace, Luke McNeill, Mark Bruce, Isabel Liko, Timothy M. Allison, Shabaz Mohammed, Doryen Bubeck, Carol V. Robinson, Gregor Anderluh, Mark I. Wallace, Bostjan Kobe, and Matthew T. Degiacomi. Crystal structure of an invertebrate cytolysin pore reveals unique properties and mechanism of assembly. *Nature Communications*, 7:11598, 2016. doi: 10.1038/ncomms11598.

Joris Postmus and Steven Abreu. Steering large language models using conceptors: Improving addition-based activation engineering. arXiv preprint arXiv:2410.16314, 2024. MINT workshop, NeurIPS 2024.

Barrett M. Powell and Joseph H. Davis. Learning structural heterogeneity from cryo-electron sub-tomograms with tomoDRGN. *Nature Methods*, 21(8):1525–1536, 2024. doi: 10.1038/s41592-024-02210-z.

Ali Punjani and David J. Fleet. 3D variability analysis: Resolving continuous flexibility and discrete heterogeneity from single particle cryo-EM. *Journal of Structural Biology*, 213(2):107702, 2021. doi: 10.1016/j.jsb.2021.107702.

Ali Punjani and David J. Fleet. 3DFlex: determining structure and motion of flexible proteins from cryo-EM. *Nature Methods*, 20(6):860–870, 2023. doi: 10.1038/s41592-023-01853-8.

Ali Punjani, John L. Rubinstein, David J. Fleet, and Marcus A. Brubaker. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nature Methods*, 14(3):290–296, 2017. doi: 10.1038/nmeth.4169.

Rishwanth Raghu, Axel Levy, Gordon Wetzstein, and Ellen D. Zhong. Multiscale guidance of protein structure prediction with heterogeneous cryo-EM data. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*, 2025.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20 (NeurIPS 2007)*, pp. 1177–1184, 2007. URL https://proceedings.neurips.cc/paper_files/paper/2007/file/013a006f03dbc5392effeb8f18fda755-Paper.pdf.

Soung-Hun Roh, Corey F. Hryc, Hyun-Hwan Jeong, Xue Fei, Joanita Jakana, George H. Lorimer, and Wah Chiu. Subunit conformational variation within individual GroEL oligomers resolved by cryo-EM. *Proceedings of the National Academy of Sciences*, 114(31):8259–8264, 2017. doi: 10.1073/pnas.1704725114.

Dan Rosenbaum, Marta Garnelo, Michal Zielinski, Charlie Beattie, Ellen Clancy, Andrea Huber, Pushmeet Kohli, Andrew W. Senior, John Jumper, Carl Doersch, S. M. Ali Eslami, Olaf Ronneberger, and Jonas Adler. Inferring a continuous distribution of atom coordinates from cryo-EM images using VAEs. arXiv preprint arXiv:2106.14108, 2021.

Litu Rout, Negin Raoof, Giannis Daras, Constantine Caramanis, Alexandros G. Dimakis, and Sanjay Shakkottai. Solving linear inverse problems provably via posterior sampling with latent diffusion models. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023.

Sam Roweis and Zoubin Ghahramani. A unifying review of linear Gaussian models. *Neural Computation*, 11(2):305–345, February 1999. ISSN 0899-7667. doi: 10.1162/089976699300016674.

Jean-Paul Ryckaert, Giovanni Ciccotti, and Herman J. C. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics*, 23(3):327–341, 1977. doi: 10.1016/0021-9991(77)90098-5.

Seyedmorteza Sadat, Otmar Hilliges, and Romann M. Weber. Eliminating oversaturation and artifacts of high guidance scales in diffusion models. In *International Conference on Learning Representations (ICLR)*, 2025.

Helen R. Saibil, Wayne A. Fenton, Daniel K. Clare, and Arthur L. Horwich. Structure and allostery of the chaperonin GroEL. *Journal of Molecular Biology*, 425(9):1476–1487, 2013. doi: 10.1016/j.jmb.2012.11.028.

A. K. Saxena and S. R. Meena. Crystal structure analysis of apo-GroEL structure. Protein Data Bank entry 4HEL, 2012. Deposited 2012-10-04, released 2013-09-25; no primary publication.

Sjors H. W. Scheres. RELION: Implementation of a Bayesian approach to cryo-EM structure determination. *Journal of Structural Biology*, 180(3):519–530, 2012. doi: 10.1016/j.jsb.2012.09.006. Companion: A Bayesian view on cryo-EM structure determination, J. Mol. Biol. 415(2):406–418, 2012.

Mark F. Schilling, Ann E. Watkins, and William Watkins. Is Human Height Bimodal? *The American Statistician*, 56(3):223–229, 2002. doi: 10.1198/00031300265.

Johannes Schwab, Dari Kimanius, Alister Burt, Tom Dendooven, and Sjors H. W. Scheres. DynaMight: estimating molecular motions with improved reconstruction from cryo-EM images. *Nature Methods*, 21(10):1855–1862, 2024. doi: 10.1038/s41592-024-02377-5.

Evan Seitz, Francisco Acosta-Reyes, Suvrajit Maji, Peter Schwander, and Joachim Frank. Recovery of conformational continuum from single-particle cryo-EM images: Optimization of ManifoldEM informed by ground truth. *IEEE Transactions on Computational Imaging*, 8:462–478, 2022. doi: 10.1109/TCI.2022.3174801.

Evan Seitz, Joachim Frank, and Peter Schwander. Beyond ManifoldEM: geometric relationships between manifold embeddings of a continuum of 3D molecular structures and their 2D projections. *Digital Discovery*, 2(3):702–717, 2023. doi: 10.1039/D2DD00128D.

Ashok Sekhar and Lewis E. Kay. NMR paves the way for atomic level descriptions of sparsely populated, transiently formed biomolecular conformers. *Proceedings of the National Academy of Sciences*, 110(32):12867–12874, 2013. doi: 10.1073/pnas.1305688110.

Yifei Shen, Xinyang Jiang, Yifan Yang, Yezhen Wang, Dongqi Han, and Dongsheng Li. Understanding and improving training-free loss-based diffusion guidance. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024.

Jay Shenoy, Miro Astore, Axel Levy, Frédéric Poitevin, Sonya M. Hanson, and Gordon Wetzstein. Modeling atomic conformational ensembles of proteins via test-time supervision of Boltz-2 on cryo-EM density maps. arXiv preprint arXiv:2605.09832, 2026.

David Silva-Sánchez, Alison M. Berezuk, Xing Zhu, Erik H. Thiede, Roy R. Lederman, and Pilar Cossio. Cryo-electron microscopy structural ensemble optimization using individual particles. *Journal of Chemical Theory and Computation*, 22(10):4922–4938, 2026. doi: 10.1021/acs.jctc.6c00053.

Bernard W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London, 1986. doi: 10.1201/9781315140919.

Raghav Singhal, Zachary Horvitz, Ryan Teehan, Mengye Ren, Zhou Yu, Kathleen McKeown, and Rajesh Ranganath. A general framework for inference-time scaling and steering of diffusion models. In *ICML 2025, PMLR 267*, pp. 55810–55827, 2025.

Bowen Song, Soo Min Kwon, Zecheng Zhang, Xinyu Hu, Qing Qu, and Liyue Shen. Solving inverse problems with latent diffusion models via hard data consistency. In *International Conference on Learning Representations (ICLR)*, 2024.

Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=9_gsMA8MRKQ.

Florence Tama, Osamu Miyashita, and Charles L. Brooks, III. Normal mode based flexible fitting of high-resolution structure into low-resolution experimental data from cryo-EM. *Journal of Structural Biology*, 147(3):315–326, 2004. doi: 10.1016/j.jsb.2004.03.002.

Daniel Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. Analysing the generalisation and reliability of steering vectors. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024.

Wai Shing Tang, David Silva-Sánchez, Julian Giraldo-Barreto, Bob Carpenter, Sonya M. Hanson, Alex H. Barnett, Erik H. Thiede, and Pilar Cossio. Ensemble reweighting using cryo-EM particle images. *The Journal of Physical Chemistry B*, 127(24):5410–5421, 2023. doi: 10.1021/acs.jpcb.3c01087.

Bogdan Toader, Fred J. Sigworth, and Roy R. Lederman. Methods for cryo-EM single particle reconstruction of macromolecules having continuous heterogeneity. *Journal of Molecular Biology*, 435(9):168020, 2023. doi: 10.1016/j.jmb.2023.168020.

Brian L. Trippe, Jason Yim, Doug Tischer, David Baker, Tamara Broderick, Regina Barzilay, and Tommi Jaakkola. Diffusion probabilistic modeling of protein backbones in 3D for the motif-scaffolding problem. In *ICLR 2023*, 2023.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. arXiv preprint arXiv:2308.10248, 2023. Earlier versions titled "Activation Addition: Steering Language Models Without Optimization".

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. doi: 10.1093/nar/gkad1084.

Gian Marco Visani, William Galvin, Michael Neal Pun, and Armita Nourmohammad. H-Packer: Holographic rotationally equivariant convolutional neural network for protein side-chain packing. In *Machine Learning in Computational Biology (MLCB)*, volume 240 of *Proceedings of Machine Learning Research*, pp. 230–249. PMLR, 2023.

Rémi Vuillemot and Sergei Grudinin. CryoARC: Atomic-resolution conformational landscapes of protein assemblies from cryo-EM single particles with evolutionary priors. *bioRxiv*, 2026. doi: 10.64898/2026.05.25.727696.

Rémi Vuillemot, Alex Mirzaei, Mohamad Harastani, Ilyes Hamitouche, Léo Fréchin, Bruno P. Klaholz, Osamu Miyashita, Florence Tama, Isabelle Rouiller, and Slavica Jonić. MDSPACE: Extracting continuous conformational landscapes from cryo-EM single particle datasets using 3D-to-2D flexible fitting based on molecular dynamics simulation. *Journal of Molecular Biology*, 435(9): 167951, 2023. doi: 10.1016/j.jmb.2023.167951.

Jimin Wang and Dennis C. Boisvert. Structural basis for GroEL-assisted protein folding from the crystal structure of $(\text{GroEL-KMgATP})_{14}$ at 2.0 Å resolution. *Journal of Molecular Biology*, 327 (4):843–855, 2003. doi: 10.1016/s0022-2836(03)00184-0.

Zi Wang, Alexander Ku, Jason Baldridge, Thomas L. Griffiths, and Been Kim. Gaussian process probes (GPP) for uncertainty-aware probing. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023.

Tomasz Włodarski, Julian O. Streit, Alkistis Mitropoulou, Lisa D. Cabrita, Michele Vendruscolo, and John Christodoulou. Bayesian reweighting of biomolecular structural ensembles using heterogeneous cryo-EM maps with the cryoENsemble method. *Scientific Reports*, 14(1):18149, 2024. doi: 10.1038/s41598-024-68468-7.

Luhuan Wu, Brian L. Trippe, Christian A. Naesseth, David M. Blei, and John P. Cunningham. Practical and asymptotically exact conditional sampling in diffusion models. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, pp. 31372–31403, 2023. doi: 10.52202/075280-1363.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. ReFT: Representation finetuning for language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Zhaohui Xu, Arthur L. Horwich, and Paul B. Sigler. The crystal structure of the asymmetric GroEL–GroES-$(\text{ADP})_7$ chaperonin complex. *Nature*, 388(6644):741–750, 1997. doi: 10.1038/41944.

Yang Yan, Shiqi Fan, Fajie Yuan, and Huaizong Shen. A comprehensive foundation model for cryo-EM image processing. *Nature Methods*, 23(1):88–95, 2026. doi: 10.1038/s41592-025-02916-8. Published online 27 November 2025; preprint bioRxiv 2024.11.04.621604; code github.com/westlake-repl/Cryo-IEF.

Haotian Ye, Haowei Lin, Jiaqi Han, Minkai Xu, Sheng Liu, Yitao Liang, Jianzhu Ma, James Zou, and Stefano Ermon. TFG: Unified training-free guidance for diffusion models. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024.

Shuxin Zheng, Jiyan He, Chang Liu, Yu Shi, Ziheng Lu, Weitao Feng, Fusong Ju, Jiaxi Wang, Jianwei Zhu, Yaosen Min, He Zhang, Shidi Tang, Hongxia Hao, Peiran Jin, Chi Chen, Frank Noé, Haiguang Liu, and Tie-Yan Liu. Predicting equilibrium distributions for molecular systems with deep learning. *Nature Machine Intelligence*, 6(5):558–567, 2024. doi: 10.1038/s42256-024-00837-3.

Ellen D. Zhong, Tristan Bepler, Bonnie Berger, and Joseph H. Davis. CryoDRGN: reconstruction of heterogeneous cryo-EM structures using neural networks. *Nature Methods*, 18(2):176–185, 2021. doi: 10.1038/s41592-020-01049-4.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. arXiv preprint arXiv:2310.01405, 2023.

A Study design

A.1 Sample sizes and pairing

Table S1 lists every reported experiment. Its counts are retained samples, or, for sequential Monte Carlo and selection, outputs that each keep one of 24 trajectories; otherwise the only attrition is the distance check on the unguided reference draw, which is not repaired (below).

Table S1: Every reported experiment with its sample sizes.

experiment	design	samples / particles
GroEL, two extreme targets	2 targets, $M = 128$	96 guided per target (4 seeds × 24); 300 unguided
GroEL, PDB structures	10 targets, $M = 128$	96 guided per target (4 seeds × 24); unguided: the $\lambda = 0$ runs (4 seeds × 24)
Lysenin	2 targets, $M = 128$	48 guided per target (2 seeds × 24); 25 unguided of 300
DppA, guidance	2 targets, $M = 128$	96 guided per target (4 seeds × 24); 281 unguided of 300
DppA, sequential Monte Carlo (§4.3)	2 targets, $M = 128$	32 populations of 24 trajectories per target, one output each
DppA, other schemes (App. C)	open target; three also closed	as in Table S2
GroEL, sequential Monte Carlo (App. C)	2 targets, $M = 128$	4 populations of 24 per target; 96 unguided at $\lambda = 0$
Guidance-scale sweeps (App. G)	one target per protein, $M = 128$	96 guided samples per value (144 on DppA)
Query imaging mismatch (App. F)	4 conditions (3 outside the dataset's distribution) × the 2 GroEL targets, $M = 128$	96 guided per condition and target (4 seeds × 24)
Code probes	639 validation conformations + 3 withheld conformations	128 simulated images each
Population inference	4 landscapes × 3 particle sets	4,000 particles each; 300 candidates, 2,900 conformations outside the library

Seeds and pairing. Each run yields 24 samples at one seed; four seeds give GroEL's 96 and DppA's 96, and two give lysenin's 48. The sampler draws x_T afresh for each batch, so samples within a seed are independent given the seed and n is not inflated by correlation. The initial noise is seeded as $\texttt{seed}+400+\lfloor 100\lambda \rfloor + 1000\cdot\texttt{batch}$, independently of the fit, with λ the guidance weight of App. B: 31.0 in every guided run, whatever its κ_x and κ_ω, and 0 in runs without guidance. Two conditions at matched $(\text{target}, \text{seed}, \text{batch}, \lambda)$ therefore draw one x_T, as the values of the position-scale sweep (App. G) do, and can be contrasted per sample with an exact two-sided McNemar test. Guided and unguided conditions are not paired, because λ enters the seed.

Unguided samples. Three sets of unguided samples are reported. The unguided reference draw is 300 structures per protein from the generator's own sampling entry point at a fixed seed (App. G), without the geometry repair. Its samples that pass the distance check supply the unguided medians of §4.2 and §4.3 and the core-RMSD threshold of the success criterion (App. F.1); all 300, failures included, are the reference of the guidance-scale rule (Eq. S8). All 300 pass on GroEL and 281 on DppA, but only 25 on lysenin, so lysenin's unguided medians (§4.3, Table S11) rest on 25 samples. The $\lambda = 0$ runs are the sampler of App. B with guidance switched off, followed by the geometry repair (App. G.5); they supply the unguided samples of Figure 2 and of the sequential Monte Carlo on GroEL (App. C). DppA's unguided pool is 240 such $\lambda = 0$ samples, from which the selection schemes of App. C draw. Because the reference draw and the pool are drawn differently, their rates of reaching DppA's open target, 9 of 281 and 12 of 240, are separate measurements, and neither set is a matched control for the guided runs.

B THE GUIDED SAMPLER

B.1 THE POTENTIAL IN WHITENED FORM

The implementation evaluates the gradient of Eq. 5 in whitened form. With $\Sigma_M(t) = S_h(t) + \Omega/M$ the covariance of Eq. 4, L_t its Cholesky factor, $J_w(t) = L_t^{-1} J(t)$ and $b(t) = L_t^{-1} \hat{g}_z$ for the pooled code $\hat{g}_z$ of the observed particles, the potential is

$$\ell_t(x_t) \;=\; -\frac{1}{2} \left\| \, b(t) - J_w(t) \big[h(x_t, t) - \overline{h}(t) \big] \, \right\|^2 . \tag{S1}$$

$\Sigma_M(t)$ enters only through a triangular solve; a covariance that is not positive definite makes the factorisation fail, and no repair is applied. Eq. S1 is Eq. 4 without its constant, which depends only on t and M and has no gradient in x_t, so its gradient is u_t of Eq. 5. The rotation channel adds $\kappa_\omega u_t$ with $\kappa_\omega = 1.00$, projected and clamped as below, and the position channel $\kappa_x P u_t$ (Eq. 6). In the implementation each scale is the product of a guidance weight λ, shared by the two channels, and a gain for that channel, set per protein. λ is 31.0 in every guided run and 0 in runs without guidance, so the gains set κ_x and κ_ω; apart from switching guidance off, λ enters only the random seed (App. A) and the rotation projection described below. The Hessian in h, $-J_w^\top J_w$, has rank at most 5 (for GroEL), so the particles constrain 5 combinations of h and the frozen generator determines the remaining 507.

B.2 STATE, SOLVER, AND SCHEDULE

BioEmu's state consists of Cα translations in nanometres and one rotation matrix per residue, diffused under a variance-preserving cosine schedule for positions and a variance-exploding IGSO(3) schedule for rotations. Sampling uses the generator's second-order solver on its grid of 50 times, uniformly spaced from $t = 0.99$ to $t = 0.001$ (49 iterations), with no injected noise, so the reverse process is the probability-flow ODE and a draw's only randomness is its initial state x_T. Each iteration evaluates the score at the predictor and the corrector time, and with guidance each evaluation is one forward and one backward pass through the frozen generator.

At each evaluation h is read from the forward pass that produces s_θ, the potential of Eq. S1 is formed, and its gradient is taken by automatic differentiation through the frozen network in the tangent space the sampler updates: translations are Euclidean, and rotations are right-multiplied by $\exp(\widehat{\delta\omega})$ with $\delta\omega$ initialised at zero, so $\partial/\partial\delta\omega$ is a vector in $\mathbb{R}^3$ per residue (an axis-angle map would begin with a vector norm, whose gradient at zero is undefined). The whitened objects J_w, b and $\overline{h}$ are tabulated at the 50 diffusion times and interpolated linearly in t, clamped outside the grid; the fitting grid is the sampler's, so interpolation acts only at the corrector time.

Behaviour at high noise. At high noise $S_h(t)$ rises to the code prior Λ of App. E (0.942 per direction at the noisiest step, the prior's value), the precision collapses, and the guidance term is negligible for the first several steps; it becomes active as the structure becomes informative, the precision spanning a factor of 2×10^5 across the schedule. This follows from the held-out residual without an imposed schedule.

Algorithm S1 Evidence-weighted guidance.

Offline, once. For each conformation c: simulate 128 images and embed them as $e_{c,v}$; forward-noise c and extract $h_{c,t}$ at 50 diffusion times.
Fit ϕ by Eq. 1; this defines the code ψ (App. E).
Fit $J(t)$ per diffusion time, and estimate $S_h(t)$ out of fold.
Measure Ω as the per-image covariance of the code across images of one conformation.

At sampling time. Encode the M observed particles as $e(I_m)$.
$\hat{g}_z \leftarrow$ the mean of $\psi(e(I_m))$
for each solver step at noise level t **do**
  $h \leftarrow h(x_t, t)$
  $\hat{g}_h \leftarrow J(t)\,(h - \bar{h}(t))$
  $W_M(t) \leftarrow \big(S_h(t) + \Omega/M\big)^{-1}$
  $u_t \leftarrow (\partial h/\partial x_t)^\top J(t)^\top W_M(t)(\hat{g}_z - \hat{g}_h)$ Eq. 5
  add $\kappa_x P u_t$ to the position and $\kappa_\omega u_t$ to the rotation score ($\kappa_\omega = 1.00$); take the step
end for

B.3 The position channel

The position update is Eq. 6, $s_{\text{used}} = s_\theta + \kappa_x P u_t$, with κ_x this channel's guidance-strength scalar (App. G) and P a smoothing operator along the residue sequence. Low-resolution refinement is routinely parameterised by a few low-frequency modes (Tama et al., 2004; Delarue & Dumas, 2004); P applies the same idea with the residue ordering as its only input.

P is the three-tap average $u_i \leftarrow (u_{i-1} + 2u_i + u_{i+1})/4$ iterated n times, with the reflecting end rule $u_0 \leftarrow (3u_0 + u_1)/4$, which makes it symmetric and doubly stochastic, so preconditioning by P is steepest ascent in the metric P^{-1} and the chain's cosine modes are its eigenvectors. Over the long-wavelength modes it passes, it is a Gaussian window in mode index with standard deviation k_0,

$$\text{gain}(k) = \cos^{2n}\!\left(\frac{\pi k}{2N_{\text{res}}}\right) \simeq \exp\!\left(-\frac{k^2}{2k_0^2}\right), \qquad k_0 = \frac{N_{\text{res}}}{\pi}\sqrt{\frac{2}{n}}, \tag{S2}$$

and k_0, in chain modes, is the quantity that transfers between proteins; the iteration count follows from it and the chain length N_{res} in the batch, $n = \text{round}\big(2(N_{\text{res}}/\pi k_0)^2\big)$. We use $k_0 = 6$ modes, which at $N_{\text{res}} = 547$ gives 1,684 iterations, a Gaussian of width 29 residues along the chain.

B.4 The rotation channel

The rotation gradient enters at $\kappa_\omega = 1.00$, uncalibrated on any protein, with no preconditioner and two per-structure operators following Sadat et al. (2025); both can only reduce the update and were introduced for sampling stability, following established approaches in the field. The first is a projection: with u_t^ω the rotation block of u_t and $\text{ref} = s_\theta + (\kappa_\omega/\lambda)\,u_t^\omega$ in the same channel, the score with the rotation term at its gain alone, as it would be at $\lambda = 1$, u_t^ω is replaced by $u_t^\omega - \langle u_t^\omega, \text{ref}\rangle\,\text{ref}/\|\text{ref}\|^2$, the inner products summed over the residues of one structure and the denominator floored at 10^{-20}. The second is a relative clamp: after κ_ω is applied, the rotation term is rescaled by $\min\big(1,\ 3\,\|s_\theta\|/\|\kappa_\omega u_t^\omega\|\big)$, in per-structure Frobenius norms.

Numerical safeguards. A non-finite potential or gradient sets that step's guidance to zero; completed samples pass through the geometry repair of App. G.

C Reweighting by sequential Monte Carlo

Gradient guidance adds $\kappa_x P u_t$ to a deterministic solver (App. B), so each trajectory ends where the guided drift from x_T takes it. On DppA that drift does not raise the rate at which samples reach the open state: 9 of 281 unguided samples reach it, and guidance at the scale of App. G reaches it in 0 of 96 samples (Table S2). The likelihood can instead reweight trajectories that the unguided sampler already produces. This appendix describes the scheme behind Figure 3c, compares it with the alternatives run on DppA, and states its limits.

Algorithm S2 Sequential Monte Carlo reweighting of one population.

Draw $x_T^{(j)}$ for $j = 1, \dots, N$; set $\log w_j \leftarrow 0$ and $g_j \leftarrow 0$.
for each solver step from t to t' **do**
  if $t \in [0.2, 0.8]$: forward-noise every $x^{(j)}$ from t to $t + c|t' - t|$
  take the unguided step ($\lambda = 0$) for every trajectory
  if t' is a check **then**
    $\ell_j \leftarrow \ell_{t'}(x^{(j)})$ Eq. 4, no gradient
    $\tau \leftarrow$ the largest value in $(0, 1]$ keeping the conditional ESS $\geq 0.5N$, else 0
    if $\tau > 0$: $\log w_j \leftarrow \log w_j + \tau\ell_j - g_j$; $g_j \leftarrow \tau\ell_j$
    if $\mathrm{ESS}(w) < 0.5N$: resample systematically, copying g_j; set $\log w_j \leftarrow 0$
  end if
end for
Repair every trajectory; $\log w_j \leftarrow \log w_j + \ell_{\mathrm{clean}}(x^{(j)}) - g_j$.
return one trajectory drawn with probability proportional to w_j.

C.1 Target law and algorithm

A population is $N = 24$ trajectories of the unguided sampler ($\lambda = 0$) from independent x_T, advanced together (Algorithm S2). The solver is made stochastic by noise injection: before every step that starts at $t \in [0.2, 0.8]$ the state is forward-noised from t to $t + c\,|\Delta t|$ with $c = 0.5$, so that trajectories copied by resampling separate again. At the 15 solver boundaries with $t \in [0.4, 0.7]$ each trajectory is scored by the emission log-likelihood ℓ_t of Eq. 4 at its current representation, one forward pass without gradient, and its twist is the tempered score $\tau\ell_t$. With g_j the twist last applied to trajectory j, a check adds $\tau\ell_t(x^{(j)}) - g_j$ to its log weight and sets $g_j = \tau\ell_t(x^{(j)})$; after the final repair (App. G.5) the log-likelihood ℓ_{clean} of the stored structure at the last diffusion time is added in place of the last twist. The intermediate terms telescope, as with the difference potential of Feynman–Kac steering (Singhal et al., 2025), so for any positive schedule of τ the final weights target

$$\pi(x_0) \;\propto\; \tilde{p}(x_0)\, L_{\mathrm{clean}}(x_0), \tag{S3}$$

the law $\tilde{p}$ of the noise-injected unguided sampler reweighted by the likelihood of the observed particle images; the twist shapes the path and the variance, not the target. At each check τ is the largest value in $(0, 1]$ that keeps the conditional effective sample size $N\big(\sum_j W_j u_j\big)^2 / \sum_j W_j u_j^2$, with W_j the normalised weights and $u_j = e^{\tau\ell_j - g_j}$, at or above $0.5N$; when no value does, the check leaves weights and twists unchanged. The population is resampled systematically whenever its effective sample size falls below $0.5N$, after which the weights restart from equal. The population's output is one trajectory drawn from its final weights.

C.2 What a population costs

A population costs one pass of the unguided sampler per trajectory plus one forward pass per trajectory and check; the checks take a few percent of the time. On one GPU a population of 24 takes 226 s, and the 32 populations per target of Figure 3c took 4.0 GPU-hours for both targets. Each population draws its initial noise exactly as the unguided pool of Table S2 draws consecutive batches, so the first populations start from the pool's x_T, and the two targets share their initial noise, injected noise and resampling uniforms. N, the window, the injected noise and the two thresholds were fixed before any population reported here was drawn and were not changed afterwards; the window and the choice of temperature were tuned on an earlier pilot population of unguided DppA trajectories that was never resampled.

C.3 The other schemes on DppA

Table S2 scores the unguided pool and seven schemes run on DppA, and Figure S1 draws three of them. Gradient guidance at the scale of App. G ($\kappa_x = 3.78$) reaches the open target on 0 of 96 samples, and at $\kappa_x = 1$ on 8 of 144, about the pool's rate of 12 in 240. The likelihood itself favours the open state: choosing among 24 samples of that pool by it, either the most likely one or one drawn in proportion to it, reaches the open target in 4 and 5 of 10 groups. The sequential Monte Carlo of Algorithm S2 reaches it in 11 of 32 populations. Two further rows separate its parts.

With the injected noise but no intermediate checks, so that each population is weighted by its final likelihood alone, 3 of 10 populations reach the open target; with guidance at $\kappa_x = 1$ as the proposal in place of the unguided sampler, 1 of 6 do. At equal cost, 24 trajectories per output, the sequential Monte Carlo is therefore not better than selection from the pool, and its resampling adds nothing measurable here. On the closed target, which no sample of the pool reaches, guidance succeeds in 2 of 96 samples, the sequential Monte Carlo in 1 of 32 populations, and selection from the pool never.

Table S2: The unguided pool and seven schemes run on DppA. An output is one sample in the per-sample rows and otherwise the one sample that a population, or a group of 24 unguided samples, yields; the trajectories column is the cost of one output. Hits count successful outputs (criterion in App. F.1); RMSD is the median core RMSD of the outputs in nm; — marks a target the scheme was not run on. The unguided pool is the sampler run at $\lambda = 0$ with the repair of App. G.5, the pool the selection rows draw from; it is not the unguided reference draw of the main text (App. A), whose medians are 0.53 and 0.20 nm.

				open, 1DPE:A		closed, 1DPP:A	
scheme	output	outputs	trajectories	hits	RMSD	hits	RMSD
unguided pool	sample	240	1	12	0.52	0	0.26
guidance, $\kappa_x = 3.78$	sample	96	1	0	0.47	2	0.12
guidance, $\kappa_x = 1$	sample	144	1	8	0.51	—	—
most likely of 24	group	10	24	4	0.31	0	0.08
drawn $\propto$ likelihood	group	10	24	5	0.35	0	0.10
injected noise, final weights only	population	10	24	3	0.22	—	—
sequential Monte Carlo	population	32	24	11	0.26	1	0.10
on the $\kappa_x = 1$ proposal	population	6	24	1	0.25	—	—

C.4 Relation to other samplers, and limits

The scheme is Feynman–Kac steering (Singhal et al., 2025) with its difference potential: the proposal is the frozen sampler with no guidance term, made stochastic by the injected noise, and the potentials are successive differences of the emission log-likelihood of the current state, so Eq. S3 is that method's tilt of the generator's distribution with the log-likelihood as the reward. The twisted diffusion sampler (Wu et al., 2023), which is asymptotically exact for a diffusion model's conditional distribution, also moves its proposal along the likelihood's gradient, which ours does not; ours is consistent, as N grows, only for Eq. S3. SMCDiff (Trippe et al., 2023) applies sequential Monte Carlo to protein backbones.

Resampling concentrates a population on few ancestors, so its trajectories are not independent; we therefore keep one output per population and count populations, not trajectories. The scheme reweights trajectories of the noise-injected unguided sampler and so rarely reaches a state that sampler seldom produces: on GroEL at $\lambda = 0$ it reaches neither target in any of 4 populations per target, and 0 of 96 unguided samples lie past the extended target along ξ, toward which gradient guidance does move the ensemble (§4.2). Reweighting by the likelihood, by this scheme or by selection among unguided samples, helps when the target is a rare but present state of the prior, as DppA's open state is; it is not a general replacement for guidance.

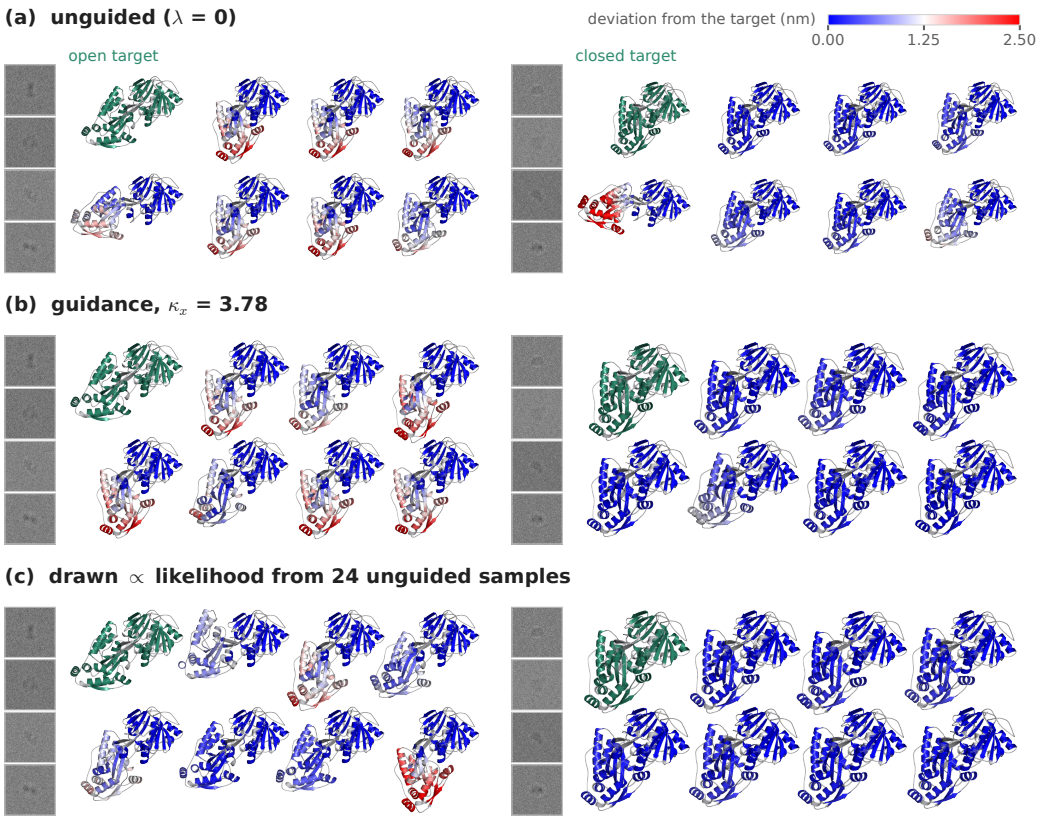

Figure S1: DppA under three schemes of Table S2, open target (1DPE:A) left and closed target (1DPP:A) right: **(a)** samples of the unguided pool, the same samples under both targets; **(b)** guided samples at $\kappa_x = 3.78$; **(c)** samples drawn from groups of 24 pool samples in proportion to the likelihood. Each block shows the target's particle images, the target (green) and 7 samples, with the camera, superposition and colouring of Figure 3c on a wider colour scale.

D ERROR DECOMPOSITION AND THE CODE DIMENSION

D.1 DECOMPOSITION OF THE TARGETING ERROR

A guided ensemble can stop short of its target because the images do not imply the target's ξ or because the sampler does not reach what they imply. Three quantities per target separate the two: the target's ξ, the ξ predicted from its observed particles, and the ξ the ensemble achieves (Table S3). The predicted ξ is read off the observed particles' pooled code by a fit that excludes the target conformation, so it is out of sample, and is quoted with its local residual standard deviation.

Table S3: The three quantities for the GroEL runs of the main text, in nm. Signs are taken in each target's required direction, so negative means short.

target	target ξ	predicted	predicted SD	achieved	predicted − target
extended	7.14	6.55	0.24	7.00	−0.587
compact	3.66	3.89	0.11	3.90	−0.231

On both targets the predicted ξ falls short, by 0.587 nm (extended) and 0.231 nm (compact), more than two local standard deviations in each case. The extended ensemble passes it, sitting 0.452 nm beyond its predicted ξ (1.88 local SD), and the compact one stops 0.011 nm short of it (0.10 SD), within its uncertainty. The gap to the target is dominated by the image side.

D.2 HOW MANY PARTICLES THE CODE NEEDS

Each GroEL target is observed through 128 particles. To measure how many the code needs, we average the codes of M of them, drawn at random without replacement, and predict ξ from the mean with the fit of Table S3. Because the prediction is linear in the mean code, adding particles leaves its expectation unchanged and only reduces its spread, which falls as $M^{-1/2}$ (Table S4); random subsets agree with this rate to within 2%. The spread drops below the prediction's local residual standard deviation, which does not decrease with M, at about 13 particles for the compact target and 36 for the extended one. A bootstrap over the particles gives 95% intervals of 11–16 and 28–42, and the dataset's images of the same two conformations give 14 and 37. Additional particles still sharpen W_M, but beyond these counts they change the predicted ξ by less than its fitting error, so the shortfall in Table S3 is not caused by the number of particles.

The extended target needs more particles because its particles scatter more: the covariance of its single-particle codes has 5.5 times the trace of Ω, compared with 0.7 for the compact target. Across the reference ensemble this scatter grows with ξ (Spearman 0.79), which a single pooled Ω (Eq. S5) does not capture. The full code converges in the same way. At the 95% level, no reference conformation other than the target remains consistent with the extended target's mean code from 64 particles on, whereas 11 remain consistent with the compact target's at 128.

Table S4: The mean code of M of a target's 128 particles. Spread: standard deviation of the ξ predicted from it over sets of M particles, in nm. Consistent: median number of the $3{,}200$ reference conformations, other than the target, whose code is consistent with it at the 95% level. The last row repeats the predicted SD of Table S3.

	compact target		extended target	
M	spread	consistent	spread	consistent
1	0.41	2,500	1.43	2,600
2	0.29	1,600	1.01	1,200
4	0.20	750	0.72	160
8	0.14	280	0.51	26
16	0.10	100	0.36	6
32	0.07	49	0.25	1
64	0.05	23	0.18	0
128	0.04	11	0.13	0
predicted SD	0.11		0.24	

D.3 THE CODE DIMENSION

The code dimension r, the rank of the reduced-rank regression, is fixed before the image map is trained and without reference to any target, coordinate or sampled draw. A separate regression can fix it because, for a given ϕ, the best decoder in Eq. 1 is the least-squares regression of h on $\phi(e)$, whose rank constraint keeps the r leading directions of the covariance of $\mathbb{E}[h \mid e]$; how many such directions the embedding supports can therefore be measured by a reduced-rank ridge regression of the same target on the embedding itself, in closed form. That regression has one row per conformation, its transformed embedding (App. E) averaged over its 128 simulated images, as the sampler averages its $M = 128$ particles, and the targets are excluded. On each of twenty replicate $80/20$ splits of the conformations, its ridge strength and rank $k \in \{1, \dots, 64\}$ are chosen jointly by an inner five-fold cross-validation on the training part, and the held-out part is evaluated once; k_{cv} is the rank with the best inner score and $k_{1\mathrm{se}}$ the smallest within one standard error of it (Table S5). We use k_{cv}, the rank with the smallest held-out error, which is what Eq. 1 is fitted to minimise; $k_{1\mathrm{se}}$, the usual parsimony rule, is listed for comparison and would give a smaller code on every protein.

Table S5: Offline selection of the code dimension over twenty replicate splits per protein; the rule is k_{cv}, with the number of replicates returning it. On GroEL the selection is unanimous and on DppA nearly so; on lysenin it is modal, the held-out R^2 being a plateau from $k = 6$ upward.

protein	k_{cv} (replicates)	$k_{1\mathrm{se}}$	used
GroEL	5 (20/20)	4	5
lysenin	7 (8/20)	3	7
DppA	6 (18/20)	3	6

E THE SHARED CODE AND ITS TWO COVARIANCES

E.1 FROM AN EMBEDDING TO A CODE

The image map is an elementwise $\arccos$ of the unit-norm encoder embedding followed by one affine map, $768 \to r$. The transformed embedding is standardised per coordinate on the fitting conformations, with the scale floored relative to its mean so that an inactive encoder dimension is not amplified; the stored normaliser applies the same two steps at sampling time.

The structural target h_c^* is conformation c's representation in the generator (App. G), averaged over the 50 diffusion times after each time is centred on its dataset mean.

Equation 1 is solved in two stages. The first projects the target onto the leading 64 principal directions of the per-time-centred structural scatter, which hold about 93.5% of its variance, so the decoder maps the code into that 64-dimensional basis. The second trains the image map and the decoder together on the projected target by AdamW (Loshchilov & Hutter, 2019) for 300 epochs in minibatches of 64 conformations, keeping the weights of the best validation epoch; the loss, not the order of the principal directions, decides which r-dimensional subspace of that basis the code spans.

Coordinates of the code. After fitting, the code is whitened so that the conformation-mean codes have zero mean and identity covariance, rotated so that $A^\top A$ is diagonal (whitening alone leaves any rotation free), ordered by descending decoder energy and sign-fixed on each column's largest entry. ψ is the code in these coordinates, the form the main text and App. B use.

Storage form. The implementation stores each image's code standardised per coordinate, $z(e) = C\psi(e) + d$, with $C = \mathrm{diag}(1/\sigma_k)$ and $d = -\mu_k/\sigma_k$ for the per-image code's mean μ_k and standard deviation σ_k; the offset is numerically zero on the reported fit. This is a storage convention, and the likelihood does not depend on it: C is diagonal, so the Cholesky factor of $C\Sigma_M(t)C^\top$ is CL_t, and the $J_w(t)$ and $b(t)$ of Eq. S1 are the same whether they are formed from z or from ψ. C and d are not fitted: the fit verifies $g_c = V^{-1}\sum_v \big(z(e_{c,v})\sigma + \mu\big)$ over all conformations in the paired dataset. The variance share of App. F is the one number reported in the stored form.

E.2 THE CODE REGRESSION $J(t)$

At each of the 50 diffusion times, $J(t)$ is a ridge regression of the reference codes on the centred representation,

$$J_k(t)^\top = \left(X_t^\top X_t + \alpha_{t,k}\,\mathbb{I}\right)^{-1} X_t^\top G_{\cdot k}, \qquad k = 1, \ldots, r, \tag{S4}$$

for the kth row $J_k(t)$ of $J(t)$ and the kth column $G_{\cdot k}$ of G, with one row per conformation, $h_{c,t} - \overline{h}(t)$ in X_t and g_c in G, one penalty per time and per code coordinate, the intercept absorbed by the centring, and the data term summed over conformations, not averaged. Each coordinate's penalty minimises that coordinate's held-out error at that time on a grid of 25 values from 10^{-2} to 10^{6}; the selected values run from 5×10^{2}–10^{6} at the noisiest time to 2×10^{-2}–2×10^{-1} at the cleanest, so the selection switches the regression off at high noise. A penalty shared across coordinates would let the noisier directions of a larger code shrink the informative rows (§3.2), which is why it is per coordinate; the dimension is set by the rule of App. D. The selection is made on the first of five folds and reused for the rest, and the folds are over conformations of the regression stage only.

E.3 THE TWO COVARIANCES

Image scatter. Ω is the pooled within-conformation covariance of the image codes about their conformation mean,

$$\Omega = \frac{1}{n_c\,(V-1)} \sum_{c} \sum_{v=1}^{V} \left[\psi(e_{c,v}) - g_c\right] \left[\psi(e_{c,v}) - g_c\right]^\top, \tag{S5}$$

over the n_c fitting conformations and all $V = 128$ images of each: a full, unshrunk $r \times r$ matrix, unbiased for the per-image covariance by the divisor $V - 1$. A reference code g_c averages V images and so carries averaging noise of covariance Ω/V; the whitening above gives the g_c identity covariance, so the conformations' underlying codes have covariance $\Lambda = \mathbb{I} - \Omega/V$, the code prior. On the reported fit $\operatorname{tr}\Omega = 37.31$ against $\operatorname{tr}\Lambda = 4.71$, and the eigenvalues of Ω span a factor of 14.0, so one image constrains the 5 code directions very unequally, and W_M carries that anisotropy.

Structural residuals and the cap. $S_h(t)$ is built at each diffusion time in five steps.

1. The out-of-fold residual $\varepsilon_c(t) = g_c - \hat{g}_{h,c}^{(-f(c))}(t)$ is formed over five conformation-grouped folds, so that all 50 times of a conformation fall in one fold.
2. Its covariance is taken about the mean with the unbiased divisor; the degrees of freedom are not adjusted for the ridge's effective parameters.
3. The image-averaging term Ω/V is subtracted, so that the result estimates the error against the underlying code rather than against its V-image average g_c.
4. The eigenvalues are floored at 10^{-8}.
5. The matrix is capped at the code prior Λ in the Loewner order,

$$S_h \;\leftarrow\; \Lambda^{1/2}\,\Pi_{\le 1}\!\left(\Lambda^{-1/2} S_h \Lambda^{-1/2}\right)\Lambda^{1/2}, \tag{S6}$$

   where $\Pi_{\le 1}$ clips eigenvalues at one, so that $S_h \preceq \Lambda$ holds as a matrix inequality, verified in the whitened basis before the fit is stored. The cap reflects that at the noisiest times a structure carries no information about its code, so its prediction error cannot exceed the prior's spread.

On the reported fit the cap is active on 4 of the 5 code directions at the noisiest time, where S_h has risen to the prior ($\operatorname{tr} S_h = 0.942$ per direction against the prior's 0.942), and on none from partway through the schedule onward.

Both covariances are properties of the paired dataset rather than of the observed particles, so the model assumes the dataset covers the particles' imaging conditions; App. F measures the guided runs when it does not. The only conformations withheld from both fits, from Ω, from S_h and from the penalty search are the 3 at which the method is evaluated, conformations 293, 363, 689 on GroEL; the fits' splits are in App. G.

F INFORMATION RETAINED BY THE CODE

§4.2 decodes each simulated image's imaging parameters and ξ from two representations of the same 128 images of each of 3,200 conformations (Table S6); the protocol below bounds what the comparison can show.

Table S6: Held-out R^2 for decoding each quantity from the fitted code and from the embedding it is built from, on the 639 validation conformations and the 3 withheld conformations. The ceiling is the best 5-dimensional linear summary of that embedding for that quantity, so the comparison is not a statement about dimensionality.

decoded from	fitted code ($r = 5$)	5-dimensional ceiling	full embedding
defocus	0.008	0.697	0.705
pixel size	0.028	0.159	0.164
viewing direction	0.000	−0.002	0.000
ξ	0.473	0.475	0.487

Splits and nulls. These are linear probes on frozen representations (Alain & Bengio, 2017): what a probe decodes bounds what the representation holds, not what the model uses. Each probe is trained on the conformations the image map was fitted on and scored on the 639 validation conformations and the 3 withheld conformations, with the ridge strength chosen on an inner split of the training conformations for every (representation, quantity) pair separately, so the ceiling never suffers from a shared choice. Held-out R^2 can be negative, and a small negative value means no signal; a permutation null per target is computed alongside and released with the results. The viewing direction is decoded as its unit vector $\hat{n}$ and the in-plane rotation angle γ as $(\cos\gamma, \sin\gamma)$, so no angle wraps.

Pose. No probe tested, linear or nonlinear, decodes the viewing direction from the raw embedding; a 2048-dimensional random Fourier feature expansion (Rahimi & Recht, 2007) does no better than the linear probe. This is a property of the dataset as much as of the encoder. The conformations are stored in arbitrary orientations (App. F.1), so one viewing direction in the simulator's frame shows each conformation from a different side, and no single direction of the embedding can follow it across conformations. Pose therefore appears as scatter among the images of one conformation, which Ω measures together with the other imaging variables and the noise. All of that scatter, pose included, is 3.2% of the variance of the 128-image mean code across conformations when each code coordinate is scaled by its per-image standard deviation, the stored form of App. E. The likelihood measures it in ψ, where the conformation means have unit variance in every direction, as $\operatorname{tr}\Omega/128$ against their trace of 5, which is 5.8%; the figure is larger because ψ does not scale down the directions in which images scatter most. The comparison shows selective suppression of the nuisance tested, not removal of every imaging variable: pixel size is partly retained, at 0.028 against a ceiling of 0.159. The probes read one image at a time, so they cannot show what a query stack with restricted orientations does to the pooled code; the next subsection measures that.

F.1 QUERY IMAGES OUTSIDE THE DATASET'S IMAGING DISTRIBUTION

Three variants of the GroEL query stacks of §4.2 re-render the 128 particles of the same two targets with one imaging parameter drawn outside the dataset's distribution and everything else unchanged: the tilt θ of the viewing direction (its angle to the stored frame's z-axis) drawn uniformly from $[80°, 100°]$ (side views) or $[0°, 30°]$ (top views) in place of the uniform distribution on $SO(3)$, or the defocus drawn from 2.5–3.5 μm against the dataset's 0.8–2.4 μm. The dataset, the fitted maps and covariances, M and the sampler are those of the main text, but all four conditions, the uniform one included, were run with one scale on both channels, $\kappa_x = \kappa_\omega = 4.04$, the setting in use before the two channels were given separate scales, with 96 samples per condition and target (4 seeds × 24). The query stack does not depend on the sampler seed, so its decoded ξ is one number per condition (Table S7).

Every quantity in Table S7 is ordered by the view geometry on both targets, top views then uniform then side views. On the extended target the decoded ξ runs from 6.32 to 6.88 nm against the target's

Table S7: The two GroEL targets under four query-image conditions, in nm. Decoded: ξ read off the pooled query code (App. D). Achieved: the mean over seeds of the per-seed median ξ with its t_3 half-width. Core RMSD: the same for the per-seed median, with its interval. The targets' values of ξ are 7.14 (extended) and 3.66 (compact); $\mathrm{ss}_{\mathrm{core}}$ is 0.89–0.91 on the extended target and 0.95 on the compact one under every condition, and the distance check passes on every sample. The uniform rows use the main text's query distribution but, like the others, the single scale $\kappa_x = \kappa_\omega = 4.04$, so they differ from §4.2.

condition	decoded	achieved	core RMSD
extended target			
top views, $\theta \in [0^\circ, 30^\circ]$	6.323	6.387 ± 0.018	1.240 [1.189, 1.290]
uniform	6.549	6.806 ± 0.020	1.211 [1.159, 1.262]
far defocus, 2.5–3.5 μm	6.766	7.214 ± 0.038	1.195 [1.133, 1.257]
side views, $\theta \in [80^\circ, 100^\circ]$	6.881	7.286 ± 0.044	1.198 [1.139, 1.257]
compact target			
top views, $\theta \in [0^\circ, 30^\circ]$	3.958	3.981 ± 0.042	0.466 [0.433, 0.498]
uniform	3.886	3.898 ± 0.036	0.440 [0.408, 0.472]
far defocus, 2.5–3.5 μm	3.990	3.948 ± 0.040	0.445 [0.413, 0.477]
side views, $\theta \in [80^\circ, 100^\circ]$	3.861	3.888 ± 0.033	0.435 [0.401, 0.468]

7.14 nm, the achieved ξ follows it from 6.39 to 7.29 nm, and the success rate rises from 0.000 under top and uniform views to 0.854 under side views (per-seed rates 0.833–0.875). A sample succeeds when its ξ lies at or beyond the target's, seen from the unguided median; its core RMSD to the target is below the tenth percentile of the unguided reference draw's (App. A); its core secondary structure is at or above the floor of App. H; and it passes the distance check. Core RMSD does not separate the conditions on the extended target; on the compact target it is worse under top views (0.466 against 0.440 nm, the intervals barely overlapping), while ξ stays at the edge of the reference ensemble under every condition.

The direction of the effect was predicted before the top-view runs were sampled: ξ is a length along the apical–equatorial axis, and a band of tilts changes how much of that axis projects into the image plane. The conformations are stored in arbitrary orientations, the two targets' axes lying 47° (extended) and 37° (compact) from the stored z-axis, and the mean projected fraction of the axis is 0.73, 0.79 and 0.85 under top, uniform and side views for the extended target and 0.62, 0.79 and 0.90 for the compact one; a view distribution that projects more of the axis carries more information about ξ.

Because the two targets lie on opposite sides of the reference ensemble along ξ, the runs can tell two kinds of change apart. A change in how much the images say about ξ moves both decodes toward their targets' values or both away from them; a shift of the code moves both decodes the same way along ξ, and so toward one target and away from the other. The tilt bands are of the first kind: top views move both decodes away from their targets and side views move both toward them. Far defocus is of the second: it moves both decodes up along ξ, by $+0.22$ nm on the extended target (toward its value) and $+0.10$ nm on the compact one (away from it). This offset raises the extended target's success rate to 0.688 (per-seed rates 0.583–0.750) and leaves the compact target's core RMSD unchanged (0.445 against 0.440 nm).

These runs qualify the requirement of §3.3 that the observed images come from imaging conditions the dataset covers: an orientation distribution outside the dataset's changes how much the stack says about ξ, monotonically in the tilt and in the predicted direction, and a defocus outside its range displaces the decode along ξ without changing the recovered structure.

G Inputs, data splits, and settings

G.1 Proteins and reference ensembles

Each protein has its own reference ensemble, code and covariances; only the procedure is shared (Table S8). Conformations are sampled from BioEmu (Lewis et al., 2025) conditioned on sequence alone, through evoformer embeddings from an automatically retrieved multiple-sequence alignment,

at the generator's denoiser settings. Samples with a broken Cα trace are rejected, side chains are rebuilt by H-Packer (Visani et al., 2023), and each structure is centred on its heavy atoms, because the simulator rotates about the origin.

Table S8: The three systems and their data splits, in conformations; every dataset has 128 simulated images and 50 diffusion times per conformation. The GroEL and lysenin steering targets are withheld from every fit, which records the indices it used, and the sampler checks that its operator saw no target. DppA's targets are PDB chains outside its dataset; its fit also withholds dataset conformations reserved as targets that this work does not use.

system	UniProt	residues	dataset	withheld	for fitting	ϕ train	ϕ val.
GroEL	P0A6F5	547	3,200	3	3,197	2,558	639
lysenin	O18423	309	3,214	6	3,208	2,566	642
DppA	P23847	507	3,240	3	3,237	2,590	647

Sequences and the withheld indices are in the released split manifest, with lysenin's two folds in the PDB (3ZXG:B (Colibus et al., 2012) / 5EC5:P (Podobnik et al., 2016)) and DppA's two targets, the unbound open chain 1DPE:A (Nickitenko et al., 1995) and the closed chain 1DPP:A bound to Gly-Leu (Dunten & Mowbray, 1995); GroEL's ξ is defined by residue windows and needs none. GroEL is one subunit of the tetradecamer, the lysenin construct carries its expression tag, and DppA is the mature chain without its signal peptide. Each system's unguided reference draw is a separate draw of 300 structures at a fixed seed (App. A).

The validation split is a fixed fraction of a seeded permutation over the whole dataset, filtered to the conformations available for fitting, so the withheld targets fall in neither split and the split does not depend on how many are withheld. The code regression and its covariance use a separate five-fold split of the same conformations, grouped by conformation and drawn independently of the image map's split, so the two overlap and neither is a held-out test of the other (App. E). The population experiment of App. I divides the whole GroEL dataset at random, independently of these splits, into its candidate library and the conformations its particles come from, so both overlap the conformations the code was fitted on.

G.2 Image simulation and encoding

Images are simulated with cryoJAX (O'Brien et al., 2026): each heavy atom is an isotropic Gaussian of unit amplitude and 1 Å^2 variance, a uniform scattering model in place of tabulated factors that makes the whole-box signal-to-noise ratio below well defined; the projection integrates the potential over each pixel, the contrast transfer function is applied with astigmatism, and Gaussian noise is added after it in real space with the zero-frequency component set to zero. There is no envelope, phase shift, detector or particle translation: the structure sits at the origin, and only its orientation, the optics and the noise are redrawn, independently per image from a range (Table S9).

Table S9: Imaging. Ranges are uniform and independent per image. Pixel size and noise variance are per system; everything else is shared.

	GroEL	lysenin	DppA
image size (px)	224		
orientation	uniform on $SO(3)$		
translation	fixed at zero		
defocus (μm)	0.8–2.4		
astigmatism (Å)	0–500, angle uniform on $[0, 360)^\circ$		
voltage (kV)	260–340		
spherical aberration (mm)	2.0–3.4		
amplitude contrast	0.08–0.12		
pixel size (Å)	1.20–1.30	0.998–1.081	0.915–0.992
box (Å) / longest chain	280/161	233/134	214/119
$\mathrm{Var}_{\mathrm{box}}$	0.012983	0.012401	0.021423
noise variance	1.82–3.38	1.74–3.22	3.00–5.57
$\mathrm{SNR}_{\mathrm{box}}$	0.005		

Signal-to-noise convention. Following the published heterogeneity benchmark, the whole box image is the signal:

$$\mathrm{SNR}_{\mathrm{box}} = \frac{\mathrm{Var}\big(\text{noiseless projection over the whole box}\big)}{\text{noise variance}}, \qquad \text{noise variance} = \frac{\mathrm{Var}_{\mathrm{box}}}{\mathrm{SNR}_{\mathrm{box}}}. \quad \text{(S7)}$$

The simulator works in arbitrary signal units, so $\mathrm{Var}_{\mathrm{box}}$ is measured per system as the median over eight conformations spanning the library, forty noiseless images each over the full optics sweep, and the per-image noise variance is drawn uniformly within $\pm 30\%$ of the centre Eq. S7 gives, the same relative spread in every dataset. The label holds only for the geometry it was measured on, since box, pixel size, defocus range, amplitude contrast and voltage all enter $\mathrm{Var}_{\mathrm{box}}$.

The operating noise level. Every reported dataset is simulated at $\mathrm{SNR}_{\mathrm{box}} = 0.005$, twice as noisy as the benchmark's standard setting. The whole-box convention understates the ratio on the particle: ninety percent of the noiseless signal power lies within 0.40 (extended target) and 0.31 (compact target) of the box diameter, so the ratio on the particle is $6.9\times$ and $11.8\times$ the quoted value; the box is $1.74\times$ the longest conformation ($1.79\times$ on DppA), as the contrast-transfer delocalisation at $2.4\,\mu$m requires. The extended target is the larger particle and carries 12.3% more whole-box signal power at any shared noise variance, so the two targets cannot be matched in signal-to-noise ratio at one setting.

Encoding. Each image is min–max normalised to $[0, 1]$, standardised with the encoder's published constants, and passed to the frozen Cryo-IEF (Yan et al., 2026) ViT-B/14 at its native 224 px with no circular mask; the embedding is the class token after the final layer normalisation, ℓ_2-normalised to $\mathbb{R}^{768}$, the classification head discarded. Images are encoded and stored in half precision in batches of 128, at a relative error of about 3×10^{-4} against full precision.

Seeding. The simulation draws one task seed per conformation and produces image i of conformation f from task seed$_f + i$, so two conformations whose task seeds differed by less than the number of images would share images. The minimum gap across every dataset is 137, so at 128 images none is shared. The observed particles are simulated from fixed seeds that are also at least 137 apart, so M can reach 137 before two particle stacks share an image; $M = 128$ stays within that margin.

G.3 Extracting the generator's representation

The structural representation is the residue-stream output of the first of the generator's eight structure-module layers, after the attention residual and before the feed-forward block, mean-pooled over residues to $\mathbb{R}^{512}$. The extraction replaces that layer's forward pass with an equivalent differentiable reimplementation and adds no forward pass. The four sites we compared have canonical correlations of 0.94–0.99 with one another, so a second site would add little.

Conformations are forward-noised on the generator's schedules at the sampler's 50 times (App. B), with one noise realisation per (conformation, time) drawn independently across times, so the time dependence of h cannot be read from differences between times; a variant with noise shared across times performed worse. The noising and the extraction depend only on the conformation, the time and the seed, so the representation half of the paired dataset is computed once per system and reused at every noise level.

G.4 Settings

Calibration of the guidance scale. The position scale κ_x sets how hard the particles pull on the sampler, and with it a trade-off between exploitation and exploration. A stronger pull carries the ensemble further toward the conformation the particles support; a weaker one keeps more of the spread the prior would produce on its own. Structure preservation limits how far exploitation can go. Up to some scale, guidance moves the ensemble as a whole, and its internal spread stays close to that under unit-weight guidance. Beyond that scale, guidance begins to reshape the ensemble itself. It either contracts the ensemble toward a point estimate, which discards the uncertainty the posterior should carry, or inflates the differences between its members' stable cores. We place κ_x where this reshaping first appears, using a rule that reads the guided ensemble's own spread and no target structure, coordinate or success rate.

Table S10: *Measured*: from data no part of this work chose. *Fitted*: by optimising a stated objective on the paired dataset. *Derived*: from other quantities by a rule fixed in advance. *Chosen*: without reference to the reported results. *Rule*: from a sweep, by a criterion with a declared tolerance that reads no target structure, coordinate or success rate. *Tuned*: from a sweep or a benchmark, and marked as such wherever it appears.

setting	value	how fixed	set on / where
code dimension r	5, 7, 6	derived	GroEL, lysenin, DppA; held-out selection for Eq. 1 (App. D)
image map ϕ	Eq. 1	fitted	2,558 conformations
structural pre-basis	64 directions	chosen	App. E
storage form C, d	from the dataset's code	derived	verified to 4.8×10^{-7}
$J(t)$	Eq. S4, one per t	fitted	3,197 conformations
$J(t)$ ridge penalty	one per t and per coordinate	fitted	a conformation-disjoint split
guidance scale κ_x	3.78, 3.46, 3.78	rule	GroEL, lysenin, DppA; variance band on one reported target, $F = 2$ set after pilot samples
$S_h(t)$	Eq. S6	measured	out-of-fold residual, five folds
Ω	Eq. S5	measured	128 images per conformation
$W_M(t)$	Eq. 4	derived	from $S_h(t)$, Ω, M
dataset noise level	0.005	chosen	the published benchmark
M	128	chosen	the seed margin above
preconditioner bandwidth	6 chain modes	tuned	sampled outcomes of GroEL compact/extended only, then transferred to other systems
rotation projection, clamp	orthogonal, threefold	tuned	Sadat et al. (2025); sampled outcomes of GroEL compact/extended only, then transferred to other systems
rotation scale κ_ω	1.00, every protein	derived	the likelihood itself
solver steps	50	chosen	the generator's own
Monte Carlo population N	24	chosen	App. C
twist window	$t \in [0.4, 0.7]$	tuned	one unguided DppA pilot population, before the reported ones
noise injection	$c = 0.5$ on $t \in [0.2, 0.8]$	chosen	App. C
selection group G	24	chosen	equal to N (App. C)
tempering, resampling	conditional ESS $0.5N$, ESS $0.5N$	chosen	App. C
repair constraint set	three distance bounds	chosen	below
candidate library	300 structures, 128 images	chosen	§4.4
kernel degrees of freedom	5	tuned	App. I
kernel widths b_i	one per candidate	measured	that candidate's images
kernel scale s	0.84 (library)	fitted	per particle source, then held
resolution limit H	0.412 nm	derived	the median-scatter rule
weight basis order	5	derived	the reported range and H
reporting width Δ	candidate spacing	derived	App. I
ρ	0.01 at 4,000 particles	tuned	App. I

The rule runs a sweep of κ_x upward from 1 on one reported target, guiding with that target's 128 particles and drawing 96 samples per value (144 on DppA). At each value it compares the spread of the guided samples with that of unguided sampling,

$$q(\kappa_x) = \frac{\tilde{d}(\kappa_x)^2}{\mathrm{median}_{\mathcal{S}}\, \tilde{d}_{\mathcal{S}}^2}, \tag{S8}$$

where $\tilde{d}(\kappa_x)$ is the median pairwise core RMSD of the guided samples, each pair superposed on the stable core. The denominator $\tilde{d}_{\mathcal{S}}$ is the same quantity for a subset $\mathcal{S}$ of the 300 unguided draws of equal size, those failing the distance check included, with the median taken over such subsets. Squaring makes q a ratio of variance-like spreads. The reference is $a = q(1)$: at $\kappa_x = 1$ the position term is the likelihood gradient at unit weight, smoothed by P, the weight the likelihood itself implies. Walking up the sweep, the selected κ_x is the first value at which q leaves the band $[a/F,\, aF]$, interpolated log-linearly between neighbouring sweep values. An exit below the band marks the onset of contraction, and an exit above it the onset of inflation.

The tolerance is set at $F = 2$. It has to exceed the one-sided 95% noise floor of q at the sweep's sample size, corrected for the finite unguided reference, so that sampling noise alone cannot carry q out of the band. That floor is 1.31 on GroEL, 1.20 on lysenin and 1.92 on DppA.

On the sweeps the rule returns 3.78 on GroEL and 3.46 on lysenin, where q exits above the band, and 3.78 on DppA, where it exits below. The inflation is in pairwise core RMSD; along ξ the GroEL ensemble still narrows at this scale (§4.2). Each value is read on a single target: the extended one on GroEL and lysenin, the open one on DppA. DppA's value is unstable under half-sampling (5–95%: 2.68–15.95) because its q leaves the band only just before returning into it at the top of the sweep.

The scale is set per protein because the whitened operator of Eq. S1 is built from each fit's J, S_h and Ω. A shared κ_x would therefore apply guidance of a different magnitude to each protein. Everything else is shared across proteins: M, the preconditioner, the rotation operators, the solver, and the rotation scale $\kappa_\omega = 1.00$ (App. B). DppA's sequential Monte Carlo samples (App. C) use no gradient and therefore no κ_x.

G.5 Geometry repair

Completed samples pass through two operations. The first is a restrained refinement: a few hundred gradient steps on the frame translations and rotations under a weighted penalty toward ideal consecutive Cα–Cα and peptide C–N separations and away from non-adjacent contacts, with a harmonic restraint toward the input structure. A finite soft penalty leaves a residual violation, so the second operation projects onto the set the validity check tests (consecutive Cα–Cα at most 0.45 nm, peptide C–N at most 0.20 nm, and at least 0.10 nm between atoms of residues more than two apart) by iterated minimum-norm projection in the SHAKE family (Ryckaert et al., 1977) and, more generally, convex feasibility (Bauschke & Borwein, 1996). Constraints are one-sided, so a satisfied constraint is left untouched. The projection aims four percent inside each bound rather than at it, so that a repaired structure does not land exactly on a threshold that the check evaluates in single precision after converting units; violations are still counted against the bounds themselves.

On the two GroEL targets the projection moves the structure by a median of 0.0000 and 0.0000 nm, a ninetieth percentile of 0.0106 and 0.0000 nm, and a maximum of 0.0703 and 0.0289 nm, recorded for every sample. Both operations are applied to the guided runs and the $\lambda = 0$ runs; the unguided reference draw passes through neither (App. A), so its pass rate is not compared with theirs.

G.6 Cost and released materials

Fitting both stages from a simulated dataset takes a few GPU-hours on one device, most of it the image simulation, done once per dataset; a guided batch costs 98 forward and 98 backward passes through the generator (App. B).

The supplementary material holds the source code for the method, including the sequential Monte Carlo reweighting and the likelihood selection of App. C; the experiment drivers; the configurations of the main runs; the system descriptors with their reference structures and the PDB chains used as targets; each system's fitted maps and covariances; the candidate library with the stored landscape fits; and the producers of the body's figures and numbers and of App. C's table and figure, which recompute them from the measured artefacts. Four are not reproducible from the archive alone: Figure 2's per-chain data are included but their scoring step is not; Figure 4a needs the molecular-dynamics ensemble, which is not redistributed; the sweeps behind κ_x are not re-run, the rule instead recomputing every κ_x, noise floor and half-sample quantile from their per-value statistics; and the query-mismatch runs of App. F.1 need rendering options and configurations the archive does not include. Figures 1, 2 and 4b–e are redrawn from inputs the archive carries. The remaining measured

artefacts (the conformation libraries and rendered particle images, the split manifest with conformation identities, fold assignments, withheld targets, seed streams and file hashes, the final coordinates and weights, and the tables behind the other figures) do not fit in a submission archive and are released with the paper; regenerating them requires re-running the stages on a GPU, on which guided draws are not bit-reproducible.

H STRUCTURAL COORDINATES AND STABLE CORES

H.1 THE STABLE CORE

Every quantity in this paper is measured over a *stable core*, defined once per protein from its reference ensemble and fixed: a residue belongs to the core if its simplified DSSP assignment (Kabsch & Sander, 1983) is helix or strand in at least half the conformations of that ensemble, with no run-length, smoothing or contiguity condition. No target or guided sample enters the mask, which is stored with the hash of the conformation file that produced it, because rebuilding it on a different ensemble would change every number measured against it. The criterion is local order rather than immobility because GroEL's apical domain is expected to move: a fluctuation criterion after a global superposition would discard the residues whose displacement the experiment is about while keeping a disordered loop that fluctuates in place. The cores hold 380 of 547 GroEL residues, 141 of 309 on lysenin and 274 of 507 on DppA; the ranges are in the released manifest.

H.2 THE MEASURED QUANTITIES

All coordinates are Cα positions in nanometres and all averages over atoms are unweighted (centroids).

GroEL. ξ is the distance between the centroids of the apical and equatorial residue sets, each intersected with the stable core:

$$\xi(x) = \left\| \operatorname{mean}_{i \in \mathcal{A} \cap \mathcal{C}} x_i \; - \; \operatorname{mean}_{i \in \mathcal{E} \cap \mathcal{C}} x_i \right\|, \tag{S9}$$

with $\mathcal{A} = [190, 373)$ and $\mathcal{E} = [0, 133) \cup [410, 525)$ in 0-indexed residue numbering (E. coli GroEL); the flexible C-terminal Gly–Met tail $[525, 547)$ is excluded. The core intersection leaves 130 of 183 apical and 180 of 248 equatorial residues. Every threshold and reported value uses the core-restricted ξ; the full-chain version differs from it by up to 0.16 nm and is not used.

Lysenin and DppA. ξ is the radius of gyration of the stable core, the root-mean-square distance of the core Cα positions from their centroid, unweighted.

Accuracy. Core RMSD superposes the sample on the target by the Kabsch algorithm (Kabsch, 1976; 1978) on the stable-core Cα atoms, with a reflection check (a mirror image can otherwise score lower than the true fit), and reports the root-mean-square deviation over the same atoms; superposing on the whole chain would let a mobile tail rotate the fit and corrupt the core value.

Retained secondary structure. $\mathrm{ss}_{\mathrm{core}}$ is, per structure, the fraction of stable-core residues whose simplified DSSP code is helix or strand; the denominator is the core size. The threshold applied to it is half the reference ensemble's core fraction: 0.484, 0.448 and 0.486 for the three systems.

H.3 TARGETS AND BASELINES

The two GroEL targets lie in the tails of the reference ensemble along ξ, the extended one at the 99.9th percentile (5.8 standard deviations from the median) and the compact one in the lowest 0.2% (1.7 SD); the lysenin targets are the extreme members of its ensemble along ξ, chosen the same way, and these four are withheld from every fit. DppA's targets are the open and closed PDB chains (App. G), which are not members of its dataset. §4.2 and §4.3 report the accuracies on all six, and Table S11 gives the lysenin and DppA targets along ξ.

Table S11: The lysenin and DppA targets along ξ, in nm, at the sample sizes of Table S1. The last column is the guided median on lysenin and the median of the sequential Monte Carlo outputs on DppA; each lies on its target's side of the unguided one.

	target's ξ	unguided median	steered median
lysenin, extended	4.07	2.58	4.42
lysenin, compact	1.85	2.58	2.21
DppA, open	2.46	2.35	2.44
DppA, closed	2.27	2.35	2.30

The PDB structures. The experimental targets of Figure 2 are the ten GroEL chains of Table S12, from seven PDB entries, aligned to the generator's 547-residue sequence; four are separate, independently resolved subunits of one crystal.

Table S12: The seven PDB entries behind the ten GroEL chains used as targets in Figure 2. Entry 4HEL has no primary publication and is credited as a database deposition.

entry	chain(s)	state	source
4HEL	A	apo, 3.2 Å (unpublished deposition)	(Saxena & Meena, 2012)
1OEL	A	apo, 2.8 Å	(Braig et al., 1995)
1KP8	A	$(\text{GroEL–KMgATP})_{14}$, 2.0 Å	(Wang & Boisvert, 2003)
4KI8	A	GroEL–ADP, relaxed state	(Fei et al., 2013)
1AON	A	GroES-bound, open (cis ring)	(Xu et al., 1997)
1XCK	E, G, I, J	apo wild type, four distinct subunits	(Bartolucci et al., 2005)
5W0S	H	cryo-EM, per-subunit variation	(Roh et al., 2017)

Selection of the samples shown. Panels are selected neither on distance to the target, which would require the answer, nor on ξ, which would manufacture the result. They are drawn at random, at a fixed seed, from the samples that pass the distance check and retain at least the floor of core secondary structure, two criteria that are properties of each protein's reference ensemble; DppA's panels are drawn from its sequential Monte Carlo outputs.

The colours are each residue's deviation after the superposition the accuracy numbers use, on the stable core. DppA's samples are instead superposed on the 142 core residues whose positions the two PDB chains share, so that its colours show the hinge motion of the other lobe.

Because a total deviation is dominated by its tail, the colour ramp covers only the core residues that hold one secondary structure in at least 98% of the ensemble (321, 50 and 224 residues), a strict subset of the stable core; the other residues are drawn faintly. Restricting the ramp changes GroEL's 90th-percentile deviation by 0.02 nm, where the deviation is the apical swing of ordered residues, and lysenin's by half a nanometre, where one mobile loop would otherwise set the scale.

On GroEL the camera is framed on the coloured residues, so the faint chain may leave the panel. On lysenin and DppA it is framed on the whole chain, because the extension that separates lysenin's two targets lies outside the coloured residues.

I Population inference

I.1 Candidates, imposed populations, and particles

300 GroEL conformations are reserved as candidates and simulated as 128 images each; the observed particles come from the remaining 2,900, held out from the candidate library but not from the paired dataset the code was fitted on (App. G). For a held-out conformation c with coordinate ξ_c the imposed weight (Table S13) is

$$q_c = \frac{p_{\text{desired}}(\xi_c)/\widehat{p}_{\text{ref}}(\xi_c)}{\sum_{c'} p_{\text{desired}}(\xi_{c'})/\widehat{p}_{\text{ref}}(\xi_{c'})}, \tag{S10}$$

where $\widehat{p}_{\text{ref}}$ is the reference ensemble's density along ξ, a 33-bin histogram over the ensemble's range, smoothed by one bin and floored. The ratio is clipped below at zero and otherwise untrimmed

(an effective-sample-size floor had no effect on the barrier's draw-to-draw scatter), and the four landscapes' effective sample sizes span 19–70% of the held-out conformations. 4,000 conformations are drawn with replacement from q, each represented by one of its 128 images chosen uniformly, so a conformation drawn several times contributes one independently chosen image per draw; a particle set whose most frequent conformation is drawn more often than it has images is rejected, because the surplus particles would be identical repeats. Here that count is at most 34.

Table S13: The four imposed landscapes. Each is a two-component Gaussian mixture in ξ, the components peak-normalised and weighted by the population shares p. "At Δ" is the imposed weights through the same reporting kernel as the estimate.

	centres (nm)	widths (nm)	shares	separation	barrier (k_BT)		effective
				(nm)	imposed	at Δ	size
separated states	4.20, 5.85	0.36, 0.36	0.50, 0.50	1.54	1.93	1.83	19%
rare state	4.20, 5.60	0.26, 0.26	0.88, 0.12	1.36	4.01	3.46	70%
deep barrier	4.20, 5.60	0.26, 0.26	0.50, 0.50	1.36	2.93	2.34	28%
minimum separation	4.20, 5.05	0.18, 0.18	0.50, 0.50	0.82	2.09	1.19	56%

Library. The candidate images are simulated with the settings of the dataset of §4.2, at $\mathrm{SNR}_{\mathrm{box}} = 0.005$, and their codes come from the 5-dimensional image map that guides the GroEL runs; nothing is refitted, so the two halves of the paper share one dataset, one code and one noise level.

I.2 The candidate distributions

Each candidate's code distribution is measured from its images by placing a kernel at each code:

$$K_i(y) = \frac{1}{V}\sum_{v=1}^{V} T_{b_i}(y - a_{i,v}), \qquad a_{i,v} = \overline{y}_i + s\,(y_{i,v} - \overline{y}_i), \tag{S11}$$

with T_{b_i} a product of univariate Student-t densities with 5 degrees of freedom, one per code coordinate, with no rotation or whitening. $b_{i,k}$ is a standard deviation, not the Student-t scale: the density is evaluated at scale $b_{i,k}/\sqrt{\nu/(\nu-2)}$, so each component has variance $b_{i,k}^2$ and differs from the Gaussian it replaces only in its tails. The widths follow Silverman's rule (Silverman, 1986) on the candidate's image codes, $b_{i,k} = |s|\,\mathrm{sd}_v(y_{i,v,k}) \cdot V^{-1/(4+r)}$, with a floor inactive on all 300×5 pairs, so one ratio holds for the whole library,

$$\frac{\mathrm{sd}(K_i)}{\mathrm{sd}(y_{i,\cdot})} = |s|\sqrt{1 - \tfrac{1}{V} + V^{-2/(4+r)}} = 0.971, \tag{S12}$$

and the fitted kernel is 2.9% narrower than the image-code scatter it is built from.

The kernel scale s. s is profiled against the fit's penalised objective (a nine-point grid, then bounded Brent refinement) on one set of 4,000 particles drawn from uniform weights over the held-out conformations, and held fixed for every imposed landscape of that library; its value is 0.84.

I.3 Resolution limit and basis

H is the structural spread below which the code cannot resolve, a property of the candidate library alone. With $\overline{y}_i$ candidate i's conformation-mean code, the image-scatter radius is the median over candidates of the root-mean-square deviation of a candidate's images from its mean, and H is the median over candidates of the standard deviation of ξ among the candidates within that radius,

$$r_{\mathrm{img}} = \operatorname*{median}_i \sqrt{\tfrac{1}{V}\textstyle\sum_v \|y_{i,v} - \overline{y}_i\|^2}, \qquad H = \operatorname*{median}_i \mathrm{sd}\Big\{\xi_j : \|\overline{y}_j - \overline{y}_i\| \le r_{\mathrm{img}}\Big\}. \tag{S13}$$

The metric is Euclidean over all 5 code coordinates; whitened variants reduce H and the estimator then reports spurious features, so they were rejected. A candidate is inside its own ball, and candidates with fewer than two neighbours are dropped from the outer median rather than imputed (2 of 300). Both aggregations are medians, so a few candidates in sparse regions cannot set the resolution

of the whole curve. On this library $r_{\mathrm{img}} = 5.49$ and $H = 0.412\,\mathrm{nm}$, and the minimum separation for reporting a barrier is $2H = 0.82\,\mathrm{nm}$.

The weights are $w(\vartheta) = \mathrm{softmax}(B\vartheta)$ with B a harmonic basis in the standardised ξ: a constant, and $\cos$ and $\sin$ at integer multiples of a fundamental whose half-period is the span between the 1st and 99th percentiles of the candidates' ξ (the reported range). The order $\lfloor \mathrm{span}/H \rfloor = 5$ keeps the finest half-period no finer than H and gives $2 \times 5 + 1 = 11$ columns; the reported range being a half-period, the two edges are not forced to agree. The softmax has no candidate base measure, so uniform ϑ gives uniform weights over candidates, not a uniform density in ξ.

I.4 Penalty and optimisation

$R = B^\top D^\top D B$, with D the second difference on candidates ordered by ξ, trace-normalised so that $\mathrm{tr}(R)$ equals the number of columns of B. The stencil is a second difference in candidate index rather than in nanometres, so where candidates are sparse it penalises a larger physical curvature, and the trace normalisation keeps the strength comparable across libraries; a constant log-weight has zero second difference, so R leaves the softmax gauge unpenalised. The objective is minimised by Levenberg–Marquardt-damped Newton iteration on the exact Hessian, which at eleven parameters costs about as much as the gradient (undamped Newton diverges from a cold start), from eleven starts of three kinds: one expectation step from uniform weights, four temperatures of a deterministic annealing schedule, and six fixed random seeds. The spread of objective values across starts is below $4 \times 10^{-15}\,k_{\mathrm{B}}T$ on all 75 fits on this library.

Selection of ρ. The strength is the weakest value at which the estimator reports no barrier on particle sets drawn from a single-basin landscape, which imposes none. In three synthetic single-basin batteries of 32 particle sets each, the unpenalised fit reports a barrier in 12, 5 and 10 sets, $\rho = 10^{-3}$ in 2, 0 and 4, and $\rho = 0.01$ in none, so two batteries select 0.01 and the third 10^{-3}. The six single-basin landscapes of the benchmark grid (25 imposed landscapes) report no barrier at any strength over 18 sets each and cannot discriminate.

The penalty's cost is missed barriers, not depth. The depth slope, the slope of the log recovered barrier against the log imposed one at the reporting width, is one for an unbiased estimator and below one when recovered depths are compressed. Across eight strengths from 0 to 1 it is flat on the benchmark grid (0.34–0.41, within its standard errors), while the imposed barriers the estimate misses rise from 11 to 49 of 51. The slope and the $0.60\,k_{\mathrm{B}}T$ bound of §4.4 describe one compression over different ranges: the three recovered landscapes carry barriers of only 1.83–$3.46\,k_{\mathrm{B}}T$ at the reporting width, where it costs at most that much, and it shows in their signs, the shallowest barrier being recovered above its value on every replicate and the two deeper ones below theirs on every reported replicate (Table S14). The error grows for barriers outside that range.

The same rule selects the penalty's form. A curvature penalty shrinks the weights toward smoothness, which a two-well landscape nearly has, whereas a flat ridge shrinks them toward the candidate library's density and fills barriers.

Both choices used the imposed landscapes, which bounds what §4.4 can claim. Of three criteria that do not use them, a Laplace approximation to the marginal likelihood selects 0.01 on 15 of the 25 landscapes, and the two L-curve criteria select 3×10^{-2} and 10^{-1}.

The objective averages the log-likelihood over particles and the penalty does not, so a constant ρ is a prior whose weight grows with the data. ρ is therefore specified in total-log-likelihood units at a reference particle count N_0 and applied as $\rho\, N_0/N$, the identity at the 4,000 particles used.

I.5 Reporting criteria

All three curves of Figure 4 are drawn at one width Δ and one grid. Δ is distinct from H (conflating them would apply the resolution limit twice): twice the 99th-percentile gap between adjacent candidate values of ξ inside the reported range, $0.139\,\mathrm{nm}$ here, so the curve is never drawn finer than the library samples. The grid runs at $\Delta/10$ across the reported range with one Δ of padding on each side, so an edge point still has the neighbours the local-minimum test needs; the density is floored before the logarithm, and a case that reaches the floor is rejected. The free-energy zero is arbitrary,

and each curve is offset onto the imposed landscape's by the constant that minimises their difference over that range.

A barrier is reported only if the two minima are at least $2H$ apart and the barrier rises at least $0.5\,k_{\mathrm{B}}T$ above the shallower minimum. The first condition is the bimodality bound for an equal-weight two-component Gaussian mixture, bimodal if and only if its components are more than twice their common width apart (de Helguero, 1904; Schilling et al., 2002), applied at the resolution H; the prominence floor binds in practice, requiring about $3.1H$ on its own. Depth is measured from the deeper minimum and prominence from the shallower one, so they coincide only for a symmetric pair.

Table S14: All 3 replicate particle sets per landscape, in $k_{\mathrm{B}}T$; "at Δ" is the imposed weights through the same kernel, and a dash marks a replicate whose prominence does not clear the 0.5 floor.

	separation / $2H$	at Δ	recovered (prominence)		
			replicate 1	2	3
separated states	1.9	1.83	2.20 (2.02)	2.13 (2.00)	1.83 (1.79)
rare state	1.7	3.46	2.87 (0.55)	— (0.38)	3.14 (0.62)
deep barrier	1.7	2.34	2.19 (1.58)	1.81 (1.32)	1.86 (1.47)
minimum separation	1.0	1.19	— (0.44)	— (0.19)	— (0.31)

The largest deviation from the imposed landscape at that width over the 8 recovered replicates (Table S14) is $0.60\,k_{\mathrm{B}}T$; on the remaining rare-state replicate the second basin's prominence falls below the floor. Across the 25 imposed landscapes on which the estimator was scored, 41 of 51 replicates whose imposed landscape carries a reportable barrier clear the floor, and only the minimum-separation landscape fails on every replicate.

Independent ensemble. Figure 4a uses the ATLAS trajectories (Vander Meersche et al., 2024) of the GroEL subunit (entry 1SX3, chains A and B Chaudhry et al., 2004; CHARMM36m Huang et al., 2017): 3,006 frames from three 100 ns replicates per chain, simulated and encoded as the candidate library was, with the library's candidates and resolution. At matched ξ the ensemble's codes are displaced from the library's by a constant vector, so the estimator explains the ensemble's particles with the candidates nearest their displaced codes rather than those at their own ξ: read as is, the recovered minimum sits 0.07 nm below the true one with an RMS deviation of $0.57\,k_{\mathrm{B}}T$ over the occupied range, and removing the displacement, measured against the ensemble's ξ, gives the landscape shown ($0.10\,k_{\mathrm{B}}T$).

Comparison curves. Particle counting assigns each particle to the candidate with the largest kernel density (an implicit uniform prior, no iteration) and counts assignments, with the same kernels, scale, grid and reporting width, so its difference from the fitted curve is the reweighting alone. The imposed landscape, the true landscape of Figure 4b–e, is the weights q_c on the held-out conformations, not the realised frequencies of one particle set, so it is identical across replicates and is what a perfect estimator would return.

Uncertainty band. The band is ± 2 standard errors from the sandwich covariance of $\hat{\vartheta}$ on that replicate's fit, propagated through Eq. 10 by the delta method (it matches a central difference to 4×10^{-9}, vanishes along the softmax gauge to 4×10^{-16}, and reproduces a 60-resample particle bootstrap at a band-mean ratio of 0.99). Two half-particle fits differ by 0.92–1.03 of what the band predicts, so it is correct as a variance statement; its coverage of the imposed landscape is 0.55–0.67 against a nominal 0.95, because the remaining error is bias. It resamples particles only, with poses, images, candidate library and s fixed, so it is not an error bar on the landscape: re-simulating the candidate images would move the curve by more than it admits.

J Extended related work

Heterogeneous reconstruction. The per-particle-latent methods of §2 (Zhong et al., 2021; Punjani & Fleet, 2021; 2023; Gilles & Singer, 2025) have relatives that add a prior meant to separate

structure from pose and defocus in the latent (Luo et al., 2023), that extend the approach to cryo-electron tomography (Powell & Davis, 2024), and that infer orientation jointly or mix over composition (Levy et al., 2022; 2024b; 2025); landscapes are also read directly off image embeddings (Dashti et al., 2014; Frank & Ourmazd, 2016). The measurement deforms the metric of any such embedding, which no embedding recovers in general (Lederman & Toader, 2023), so latent distance and density are not conformational distance and density; linear-subspace methods that deconvolve the embedding density recover the density (Gilles & Singer, 2025). Nor is an image-read latent pose-invariant: a 90° in-plane rotation of the image can change the conformation that cryoDRGN (Zhong et al., 2021) assigns to it (Edelberg & Lederman, 2023), and manifold methods recover a geometry that varies with viewing direction and is distorted by the contrast transfer function (Seitz et al., 2022). How much survives is contested (Dai et al., 2026); measurement noise, more than particle count, is argued to set the resolution at which populations are learnable (Mattingly et al., 2026). On CryoBench's molecular-dynamics dataset, of the kind Roodmus generates (Joosten et al., 2024), every method tested overlaps the true conformational neighbourhoods by under 35% at small-to-intermediate radii (Jeon et al., 2024), and in a community challenge the motions the methods identified qualitatively resembled the ground truth, but most submissions returned two modes of three, missing the sparsely populated middle state (Astore et al., 2025).

Structure-space methods. In the family that deforms a reference structure (Jin et al., 2014; Nakane et al., 2018; Chen & Ludtke, 2021; Vuillemot et al., 2023; Schwab et al., 2024; Li et al., 2024; Ducrocq et al., 2025) the representable motion lies in a class each method fixes, whether rigid bodies, contiguous segments, a force field, pseudo-atoms fitted to one averaged map, a smooth deformation of one structure or an adaptively relaxed elastic network on one atomic model, so motion outside that class is unrepresentable or admissible only against a penalty the method imposes. In two control experiments with DynaMight (Schwab et al., 2024), one domain of the atomic model regularising the deformations was replaced by the model of a different protein. In both, the reconstructed density of that domain resembled the replacement more than the correct model; in the more heterogeneous of the two domains the incorrect model also improved local resolution about as much as the correct one did, and in the other it gave lower local resolution than the correct one. DynaMight's authors kept the atomic-model regulariser out of the graphical user interface and, because correct and incorrect models would be difficult to distinguish, cautioned against its use. Li et al. (2024) instead decode the density separately from the regularised structure and report maps robust to a deleted or rotated subunit, but need the input atomic model to cover the flexible regions and, like DynaMight, take each particle's pose as known. Predicting coordinates outright avoids the hypothesis class and was proposed on simulated data in 2021 (Rosenbaum et al., 2021); the supervised form needs ground-truth structures per system and has been shown only on synthetic data (Abid & Jonić, 2026), and the self-supervised form takes a pretrained folding prior as its hypothesis class, fine-tuning its structure module per dataset by matching images rendered from its predictions to the particles, experimental as well as synthetic (Vuillemot & Grudinin, 2026). The per-particle Bayesian family infers over a free-energy profile (Giraldo-Barreto et al., 2021), over ensemble weights (Tang et al., 2023), or, closest to ours, over each particle's conformation with the posterior amortised by a neural density estimator trained on physics-simulated images (Dingeldein et al., 2025); such a posterior can be diagnosed as incompatible with every candidate, as cryoSBI's is for contaminants, whose posteriors peak at an end of the template range, but it cannot propose what is missing. The one method in this family that moves the structures alternates between optimising a fixed-size set of structures and their population weights against a particle-image likelihood, with a force-field prior imposed by an approximate projection (steered molecular dynamics toward each structure's unconstrained update) (Silva-Sánchez et al., 2026); it optimises a set where we sample a distribution.

Generative ensemble priors. Generative models now sample conformational ensembles (Lewis et al., 2025; Jing et al., 2024; Lu et al., 2024; Zheng et al., 2024) from a sequence or a starting structure; an imposed scalar property enters as sampling-time guidance from a property predictor, shown on carbon allotropes with target band gaps (Zheng et al., 2024). Experiment-guided sampling turns NMR, SAXS or cryo-EM observables into a correction on a pretrained sampler's score (Liu et al., 2025); that channel carries a measurement of the ensemble, where ours carries individual particles. Conditioning a frozen structure model on a reconstructed map (Levy et al., 2024a; Raghu et al., 2025; Maddipatla et al., 2026a), or fine-tuning it on maps (Shenoy et al., 2026), places the method after picking, pose estimation, refinement and classification. Three methods optimise

a frozen predictor's input or internal embeddings against an experimental likelihood (Fadini et al., 2026; Maddipatla et al., 2026b; Li et al., 2026); ROCKET's cryo-EM route steers AlphaFold2's co-evolutionary embeddings (Jumper et al., 2021) without retraining, from two half-maps and a docked molecular-replacement solution, and all three ascend an objective whose likelihood weight is set by hand, any posterior being over the conditioning representation rather than the generator's per-step state. CoCoFold reaches a pretrained model with raw particles, but fine-tunes AlphaFold's structure module against them under poses and contrast transfer parameters from an upstream workflow (Liao et al., 2026).

Guidance and the likelihood covariance. Guidance of diffusion models runs from classifier and classifier-free guidance (Dhariwal & Nichol, 2021; Ho & Salimans, 2022) to a training-free family that leaves the model frozen and differentiates an objective through the denoiser's clean-signal estimate, with a known measurement operator (Chung et al., 2023b; Song et al., 2023) or an off-the-shelf predictor (Bansal et al., 2024; Ye et al., 2024), extended to latent diffusion models through a decoder, either by gradient steps (Rout et al., 2023) or by optimising the latent from the clean-signal estimate to fit the measurement and then resampling it back to the current noise level (Song et al., 2024); blind variants put a second diffusion prior on the operator (Chung et al., 2023a) or alternate between operator and image in an approximate Gibbs scheme (Murata et al., 2023). The tractable surrogates replace the clean signal's distribution given the noisy state with a point mass at its posterior mean (Chung et al., 2023b) or a two-moment Gaussian (Boys et al., 2024). Both approximate only that distribution and keep the measurement likelihood, which for a raw particle is the pose marginal of §2. We take that marginal to be the reason the cryo-EM methods above condition on a reconstructed map instead.

Non-isotropic likelihood covariances exist in this literature: Song et al. (2023) build one from the known operator and a scalar denoiser variance, Kawar et al. (2022) diagonalise the operator and work per singular value, which rescales isotropic measurement noise by the inverse singular value, and Boys et al. (2024) replace the scalar denoiser variance with the denoiser's covariance from Tweedie's formula; where the denoiser's covariance is a hand-crafted isotropic one, as in Chung et al. (2023b) and Song et al. (2023), the limitation has been noted and the optimal choice characterised (Peng et al., 2024). Ours sits on the observation side, is measured (a linear map out of the frozen generator's hidden state, with the empirical per-image covariance of a pretrained encoder's embedding as its noise, where the works above derive their anisotropy from a known operator or the denoiser's covariance) and depends on the number of observations, which no likelihood covariance in this line does. A pretrained model's internal representations are an established semantic space, read and manipulated without retraining (Kwon et al., 2023; Park et al., 2023; Epstein et al., 2023; Turner et al., 2023; Li et al., 2023; Zou et al., 2023), and heads trained on a frozen denoiser's features have been differentiated back into sampling (Graikos et al., 2023; Luo et al., 2024). What is differentiated there is a point, an offset, a mean-difference direction (Arditi et al., 2024) or a regressor, whose per-input effect is heterogeneous (Tan et al., 2024) and which descends a loss rather than following a conditional score (Shen et al., 2024); gating it on the input (Lee et al., 2025), fitting a region rather than a vector (Postmus & Abreu, 2024) or training a low-rank intervention (Wu et al., 2024) enriches the object without giving it a density, and probes that place a posterior over a concept given frozen activations are discriminative and read-only (Wang et al., 2023). No prior work conditions a fully frozen generative ensemble prior on raw particle images through a fitted, never inverted measurement model over the prior's internal state whose covariance is measured and depends on the number of particles.