# FOLD'EM: DIRECT ATOMIC STRUCTURE INFERENCE FROM RAW CRYO-EM PARTICLES

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Single-particle cryo-electron microscopy (cryo-EM) has become a widely adopted technique for biomolecular structure determination. The conventional cryo-EM computational pipeline first combines many particle images to reconstruct an electrostatic potential (ESP) map and then fits an atomic model to the recovered map. Density reconstruction has high sample complexity, requiring large numbers of particle images and making structure determination high-cost and low-throughput, particularly for structurally heterogeneous samples. Downstream atomic model building, in turn, becomes increasingly difficult as the resolution of the reconstructed map deteriorates. Protein structure prediction models provide strong sequence-derived priors on atomic structure, and experiment-guided approaches can use these priors to recover structures consistent with experimental measurements. Yet, in cryo-EM, such priors are typically integrated only after density reconstruction during atomic model fitting. We introduce Fold'EM, an inference-time framework that combines priors from protein structure generative models directly with cryo-EM particle images to determine atomic models from a small number of single particle images, bypassing both intermediate density reconstruction and downstream model building against the reconstructed map. Across synthetic benchmarks and experimental cryo-EM data, Fold'EM recovers accurate atomic structures both with known particle orientations and in an *ab initio* setting where orientations are inferred jointly with structure. In heterogeneous datasets, Fold'EM further resolves distinct conformational states from mixed particle populations without separately reconstructing a density map and building an atomic model for each state. We believe these results open new avenues for structure determination in the low-sample regime and for characterizing low-population conformational states directly from cryo-EM particles.

## 1 INTRODUCTION

Single-particle cryo-electron microscopy (cryo-EM) enables high-resolution structure determination of large macromolecular assemblies without crystallization (Fernandez-Leiro & Scheres, 2016), and can capture molecules in different conformations to reveal structural mechanisms (Ourmazd, 2019; Mäeots & Enchev, 2022). Individual particle images, however, are extremely noisy because radiation damage limits the electron dose. Structure determination therefore relies on aligning and averaging many observations into a three-dimensional electrostatic potential (ESP) map (Scheres, 2012; Kimanius et al., 2024), which must subsequently be interpreted as an atomic model, a step that remains difficult where local resolution is poor (Jamali et al., 2024; Fadini et al., 2026).

This averaging creates a tension between spatial resolution and biological completeness. Combining different conformations blurs structural features, whereas separating them leaves fewer observations per state (Frank & Ourmazd, 2016). Rare and transient states may therefore be merged with abundant conformations, discarded during particle selection, or remain too sparsely represented for independent reconstruction. For scarce samples, the required observations may simply not be available.

Methods for heterogeneous reconstruction can recover variability beyond a small number of discrete classes (Frank & Ourmazd, 2016; Zhong et al., 2021; Gilles & Singer, 2025), but many rely on upstream particle alignment and therefore do not eliminate the need for large particle collections. Conventional reconstruction also uses relatively little prior knowledge about molecular structure or the sequence of the sample. Blush regularization in RELION illustrates the value of stronger priors:



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.

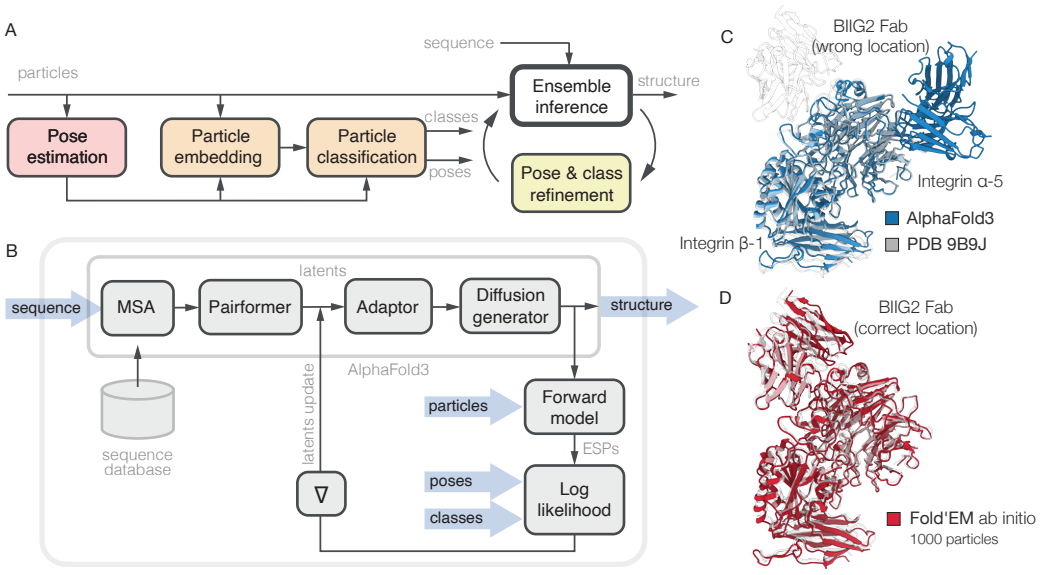


Figure 1: **Fold'EM infers atomic structures directly from sparse cryo-EM particle images.** (A) Fold'EM alternates structural inference with pose and class refinement. (B) Particle evidence is back-propagated into intermediate Pairformer states $\mathbf{H}$, filtering experimental updates through the remaining Pairformer blocks before AlphaFold3 diffusion. (C) For the integrin–BIIG2 Fab complex (PDB `9B9J`), AlphaFold3 predicts the individual components but places the Fab at the wrong binding site. (D) Ab initio Fold'EM recovers the bound complex and epitope from only $1{,}000$ particles—$0.26\%$ of those used for the deposited reconstruction—with RMSD $2.49$ Å, TM-score $0.956$, and FSC resolution $3.52$ Å (deposited map: $2.6$ Å).

a denoiser trained on experimental cryo-EM maps improves reconstruction when particle signal is weak (Kimanius et al., 2024).

AlphaFold and related models provide much stronger sequence-conditioned structural priors learned from evolutionary and structural data (Jumper et al., 2021; Baek et al., 2021). Recent methods have shown that experimental evidence can steer these models toward conformations missed by unguided prediction (Maddipatla et al., 2026b; Fadini et al., 2026; Raghu et al., 2026). In cryo-EM, however, such guidance has primarily operated on reconstructed maps, after particle selection, alignment, and averaging. We therefore ask: *can structural priors be introduced directly at the particle level, enabling atomic inference from sparse cryo-EM observations without requiring a high-resolution intermediate reconstruction?*

### 1.1 Contributions and Main Results

We introduce **Fold'EM**, an inference-time framework that combines sequence-conditioned structural priors with cryo-EM particle images. Rather than reconstructing a high-resolution density and subsequently building an atomic model, Fold'EM jointly infers atomic structures and particle-specific latent variables, including pose and conformational assignment (Figure 1).

**Particle vs. density guidance.** For a homogeneous sample with known poses, we show that fitting individual particles is equivalent to fitting a reconstructed density only when density mismatch is measured under the experiment-induced metric defined by the imaging operators and noise model. We show that the usual density-space $L_2$ objective does not, in general, reproduce the particle likelihood and when poses or class assignments are uncertain, a single consensus density may additionally discard particle-specific information required to infer these variables (Figure A2). This motivates introducing structural priors directly at the particle level.

**Improved experimental guidance.** Existing approaches guide AlphaFold by optimizing its terminal representation (Maddipatla et al., 2026a; Li et al., 2026) or the diffusion trajectory in coordinate space (Maddipatla et al., 2026b; Raghu et al., 2026). Fold'EM instead optimizes an intermediate Pairformer state. Passing its updates through the remaining Pairformer blocks induces a sequence- and state-

dependent geometry in the terminal conditioning space, substantially strengthening experimental guidance (Figure A1).

**Low-sample structure inference.** With known poses, Fold'EM recovers accurate structures on both experimental (Figure 2; Tables A3 – A4) and synthetic (Tables A5–A6) datasets using fewer than $1\%$ of the particles used for conventional reconstruction. On experimental homogeneous data from EMPIAR, as few as $0.1\%$ suffice to reach comparable FSC and atomic RMSD; heterogeneous results are shown in Figures 3 and A3, Tables A15–A16 and A19–A20. In continuously heterogeneous data, we further find that conformational organization in particle space is reflected in the AlphaFold representation, allowing a continuous trajectory to be represented by a small number of optimized anchor states (Figure 4).

***Ab initio* Fold'EM.** When poses and classes are unknown, Fold'EM alternates structural inference with refinement of these nuisance variables. For low-sample heterogeneous data, we introduce a RECOVAR-inspired residual-subspace estimator, B-RECOVAR, that provides particle assignments and confidence (Figure A4, and Table A24). On experimental data, ab initio Fold'EM reaches comparable structural accuracy using approximately $3\%$ of the particles used in the deposited reconstruction (Figure 2, Tables A9–A10) and resolves distinct heterogeneous states from approximately $5\%$ (experimental: Figure 3, Tables A17–A18; synthetic: Tables A21–A22).

Together, these results show that strong structural priors can substantially reduce the particle counts needed for atomic and conformational inference and suggest a route toward characterizing low-population states directly from cryo-EM particle images.

## 2 Problem formulation

**Notation.** Let $\mathbf{a}$ denote the molecular sequence specification supplied to AlphaFold3 and $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_m\}$ the resulting atomic coordinates, $\mathbf{x}_i \in \mathbb{R}^3$. We view AlphaFold3 as a sequence-conditioned latent model: the Pairformer produces $\mathbf{Z} = f(\mathbf{a})$ and a diffusion generator samples $\mathbf{X} \sim p(\mathbf{X} \mid \mathbf{Z})$. We denote $\mathbf{Z} = (\mathbf{s}, \mathbf{z})$, with $\mathbf{s} \in \mathbb{R}^{|\mathbf{a}| \times c_s}$ and $\mathbf{z} \in \mathbb{R}^{|\mathbf{a}| \times |\mathbf{a}| \times c_z}$. For Pairformer blocks $f = f_N \circ \cdots \circ f_1$, we split the trunk as $f = g \circ h$, where $\mathbf{H} = h(\mathbf{a})$ and $g$ contains the final $\ell$ blocks, so $\mathbf{Z} = g(\mathbf{H})$. Pairformer blocks preserve the tensor shapes of the single and pair representations, allowing $\mathbf{H}$ and $\mathbf{Z}$ to be optimized in corresponding ambient spaces. We use $\ell = 3$ throughout.

**Inference objective.** Let $\boldsymbol{\mathcal{Y}} = \{\mathbf{y}_k\}_{k=1}^K$ be the observed particles and $\boldsymbol{\mathcal{H}} = \{\mathbf{H}^j\}_{j=1}^n$ the latent states of an $n$-member structural ensemble. Particle $k$ has nuisance variables $\boldsymbol{\eta}_k = (\boldsymbol{\phi}_k, c_k)$, where $\boldsymbol{\phi}_k$ is its pose and $c_k$ its structural class; all other imaging parameters are assumed known and suppressed. Conditional independence gives

$$p(\boldsymbol{\mathcal{Y}} \mid \boldsymbol{\mathcal{X}}, \boldsymbol{\eta}) = \prod_{k=1}^{K} p(\mathbf{y}_k \mid \mathbf{X}^{c_k}, \boldsymbol{\phi}_k).$$

For continuous heterogeneity, $c_k$ may instead denote barycentric coordinates and the conditioning state is $\mathbf{H}(c_k) = \sum_j c_k^j \mathbf{H}^j$. We relax the deterministic state $\mathbf{H}_0 = h(\mathbf{a})$ using $p(\boldsymbol{\mathcal{H}} \mid \mathbf{a}) = \prod_{j=1}^n \mathcal{N}(\mathbf{H}^j; \mathbf{H}_0, \sigma^2 \mathbf{I})$. The corresponding marginal MAP objective is

$$\mathcal{J}(\boldsymbol{\mathcal{H}}, \boldsymbol{\eta}) = \log \mathbb{E}_{\boldsymbol{\mathcal{X}} \sim p(\boldsymbol{\mathcal{X}} \mid g(\boldsymbol{\mathcal{H}}))} p(\boldsymbol{\mathcal{Y}} \mid \boldsymbol{\mathcal{X}}, \boldsymbol{\eta}) + \log p(\boldsymbol{\mathcal{H}} \mid \mathbf{a}) + \log p(\boldsymbol{\eta}). \tag{1}$$

Because the AlphaFold generator is stochastic, we optimize the Jensen lower bound and approximate

$$(\boldsymbol{\mathcal{H}}^*, \boldsymbol{\eta}^*) = \arg\max \mathbb{E}_{\boldsymbol{\mathcal{X}} \sim p(\boldsymbol{\mathcal{X}} \mid g(\boldsymbol{\mathcal{H}}))} \log p(\boldsymbol{\mathcal{Y}} \mid \boldsymbol{\mathcal{X}}, \boldsymbol{\eta}) + \log p(\boldsymbol{\mathcal{H}} \mid \mathbf{a}) + \log p(\boldsymbol{\eta}). \tag{2}$$

**Optimizing $\mathcal{H}$ induces a learned geometry in $\mathcal{Z}$.** For one ensemble member, let $\mathbf{Z} = g(\mathbf{H})$, $\mathbf{J}(\mathbf{H}) = \partial \mathbf{Z} / \partial \mathbf{H}$, and let $L(\mathbf{Z})$ denote the experimental data term in equation 2. Then

$$\nabla_{\mathbf{H}} L = \mathbf{J}^\top \nabla_{\mathbf{Z}} L, \qquad \Delta \mathbf{Z} \approx \gamma \, \mathbf{J}\mathbf{J}^\top \nabla_{\mathbf{Z}} L.$$

Thus Euclidean gradient ascent in $\mathbf{H}$ induces, locally, a positive-semidefinite preconditioner $\mathbf{J}\mathbf{J}^\top$ in terminal conditioning space. Directions to which the remaining Pairformer blocks are sensitive are amplified, while insensitive directions are suppressed; the geometry therefore depends on the sequence and current representation.

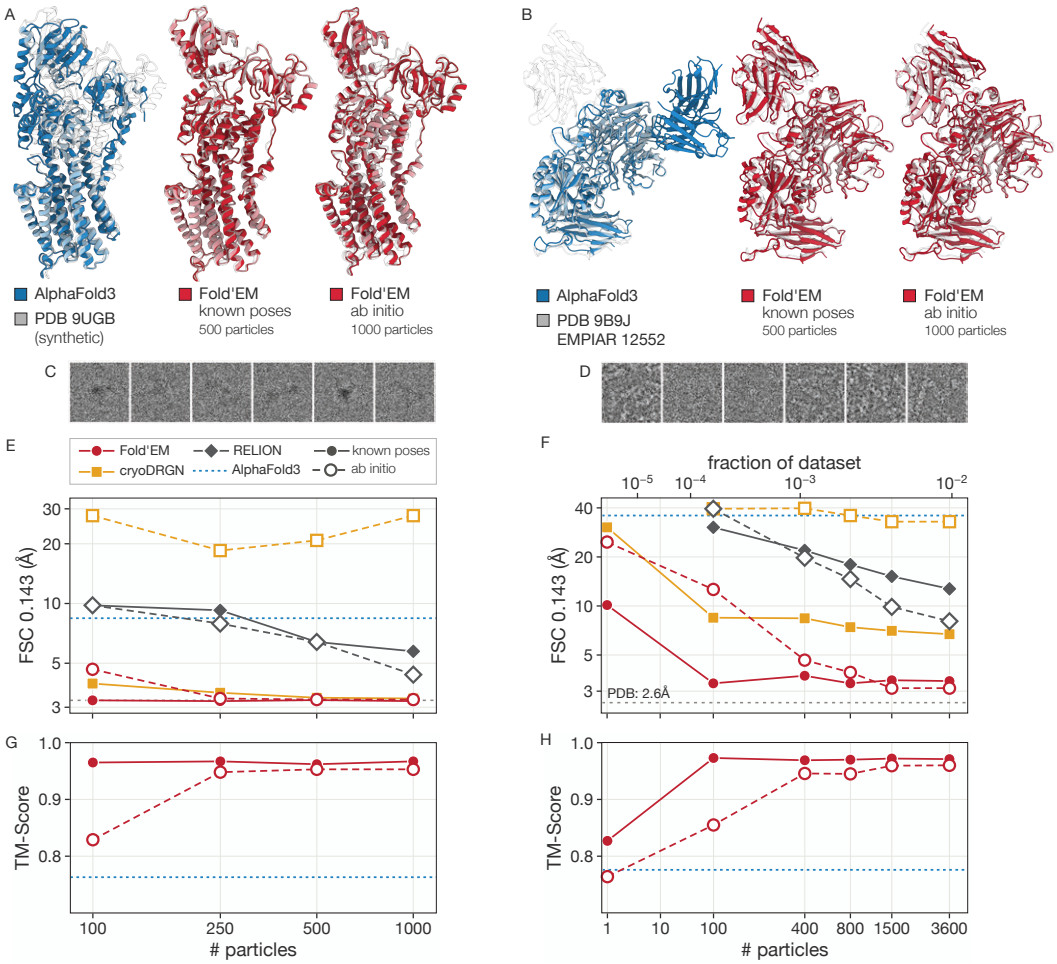


Figure 2: **Fold'EM recovers atomic structures from $0.03\%$ of the deposited particles with known poses and from $0.27\%$ ab initio.** Each column is one target: a synthetic dataset on the left (A, C, E, G) and an experimental one on the right (B, D, F, H). Rows show the structures (A, B; AlphaFold3 in blue, Fold'EM in red, deposited model in grey), representative particles (C, D), FSC resolution (E, F) and TM-score against the deposited model (G, H) as a function of the number of particles; filled markers denote known poses, open markers ab initio, and the dotted line the AlphaFold3 prediction. *Left:* Pma1, the yeast plasma-membrane $H^+$-ATPase, in its activated (E2P) state (9UGB); synthetic particles at SNR $0.005$. AlphaFold3 predicts the transmembrane domain but misplaces the cytosolic A-, N- and P-domains, whose arrangement changes along the catalytic cycle (TM-score $0.70$). Fold'EM recovers the deposited state, reaching $3.3$ Å and a TM-score of $0.97$ from $100$ particles with known poses, and $3.3$ Å from $500$ particles ab initio. *Right:* Integrin $\alpha_5\beta_1$–BIIG2 (9B9J, EMPIAR-12552), where AlphaFold3 predicts both component folds but places the Fab at the wrong site (TM-score $0.78$); antibody–antigen pairs provide little paired coevolutionary information. Fold'EM recovers the bound complex and epitope, although not the interface variable loops, reaching $3.5$ Å and a TM-score of $0.97$ from $100$ particles with known poses and $3.5$ Å from $1{,}000$ particles ab initio; no baseline reaches $6$ Å, and Fold'EM improves over AlphaFold3 even at the smallest particle counts. Sparse particle evidence therefore corrects the global arrangement largely supplied by the structural prior. Resolution (Å; FSC = 0.143).

The same view applies to the prior. Let $\mathbf{J}_0 = \mathbf{J}(\mathbf{H}_0)$ and $\mathbf{Z}_0 = g(\mathbf{H}_0)$. Linearization around $\mathbf{H}_0$ maps $\mathbf{H} \sim \mathcal{N}(\mathbf{H}_0, \sigma^2\mathbf{I}) \Longrightarrow \mathbf{Z} \approx \mathcal{N}(\mathbf{Z}_0, \sigma^2\mathbf{J}_0\mathbf{J}_0^\top)$. Hence an isotropic relaxation in precursor space becomes a sequence-dependent anisotropic relaxation in terminal space. Intuitively, experimental perturbations are filtered through transformations learned by the final Pairformer blocks rather than

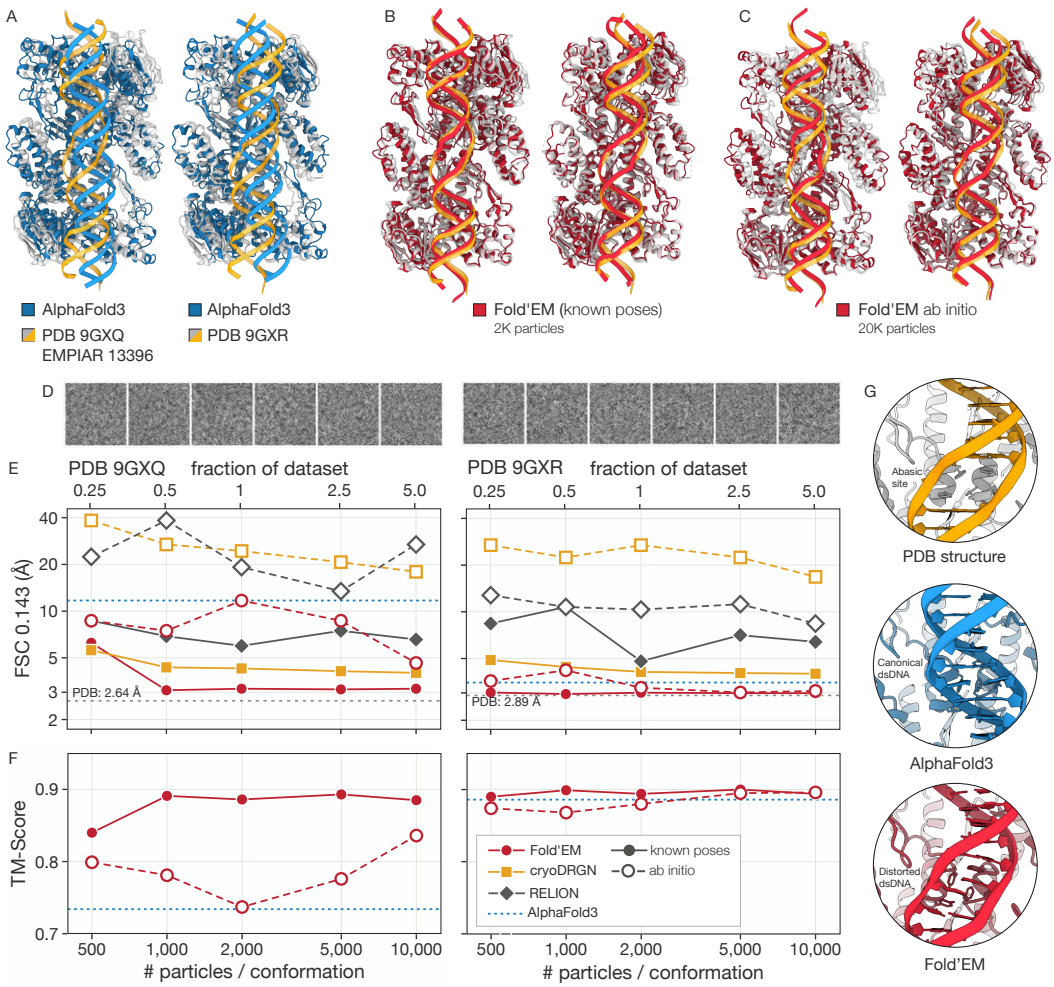


Figure 3: **Fold'EM resolves both UvrA–DNA conformations from $0.5\%$ of the deposited particles with known poses and from $5\%$ ab initio.** EMPIAR-13396 contains two states of UvrA bound to DNA carrying an abasic site (PDB 9GXQ/9GXR); damage recognition involves insertion of the Y62/Q65/F66 wedge into the duplex. (A) AlphaFold3 recovers 9GXR but not 9GXQ (TM-score 0.734). (B,C) Fold'EM resolves both states from 1,000 particles per state with known poses and from 10,000 particles per state ab initio (left: 9GXQ, right: 9GXR). (D) Representative particles. (E) FSC versus particles per state (filled: known poses; open: ab initio). With known poses Fold'EM reaches 3.1/3.0 Å from 1,000 particles per state; ab initio it reaches 3.0 Å on 9GXR and 4.6 Å on 9GXQ, while RELION and cryoDRGN-AI remain above 8 Å. (F) TM-score against the deposited models. (G) AlphaFold3 retains a fully paired duplex, whereas Fold'EM recovers the wedge insertion and disrupted DNA, but not the flipped-out nucleotide, as expected because the lesion was supplied to both models as intact DNA. Resolution (Å; FSC = 0.143).

applied directly to the terminal representation. Figure A1 and Appendix F.1 empirically demonstrate this.

**Advantages of particle-domain fitting.** Consider the linearized image formation model $\mathbf{y}_k = \mathbf{A}_k\rho + \varepsilon_k$, $\varepsilon_k \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_k)$, where $\mathbf{A}_k$ are known operators encapsulating the pose, CTF, and the sampling of a particle $k$. Define $\mathbf{Q} = \sum_k \mathbf{A}_k^* \mathbf{\Sigma}_k^{-1} \mathbf{A}_k$, $\mathbf{q} = \sum_k \mathbf{A}_k^* \mathbf{\Sigma}_k^{-1} \mathbf{y}_k$, and let $\tilde{\rho} = \mathbf{Q}^\dagger \mathbf{q}$ denote the reconstructed density. Then

$$\frac{1}{2}\sum_k \|\mathbf{y}_k - \mathbf{A}_k\rho(\mathbf{H})\|^2_{\mathbf{\Sigma}_k^{-1}} = \frac{1}{2}\|\rho(\mathbf{H}) - \tilde{\rho}\|^2_{\mathbf{Q}} + \text{const.}$$

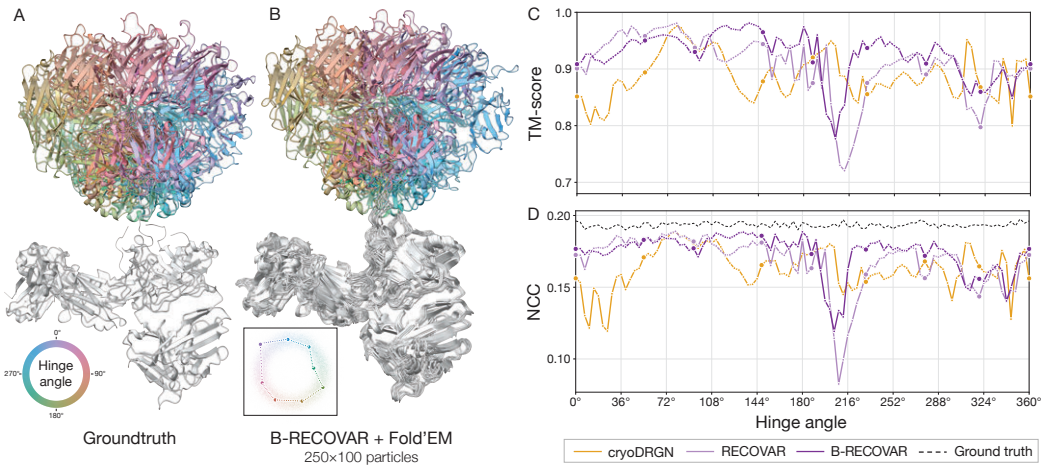


Figure 4: **Fold'EM resolves continuous heterogeneity of an IgG antibody complex.** (A) Shown as ensembles are 10 of the 100 ground-truth CryoBench IgG-1D synthetic models, colored by hinge angle; and (B) the corresponding structures recovered by Fold'EM with B-RECOVAR particle classification. The inset depicts the two leading principal components of the optimized Pairformer states, overlaid on the B-RECOVAR embedding of the particles. (C) TM-score and (D) Fourier NCC plotted as a function of hinge angle, for the three conformational coordinates.

Thus, with known poses, particle fitting and density fitting are equivalent only when the latter uses the experiment-induced metric $\mathbf{Q}$.

This metric is generally not the ordinary map-space $L_2$ norm. Under equal white noise and uniformly distributed viewing directions, for example, one has $\widehat{\mathbf{Q}\rho}(\boldsymbol{\xi}) \propto \|\boldsymbol{\xi}\|^{-1}\hat{\rho}(\boldsymbol{\xi})$ in the Fourier domain or, equivalently, $\mathbf{Q} \propto (-\Delta)^{-1/2}$, very far from $\mathbf{Q} = \mathbf{I}$. Consequently, Euclidean density guidance does not generally reproduce the particle likelihood even when poses are known (Maddipatla et al., 2026b; Raghu et al., 2026).

With uncertain poses or classes, the distinction is stronger because the imaging operators themselves are latent. For the scalar analogue $y_k = \phi_k \rho + \varepsilon_k$, with $\phi_k \sim \mathcal{N}(1, \tau^2)$ and $\varepsilon_k \sim \mathcal{N}(0, \sigma^2)$, one has $y_k \mid \rho \sim \mathcal{N}(\rho, \sigma^2 + \tau^2\rho^2)$. Its likelihood depends on both $\sum_k y_k$ and $\sum_k y_k^2$, so the sample mean alone is not sufficient. Likewise, collapsing particle observations into a single consensus map can discard information needed to refine pose or class. Particle-domain inference therefore preserves both the correct image-space metric and particle-specific information required for joint nuisance inference.

## 3 Methods

Fold'EM jointly infers atomic structures and particle-specific nuisance variables from cryo-EM particle images. Each ensemble member is represented by an intermediate Pairformer state, while nuisance variables comprise pose and, for heterogeneous samples, structural class. Fold'EM first establishes a common reference frame and initial class assignments, then alternates structural inference with nuisance-variable refinement. Reconstructed volumes, when used, serve only for initialization and class-specific frame alignment; structural guidance itself operates on individual particles (Figure 1; Algorithm 1).

### 3.1 Inputs, Initialization, and Reference Frame

**Inputs.** Fold'EM takes particle images and the molecular sequence specification $\mathbf{a}$. We assume centered particles and known preprocessing quantities, including pixel/box size and CTF parameters. AlphaFold3 constructs its MSA and unguided prediction from $\mathbf{a}$; no deposited atomic model or high-resolution experimental reconstruction is supplied. The number of classes $n$ is specified for heterogeneous data and $n = 1$ otherwise. Deposited poses or class labels are used only in the corresponding known-pose or known-class evaluations.

**Pose estimation.** In the ab initio setting, we initialize poses in three stages: (i) VDAM (Kimanius et al., 2021) jointly estimates a coarse volume and poses from random initialization; (ii) these poses define a Wiener back-projected consensus density; and (iii) an AlphaFold3 prediction is docked into this density by $L_1$ map guidance. Because the cryo-EM likelihood is invariant to a global rigid transformation of structure and poses, the docked structure fixes the global $\mathrm{SE}(3)$ gauge; all subsequent poses are expressed in this frame.

**Particle embedding and classification.** When class labels are unknown, we estimate them from the particles without state-specific atomic references. We consider cryoDRGN (Zhong et al., 2021) and RECOVAR (Gilles & Singer, 2025), but find that their embeddings can become unreliable when nuisance variation is comparable to the structural differences of interest.

We therefore introduce B-RECOVAR, a bootstrapped residual-subspace estimator inspired by RECOVAR. For two states it (i) fits deviation volumes to residuals from a consensus reconstruction using the current poses and a solvent mask; (ii) refines assignments by cross-validated comparison with class reconstructions; and (iii) combines randomized subsamples and frequency bands using a label-free spectral meta-learner (Parisi et al., 2014). Additional modes handle more than two states and provide continuous variability coordinates. The ensemble margin supplies a particle-level confidence score used to prioritize reliable assignments during structural inference.

After classification, an auxiliary consensus volume is reconstructed for each class and used only to establish its reference frame: the structural prior is docked into each volume and particle poses are re-synchronized accordingly. The resulting particles, class references, and poses are passed to structural inference. Algorithm 2 and Appendix E.4 give the full procedure.

### 3.2 Ensemble Inference

Equation 2 couples the structural states $\boldsymbol{\mathcal{H}}$ to the particle nuisance variables $\boldsymbol{\eta}$. We approximately optimize the two blocks by alternating updates. Depending on the evaluation setting, known poses or classes are held fixed rather than updated.

**Step 1: structure inference given nuisance variables.** At outer iteration $r$, we hold $\boldsymbol{\eta}^{(r)} = \{(\boldsymbol{\phi}_k^{(r)}, c_k^{(r)})\}_{k=1}^K$ fixed and optimize

$$\boldsymbol{\mathcal{H}}^{(r+1)} \approx \arg\max_{\boldsymbol{\mathcal{H}}} \left[ \mathbb{E}_{\boldsymbol{\mathcal{X}} \sim p(\boldsymbol{\mathcal{X}} \mid g(\boldsymbol{\mathcal{H}}))} \log p(\boldsymbol{\mathcal{Y}} \mid \boldsymbol{\mathcal{X}}, \boldsymbol{\eta}^{(r)}) + \log p(\boldsymbol{\mathcal{H}} \mid \mathbf{a}) \right]. \tag{3}$$

The particle term factorizes as

$$\log p(\boldsymbol{\mathcal{Y}} \mid \boldsymbol{\mathcal{X}}, \boldsymbol{\eta}^{(r)}) = \sum_{k=1}^{K} \log p(\mathbf{y}_k \mid \mathbf{X}^{c_k^{(r)}}, \boldsymbol{\phi}_k^{(r)}),$$

so each ensemble member is guided by the particles currently assigned to it. The implemented particle likelihood and image formation model are described in Appendix E.3.

We solve equation 3 by an IT-Opt-style inference-time update (Maddipatla et al., 2026a), but optimize the intermediate state $\boldsymbol{\mathcal{H}}$ rather than the terminal Pairformer representation. Let $\mathcal{R}_\theta$ denote one step of the pretrained AlphaFold3 reverse diffusion process,

$$\mathbf{X}_{t-1} = \mathcal{R}_\theta(\mathbf{X}_t; g(\mathbf{H}), t, \boldsymbol{n}_t), \quad \boldsymbol{n}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{4}$$

and let $D_\theta$ denote its current denoised structure estimate. For each structural update we perform $Q$ independent reverse-diffusion runs, retaining the optimized $\boldsymbol{\mathcal{H}}$ between runs. At each reverse step,

$$\boldsymbol{\mathcal{H}}_{t-1}^q = \boldsymbol{\mathcal{H}}_t^q + \gamma_h \nabla_{\boldsymbol{\mathcal{H}}_t^q} \left[ \log p\Big(\boldsymbol{\mathcal{Y}} \mid D_\theta(\boldsymbol{\mathcal{X}}_t^q; g(\boldsymbol{\mathcal{H}}_t^q), t), \boldsymbol{\eta}^{(r)}\Big) + \lambda_p \log p(\boldsymbol{\mathcal{H}}_t^q \mid \mathbf{a}) \right], \tag{5}$$

where $\gamma_h$ denotes the step size. The updated state conditions the next reverse step. Resampling diffusion noise while retaining $\boldsymbol{\mathcal{H}}$ encourages the latent state to explain the particles across stochastic AlphaFold realizations. After the final run, $g(\boldsymbol{\mathcal{H}}^{(r+1)})$ is decoded to obtain the updated ensemble $\boldsymbol{\mathcal{X}}^{(r+1)}$.

**Step 2: nuisance inference given structure.** Holding $\boldsymbol{\mathcal{X}}^{(r+1)}$ fixed, we update

$$\boldsymbol{\eta}^{(r+1)} = \arg\max_{\boldsymbol{\eta}} \left[ \log p(\boldsymbol{\mathcal{Y}} \mid \boldsymbol{\mathcal{X}}^{(r+1)}, \boldsymbol{\eta}) + \log p(\boldsymbol{\eta}) \right]. \tag{6}$$

With factorized nuisance priors, each particle can be updated independently:

$$(\boldsymbol{\phi}_k^{(r+1)}, c_k^{(r+1)}) = \arg\max_{\boldsymbol{\phi}, c} \left[ \log p(\mathbf{y}_k \mid \mathbf{X}^{c,(r+1)}, \boldsymbol{\phi}) + \log p(\boldsymbol{\phi}, c) \right].$$

Pose refinement aligns each particle to the current structural ensemble, while class refinement assigns it to the member that best explains its image. Poses remain expressed in the docked reference frame, fixing the global $\mathrm{SE}(3)$ gauge. The two steps are then alternated. Exact coordinate ascent would monotonically increase the objective of equation 2; however, since Fold'EM uses stochastic diffusion, particle minibatches, and finite gradient updates, the coordinate-ascent view serves as an organizing principle rather than a monotonicity guarantee.

## 4 EXPERIMENTS

**Particle guidance vs. density guidance.** Given the same particles, one may guide a structure directly against their images or first reconstruct a density and guide against that map. As shown in Section 2, these are equivalent only for known poses and the experiment-induced metric $\mathbf{Q}$; with uncertain poses, a consensus map need not be sufficient. We test this on synthetic 8U4E, the largest target in Table A2, using 500 particles at SNR 0.005 (Figure A2). AlphaFold3 starts at C$\alpha$ RMSD 22.99 Å. Guidance against the ab initio consensus map improves this only to 17.95 Å, whereas Fold'EM, using the same particles while jointly refining their poses, reaches 6.88 Å (TM-score 0.809; Table A12) and is the only model whose domains follow the map envelope. Thus the coarse map remains useful for initialization, but is a poor substitute for the particles during inference.

**Synthetic homogeneous reconstruction.** We first consider 9UGB, a Pma1 conformation mispredicted by AlphaFold3 (TM-score 0.699, C$\alpha$ RMSD 8.96 Å; Figure 2A). With known poses, only 100 particles at SNR 0.005 suffice for Fold'EM to reach 3.25 Å FSC resolution, TM-score 0.965, and C$\alpha$ RMSD 1.01 Å, essentially matching the deposited model (3.39 Å; Tables A5 and A6). Increasing the particle count to $1,000$ changes these values little, consistent with the prior supplying most molecular detail while the particles select the realized conformation.

The result extends to the other synthetic targets in Table A2. From 100 particles at SNR 0.005, Fold'EM reaches 3.45, 3.80, and 4.18 Å on 9UGC, 8U4B, and 8U4E, with TM-scores 0.983, 0.979, and 0.973, respectively. In contrast, cryoDRGN does not pass 5.8 Å on the two larger targets and RELION does not pass 5 Å on any target. Ab initio, 500 particles give similar accuracy on 9UGB: 3.25 Å at SNR 0.005, TM-score 0.976, and C$\alpha$ RMSD 1.57 Å, compared with 8.93 and 27.70 Å for RELION ab initio and cryoDRGN-AI (Tables A11 and A12). The larger 8U4B/8U4E targets remain harder ab initio (5.87/11.38 Å at SNR 0.01), but Fold'EM still substantially improves their unguided predictions (TM-score 0.917/0.810 versus 0.798/0.477).

**Experimental homogeneous reconstruction.** We next study integrin $\alpha_5\beta_1$ bound to the BIIG2 Fab (EMPIAR-12552; PDB 9B9J). AlphaFold3 predicts the individual components but places the Fab at the wrong site, illustrating the difficulty of antibody–antigen interfaces where paired coevolutionary information is limited (Figure 2B). Fold'EM identifies the correct epitope from 400 particles with ab initio-derived poses, compared with 374,814 particles used in the deposited reconstruction. With fewer than 1,500 particles it reaches better than 3.5 Å FSC resolution in both the known-pose and ab initio settings (Figure 2F), and exceeds TM-score 0.9 from 400 particles (Figure 2H). Here the prior supplies the component folds, while sparse particle evidence corrects their relative arrangement and thereby reveals the epitope. The interface variable loops, however, are not recovered with atomic accuracy. Results on additional experimental targets are reported in Tables A3, A4, A9, and A10.

**Discrete heterogeneity.** We first test a balanced synthetic mixture of 9UGB/9UGC at SNR 0.01 and 0.005 (Figure A3). AlphaFold3 recovers neither state (TM-score 0.699/0.661). With known poses, Fold'EM resolves both from 250 particles per state at SNR 0.005, reaching 3.3/3.5 Å and

TM-scores $0.96/0.98$ (Tables A19 and A20); RELION remains above $7.9$ Å and cryoDRGN requires substantially more particles. Ab initio, $500$ particles per state suffice for $3.4/3.6$ Å and TM-scores $0.95/0.94$ (Tables A21 and A22); neither RELION nor cryoDRGN-AI passes $7.9$ Å. At $250$ particles per state and SNR $0.005$, 9UGB degrades to $7.6$ Å, marking the low-signal limit of joint pose, class, and structure inference.

We then withhold both poses and labels for the experimental UvrA–DNA mixture EMPIAR-13396 (Figure 3). B-RECOVAR initializes the classes, after which Fold'EM alternates structure, pose, and class refinement. The two states differ substantially in prior difficulty: AlphaFold3 predicts 9GXR ($3.54$ Å, TM-score $0.886$) but not 9GXQ ($11.69$ Å, $0.734$). Fold'EM reaches $3.58$ Å on 9GXR from $500$ particles per state and $3.02$ Å from $5{,}000$, while 9GXQ is recovered from $10{,}000$ particles per state, about $5\%$ of its deposited stack, at $4.63$ Å and TM-score $0.836$. RELION and cryoDRGN-AI remain above $8$ Å on both states; with known poses Fold'EM resolves both from $1{,}000$ particles per state (Tables A17, A18, and A15). The additional ab initio data are therefore needed mainly for the state that the prior does not already supply.

The recovered difference is functionally meaningful. UvrA recognizes damaged DNA by inserting residues Y62, Q65, and F66 into the duplex. AlphaFold3 predicts a largely canonical duplex, whereas Fold'EM recovers the wedge insertion and associated DNA distortion (Figure 3G). It does not recover the flipped-out abasic nucleotide, which was not explicitly represented as an abasic site in the input molecular specification.

**Continuous heterogeneity on IgG-1D.** Finally, we test continuous motion on the CryoBench IgG-1D dataset (Jeon et al., 2024), containing $100$ structures along a $360°$ Fab-hinge trajectory. Using cryoDRGN, RECOVAR, or B-RECOVAR as a label-free conformational coordinate, we optimize eight independent anchor states and linearly interpolate their Pairformer representations (Appendix F.3). Across all three coordinates, the $100$ decoded interpolants attain TM-score and Fourier NCC comparable to the anchor structures (Figure 4). Moreover, the leading principal components of the optimized Pairformer states preserve the ordering in hinge angle despite the anchors being optimized independently (Figure A5). Thus the conformational organization of the particles is mirrored in the AlphaFold representation, allowing a continuous trajectory to be represented by a small number of experimentally optimized states.

## 5 Conclusion

We introduced Fold'EM, an inference-time framework that infers atomic structures directly from cryo-EM particle images by guiding AlphaFold3 with particle-level evidence, bypassing density reconstruction and model building. We showed that density fitting reproduces the particle likelihood only under the experiment-induced metric and that, under pose or class uncertainty, a consensus density need not be a sufficient statistic; and that optimizing intermediate rather than terminal Pairformer states induces a learned geometry that substantially strengthens guidance. Fold'EM recovers accurate structures from fewer than $1\%$ of the particles used by conventional reconstruction with known poses, roughly $3\%$ *ab initio*, and roughly $5\%$ for heterogeneous samples, recovering features missed by the unguided prediction, such as the Integrin $\alpha_5\beta_1$–BIIG2 binding site and the distorted DNA of UvrA. Conformational organization in particle space is also reflected in the AlphaFold representation, so that continuous motions can be represented by a few anchor states. Together, these results suggest a shorter route to structure determination in which the prior supplies what is already known and the experiment determines what it does not: assembly, conformation, and interaction geometry.

**Limitations.** Fold'EM is not yet an end-to-end replacement for the cryo-EM pipeline: inputs are curated particle collections with known CTF parameters, and *ab initio* inference relies on accurate pose estimates for initialization. Components absent from the model, such as glycans, ligands or flexible regions, are treated as unexplained signal, and fine local details, such as interface loops, remain harder to recover than global assemblies and major conformational changes. Uncurated particles, unmodelled density, better pose initialization, and finer local refinement are natural directions for future work.

## AI USE STATEMENT

Generative AI tools, primarily OpenAI ChatGPT and Anthropic Claude, were used during manuscript preparation for language editing, improving clarity and organization, checking mathematical exposition, assisting with code development and debugging, and literature and reference formatting. All scientific claims, derivations, code, experimental results, interpretations, and citations were reviewed and verified by the authors. The authors take full responsibility for the content of the paper and full credit for its scientific claims.

## ETHICS STATEMENT

All authors have read and complied with the ICLR Code of Ethics. This work does not involve human subjects or sensitive personal data, and poses no identifiable risks to individuals or their security. All datasets used in this work were either synthetically generated from protein structures available in the Protein Data Bank (PDB) (Burley et al., 2017) or derived from publicly available cryo-EM particle datasets in EMPIAR (Iudin et al., 2016).

## REPRODUCIBILITY STATEMENT

All hyperparameters and implementation details are provided in Appendix E.5, and the full Fold'EM procedure is given as pseudocode in Algorithm 1. The hardware resources, model weights and software versions are listed in Appendix E.2.1. The datasets, their pre-processing and the baseline protocols are described in Appendices C.1, E.1 and E.2. The image formation model and likelihood are derived in Appendix E.3, the stereochemical losses and relaxation in Appendices E.2.2 and E.2.3, and the evaluation metrics in Appendix E.2.4. The code will be publicly released upon acceptance.

## REFERENCES

EMDB—the electron microscopy data bank. *Nucleic Acids Research*, 52(D1):D456–D465, 2024.

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

Minkyung Baek et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373:871–876, 2021. doi: 10.1126/science.abj8754.

Stephen K Burley, Helen M Berman, Gerard J Kleywegt, John L Markley, Haruki Nakamura, and Sameer Velankar. Protein data bank (PDB): the single global macromolecular structure archive. *Protein Crystallography: Methods and Protocols*, pp. 627–641, 2017.

Alisia Fadini, Minhuan Li, Airlie J McCoy, Suresh Banjara, Hiroki Okumura, Eve Napier, Pietro Fontana, Amir R Khan, Luca Jovine, Thomas C Terwilliger, et al. AlphaFold as a prior: experimental structure determination conditioned on a pretrained neural network. *Nature Methods*, pp. 785–795, 2026.

Rafael Fernandez-Leiro and Sjors H. W. Scheres. Unravelling biological macromolecules with cryo-electron microscopy. *Nature*, 537:339–346, 2016. doi: 10.1038/nature19948.

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.

Viktor Hornak, Robert Abel, Asim Okur, Bentley Strockbine, Adrian Roitberg, and Carlos Simmerling. Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins: Structure, Function, and Bioinformatics*, 65(3):712–725, 2006.

Andrii Iudin, Paul K Korir, José Salavert-Torres, Gerard J Kleywegt, and Ardan Patwardhan. EMPIAR: a public archive for raw electron microscopy image data. *Nature Methods*, 13(5):387–388, 2016.

Kiarash Jamali, Lukas Käll, Rui Zhang, Alan Brown, Dari Kimanius, and Sjors HW Scheres. Automated model building and protein identification in cryo-EM maps. *Nature*, 628(8007): 450–457, 2024.

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*, volume 37, pp. 89468–89512, 2024.

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

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

Dari Kimanius, Liyi Dong, Grigory Sharov, Takanori Nakane, and Sjors HW Scheres. New tools for automated cryo-EM single-particle analysis in RELION-4.0. *Biochemical Journal*, 478(24): 4169–4185, 2021.

Dari Kimanius et al. Data-driven regularization lowers the size barrier of cryo-EM structure determination. *Nature Methods*, 21:1216–1221, 2024. doi: 10.1038/s41592-024-02304-8.

Diederik P Kingma and Jimmy Ba. Adam: a method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Harold W Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955.

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.

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

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

Dorothee Liebschner, Pavel V Afonine, Matthew L Baker, Gábor Bunkóczi, Vincent B Chen, Tristan I Croll, Bradley Hintze, L-W Hung, Swati Jain, Airlie J McCoy, et al. Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallographica Section D: Structural Biology*, 75(10):861–877, 2019.

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

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*, 2026b. doi: 10.1038/s41587-026-03166-5.

M.-E. Mäeots and R. I. Enchev. Structural dynamics: review of time-resolved cryo-EM. *Acta Crystallographica Section D: Structural Biology*, 78:927–935, 2022. doi: 10.1107/S2059798322006155.

Abbas Ourmazd. Cryo-EM, XFELs and the structure conundrum in structural biology. *Nature Methods*, 16:941–944, 2019. doi: 10.1038/s41592-019-0587-4.

Fabio Parisi, Francesco Strino, Boaz Nadler, and Yuval Kluger. Ranking and combining multiple predictors without labeled data. *Proceedings of the National Academy of Sciences*, 111(4):1253–1258, 2014.

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

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.

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*, volume 38, pp. 10175–10203, 2026.

Alexis Rohou and Nikolaus Grigorieff. CTFFIND4: fast and accurate defocus estimation from electron micrographs. *Journal of Structural Biology*, 192(2):216–221, 2015.

Peter B Rosenthal and Richard Henderson. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *Journal of Molecular Biology*, 333(4):721–745, 2003.

Sjors H. W. Scheres. RELION: implementation of a bayesian approach to cryo-EM structure determination. *Journal of Structural Biology*, 180:519–530, 2012. doi: 10.1016/j.jsb.2012.09.006.

RH Wade. A brief look at imaging and contrast transfer. *Ultramicroscopy*, 46(1–4):145–156, 1992.

Christopher J Williams, Jeffrey J Headd, Nigel W Moriarty, Michael G Prisant, Lizbeth L Videau, Lindsay N Deis, Vishal Verma, Daniel A Keedy, Bradley J Hintze, Vincent B Chen, et al. MolProbity: more and better reference data for improved all-atom structure validation. *Protein Science*, 27(1):293–315, 2018.

Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004.

Yuxuan Zhang, Chengyue Gong, Hanyu Zhang, Wenzhi Ma, Zhenyu Liu, Xinshi Chen, Jiaqi Guan, Lan Wang, Yanping Yang, Yu Xia, and Wenzhi Xiao. Protenix-v1: toward high-accuracy open-source biomolecular structure prediction. *bioRxiv*, 2026. doi: 10.64898/2026.02.05.703733. URL https://www.biorxiv.org/content/early/2026/02/22/2026.02.05.703733.1.

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.

Jasenko Zivanov, Takanori Nakane, Björn O Forsberg, Dari Kimanius, Wim JH Hagen, Erik Lindahl, and Sjors HW Scheres. New tools for automated high-resolution cryo-EM structure determination in RELION-3. *eLife*, 7:e42166, 2018.

# A Overview of the appendix

The appendix is organized as follows.

- **Appendix B: Additional figures.** The comparison of Fold'EM guidance with inference-time optimization and coordinate-space guidance (Figure A1), the additional heterogeneity figures (Figure A3), the evaluation of B-RECOVAR class assignment and its confidence score (Figure A4), and the latent geometry of the optimized Pairformer states on IgG-1D (Figure A5).
- **Appendix C: Additional tables.** The experimental and synthetic datasets (Appendix C.1), and the full quantitative results for homogeneous reconstruction with known poses (Appendix C.2), homogeneous *ab initio* reconstruction (Appendix C.3), heterogeneous reconstruction (Appendix C.4), and continuous heterogeneity on IgG-1D (Appendix C.5).
- **Appendix D: Algorithms.** Pseudocode for Fold'EM (Algorithm 1) and for the residual-subspace estimator of B-RECOVAR (Algorithm 2).
- **Appendix E: Additional details.** Data pre-processing and baselines (Appendices E.1 and E.2), software and hardware (Appendix E.2.1), the stereochemical losses and relaxation (Appendices E.2.2 and E.2.3), the evaluation metrics (Appendix E.2.4), the image formation model used to render particles from atomic structures, the contrast transfer function, and likelihood computation (Appendix E.3), the B-RECOVAR class-assignment method (Appendix E.4) and the hyperparameters and implementation details (Appendix E.5).
- **Appendix F: Additional experiments.** The comparison of Fold'EM with other experiment-guided AlphaFold methods (Appendix F.1), the protocol of the heterogeneous experiments on experimental UvrA data (Appendix F.2) and the protocol and full results of the continuous-heterogeneity experiment on IgG-1D (Appendix F.3).
- **Appendix G: Related works.** An extended discussion of related work.

## B ADDITIONAL FIGURES

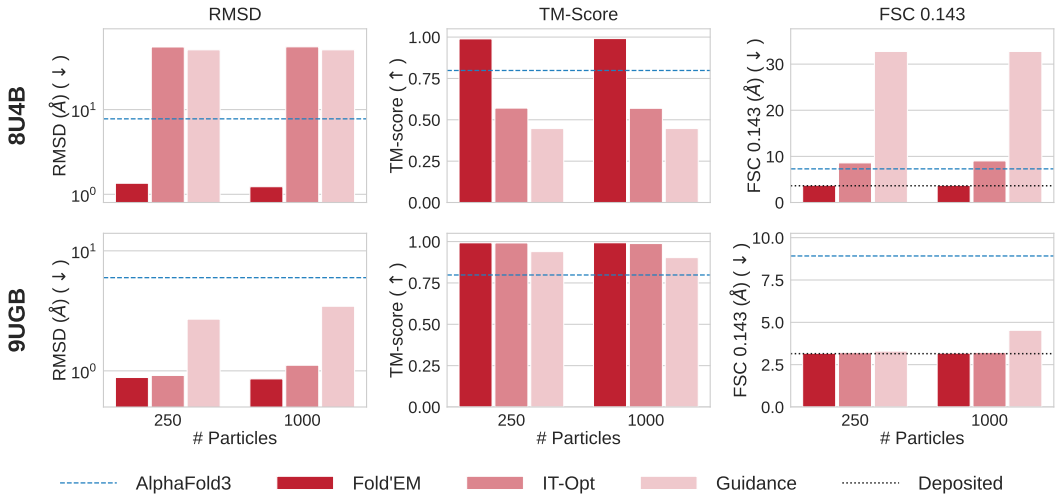


Figure A1: **Optimizing the intermediate Pairformer state improves particle-level guidance.** We compare three ways of steering AlphaFold3 with the same particle likelihood, poses and particle subsets: Fold'EM, which optimizes the intermediate Pairformer state $\mathcal{H}$ and passes the updates through the final $\ell$ Pairformer blocks (Section 2); IT-Opt (Maddipatla et al., 2026a), which optimizes the terminal embedding $\mathcal{Z}$ directly under an isotropic prior; and coordinate-space guidance of the reverse diffusion process (Maddipatla et al., 2026b). *Rows:* synthetic `8U4B` and `9UGB` from Table A2 at SNR $0.005$ with known poses, using $250$ and $1000$ particles. *Columns:* C$\alpha$-RMSD ($\downarrow$, log scale) and TM-score ($\uparrow$) against the deposited PDB model; the dashed line marks the AlphaFold3 prediction. Across both examples, Fold'EM improves drastically over both alternatives, reaching near-atomic agreement with the deposited model. The plot confirms that the learned geometry induced by optimizing $\mathcal{H}$ keeps particle-level guidance stable at low SNR and on large assemblies.

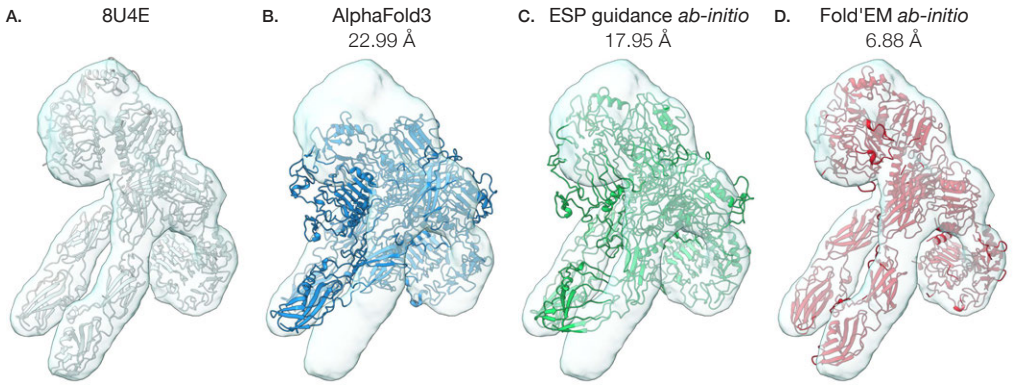


Figure A2: **Guiding with the particles recovers the structure, whereas guiding with their reconstructed ESP fails.** Ab initio inference on synthetic `8U4E` (3,304 residues) from 500 particles at SNR 0.005. The grey surface is the consensus density reconstructed from these particles at the estimated poses, and the numbers are C$\alpha$ RMSD to the deposited model. (A) The deposited model. (B) AlphaFold3 misplaces the domains of the assembly. (C) ESP guidance optimizes the same Pairformer state with the same particles, poses and optimizer, but replaces the particle likelihood by a loss against the consensus density, and improves the prediction without recovering the arrangement. (D) Fold'EM, guided by the particle likelihood, recovers it, with a TM-score of 0.81 against 0.48 for AlphaFold3. As argued in Section 2, a density loss reproduces the particle likelihood only under the experiment-induced metric, and with estimated poses the consensus density is not a sufficient statistic for the particles, so the information needed to refine the poses together with the structure is lost.

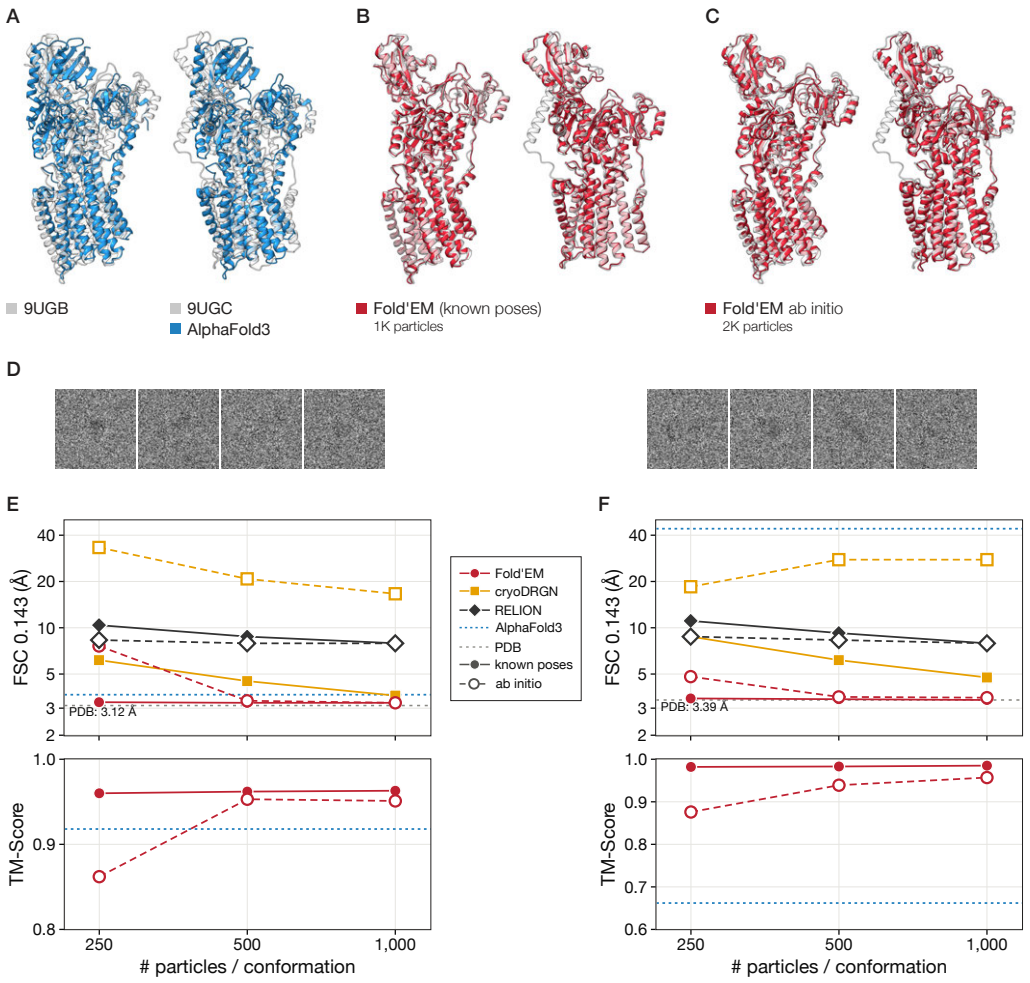


Figure A3: **Fold'EM resolves both states of a synthetic 9UGB/9UGC mixture from 250 particles per state with known poses, and from 500 ab initio.** Pma1 is the plasma-membrane proton pump of yeast, a P-type ATPase whose cytosolic A-, N- and P-domains rearrange along the catalytic cycle; 9UGB is the activated (E2P) state and 9UGC the autoinhibited state. Particles are rendered from the two deposited conformations, approximately 9 Å apart in C$\alpha$ RMSD, and mixed in equal proportion at SNR 0.005. In every panel, 9UGB is on the left and 9UGC on the right. (A) AlphaFold3 (blue) over the deposited models (grey) recovers neither state (TM-score 0.70 and 0.66): the transmembrane domain is predicted correctly and the error is confined to the cytosolic domains, whose arrangement depends on the state of the cycle. (B) Fold'EM (red) with known poses resolves both from the mixed stack of 1,000 particles (500 per state). (C) Fold'EM ab initio resolves both from 2,000 particles (1,000 per state). (D) Representative particles of each state. (E, F) FSC resolution (top) and TM-score against the deposited model (bottom) as a function of the number of particles per state, for 9UGB (E) and 9UGC (F); filled markers denote known poses and open markers ab initio, the dashed grey line is the resolution of the deposited model and the dotted blue line the AlphaFold3 prediction.

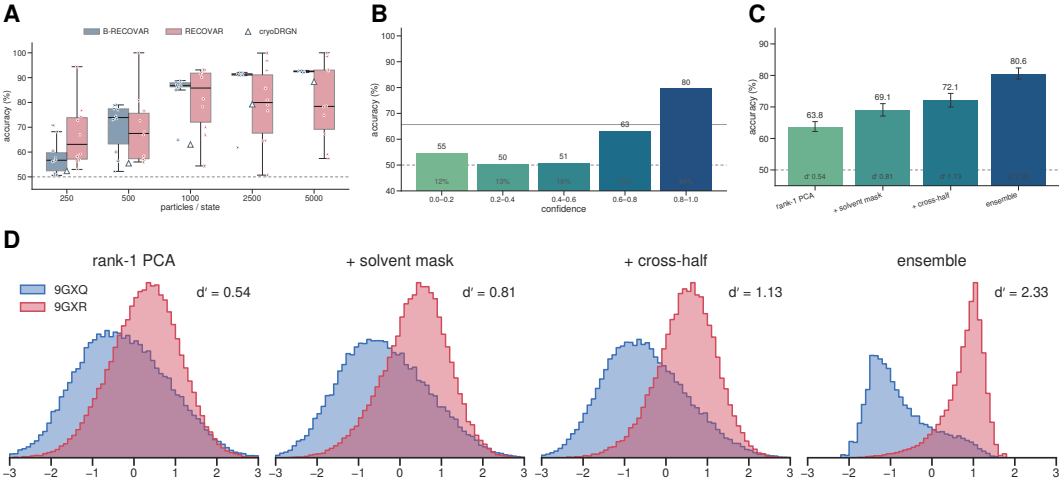


Figure A4: **B-RECOVAR assigns particles to conformational states more accurately and more reliably than RECOVAR and cryoDRGN at low particle counts, and its confidence score orders the assignments by their reliability.** Label-free class assignment on the UvrA protein–DNA complex (EMPIAR-13396, 9GXQ vs. 9GXR), scored against the deposited labels. **(A)** Accuracy as a function of the number of particles per state, from 250 to 5,000, that is, 0.13–2.6% of the particles deposited for each state. Boxes summarize ten random draws per size, shared by B-RECOVAR and RECOVAR; triangles mark cryoDRGN. The RECOVAR estimator is ill-conditioned at these sample sizes and its accuracy ranges from chance to near-perfect between draws of the same size; B-RECOVAR trails it at 250 particles but is both more accurate and markedly more stable from 500 particles onward, and cryoDRGN remains below B-RECOVAR throughout. **(B)** Accuracy of B-RECOVAR stratified by its confidence score $\kappa_k$, the normalized margin of the ensemble vote (Appendix E.4). The percentage in each bar is the fraction of particles in that bin; the solid line is the accuracy over all particles and the dashed line chance. Accuracy increases monotonically with confidence, so the score ranks the particles by the reliability of their label; Fold'EM uses this ordering to guide each ensemble member with the more confident half of its particles (Table A24). **(C, D)** Ablation of the estimator at 1000 particles per state. The base estimator is a single rank-one fit to the consensus-subtracted residuals, a one-dimensional RECOVAR fit without noise correction, and is improved in turn by restricting the mode to a solvent mask, by the cross-validated refinement of the labels, and by ensembling replicates over particle subsamples, frequency bands and initializations. **(C)** Accuracy of each stage, with error bars over repeated draws. **(D)** Distribution of the per-particle class score at each stage for the particles of either state, with their separation $d'$, the difference of the two means in units of the pooled standard deviation. Every step widens the separation between the states, and the ensemble accounts for the largest gain.

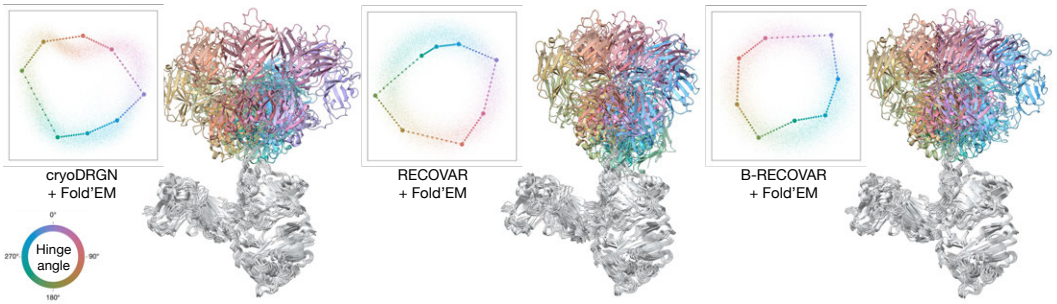


Figure A5: **Optimized Pairformer states preserve the hinge-angle ordering for all three conformational coordinates.** We show the geometry and the generated structures for each of the three particle classifiers: cryoDRGN, RECOVAR and B-RECOVAR. Left of each panel: the two leading principal components of the optimized Pairformer states, with the eight anchors as large dots and the interpolated states as small dots, overlaid on the per-particle coordinates of the corresponding method mapped into the same plane; all are coloured by the hinge angle of the state they represent, as indicated by the colour wheel. Right of each panel: a subset of 10 structures decoded from the interpolated states, superimposed on the fixed body of the antibody and coloured by hinge angle (top), and in white to show the sweep of the moving Fab about the hinge (bottom). In all three cases the anchors, ordered by hinge angle, form a simple closed ring in the representation and the decoded structures trace the full rotation of the Fab.

# C ADDITIONAL TABLES

## C.1 DATASETS

Table A1: **Experimental cryo-EM datasets from EMPIAR (Iudin et al., 2016).** For each PDB entry, we report the corresponding EMPIAR and EMDB entries (wwp, 2024), reported resolution, molecular weight (kDa), and number of particles in original dataset.

| PDB ID | EMPIAR ID | EMD ID | Resolution(Å) | # Particles | Protein size (kDa) |
|---|---|---|---|---|---|
| 9B9J | 12552 | 44386 | 2.60 | 374814 | 115.11 |
| 9B77 | 12002 | 44302 | 2.74 | 490807 | 141.72 |
| 8PWH | 11665 | 17993 | 3.17 | 708609 | 162.75 |
| 9GXR | 13396 | 51671 | 3.04 | 197058 | 162.70 |
| 9GXQ | 13396 | 51672 | 2.74 | 195233 | 162.70 |

Table A2: **Synthetic datasets constructed using PDB-deposited structures (Burley et al., 2017).** For each entry, we report the corresponding EMDB entry (wwp, 2024), reported resolution, molecular weight (kDa), and amino acid sequence length.

| PDB ID | EMD ID | Resolution(Å) | Seq Length | Protein size (kDa) |
|---|---|---|---|---|
| 9UGB | 64135 | 3.25 | 818 | 88.97 |
| 9UGC | 64136 | 3.52 | 818 | 99.62 |
| 8U4B | 41877 | 3.80 | 2764 | 312.70 |
| 8U4E | 41880 | 4.20 | 3304 | 373.10 |
| 1HZH | CryoBench (IgG-1D) | 2.70 | 1331 | 146.96 |

### C.2 HOMOGENEOUS RECONSTRUCTION WITH KNOWN POSES

Table A3: **Resolution at Fourier shell correlation (FSC)** $= 0.143$ **(↓, Å) on the experimental cryo-EM datasets from Table A1.** For Fold'EM, AlphaFold3 and PDB structures, FSC is computed using `phenix.model_map`. All baselines at different fraction of the original dataset 500 particles with known poses and are evaluated against the deposited EMD map. Entries highlighted in **bold** indicate the best result (excluding PDB).

| Fraction of dataset | Method | 9B9J | 9B77 | 8PWH | 9GXQ | 9GXR |
|---|---|---|---|---|---|---|
| 1/1000 | Fold'EM | **3.73** | **2.88** | **6.21** | **3.08** | **3.18** |
| | cryoDRGN | 8.40 | 5.70 | 7.56 | 9.87 | 4.33 |
| | RELION | 21.94 | 11.48 | 14.12 | 14.15 | 17.92 |
| | AlphaFold3 | 21.90 | 5.96 | 64.96 | 11.69 | 3.54 |
| | PDB | 2.38 | 2.49 | 3.06 | 2.64 | 2.89 |
| 1/500 | Fold'EM | **3.73** | **2.85** | **6.63** | **3.02** | **3.20** |
| | cryoDRGN | 7.41 | 4.87 | 7.55 | 9.60 | 3.90 |
| | RELION | 17.95 | 11.48 | 15.47 | 14.15 | 14.93 |
| | AlphaFold3 | 21.90 | 5.96 | 64.96 | 11.69 | 3.54 |
| | PDB | 2.38 | 2.49 | 3.06 | 2.64 | 2.89 |
| 1/250 | Fold'EM | **3.35** | **2.73** | **6.63** | **3.95** | **3.09** |
| | cryoDRGN | 7.04 | 4.87 | **6.63** | 9.27 | 3.73 |
| | RELION | 15.19 | 11.07 | 12.99 | 12.22 | 12.22 |
| | AlphaFold3 | 21.90 | 5.96 | 64.96 | 11.69 | 3.54 |
| | PDB | 2.38 | 2.49 | 3.06 | 2.64 | 2.89 |
| 1/100 | Fold'EM | **3.50** | **2.73** | 7.06 | **3.09** | **3.26** |
| | cryoDRGN | 6.71 | 3.92 | **5.16** | 9.27 | 3.73 |
| | RELION | 12.74 | 7.21 | 10.15 | 10.75 | 9.60 |
| | AlphaFold3 | 21.90 | 5.96 | 64.96 | 11.69 | 3.54 |
| | PDB | 2.38 | 2.49 | 3.06 | 2.64 | 2.89 |
| 1/50 | Fold'EM | **3.46** | **2.75** | 6.13 | **3.36** | **3.45** |
| | cryoDRGN | 6.46 | 3.87 | **5.01** | 8.67 | 3.58 |
| | RELION | 12.34 | 6.74 | 9.02 | 7.07 | 7.07 |
| | AlphaFold3 | 21.90 | 5.96 | 64.96 | 11.69 | 3.54 |
| | PDB | 2.38 | 2.49 | 3.06 | 2.64 | 2.89 |

Table A4: **Quantitative evaluation of C$\alpha$ RMSD (↓, Å) and TM-score (↑) on the experimental cryo-EM datasets from Table A1.** Fold'EM structures are obtained using different fractions of the original dataset. Entries highlighted in **bold** indicate the best result.

| | C$\alpha$ RMSD (↓, Å) | | | | | TM-score (↑) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Fraction of dataset | 9B9J | 9B77 | 8PWH | 9GXQ | 9GXR | 9B9J | 9B77 | 8PWH | 9GXQ | 9GXR |
| 1/1000 | 1.96 | 1.90 | 4.340 | 2.20 | 1.87 | 0.969 | 0.972 | 0.900 | 0.877 | 0.899 |
| 1/500 | 1.97 | 1.77 | 4.989 | **1.99** | 1.39 | 0.970 | 0.979 | 0.897 | **0.893** | 0.902 |
| 1/250 | **1.88** | **1.71** | 4.597 | 2.48 | **1.23** | **0.972** | **0.981** | 0.885 | 0.880 | **0.904** |
| 1/100 | 1.92 | 1.74 | 5.349 | 2.02 | 1.35 | 0.971 | 0.980 | 0.884 | **0.893** | 0.903 |
| 1/50 | **1.88** | **1.71** | **4.174** | 2.09 | 1.47 | **0.972** | **0.981** | **0.909** | 0.885 | 0.897 |
| AlphaFold3 | 21.90 | 3.87 | 33.16 | 5.96 | 2.01 | 0.776 | 0.916 | 0.397 | 0.734 | 0.886 |

Table A5: **Resolution at Fourier shell correlation (FSC)** $= 0.143$ **(↓) on the synthetic cryo-EM datasets from Table A2.** For Fold'EM, AlphaFold3 and PDB structures, FSC is computed using `phenix.model_map`. Fold'EM structures are generated using $100, 250, 500, \& 1000$ particles at SNR $0.01$ and $0.005$ with known poses and are evaluated against the deposited EMD map. Entries highlighted in **bold** indicate the best result (excluding PDB).

| # particles | Method | 9UGB | | 9UGC | | 8U4B | | 8U4E | |
|---|---|---|---|---|---|---|---|---|---|
| | | SNR 0.01 | SNR 0.005 | SNR 0.01 | SNR 0.005 | SNR 0.01 | SNR 0.005 | SNR 0.01 | SNR 0.005 |
| 100 | Fold'EM | **3.24** | **3.25** | **3.59** | **3.45** | **3.76** | **3.80** | **4.04** | **4.18** |
| | cryoDRGN | 3.63 | 3.95 | 3.81 | 3.96 | 7.85 | 9.80 | 9.47 | 10.03 |
| | RELION | 9.79 | 9.79 | 9.79 | 9.79 | 17.78 | 16.84 | 19.20 | 18.29 |
| | AlphaFold3 | 8.21 | | 19.29 | | 7.29 | | 8.84 | |
| | PDB | 3.39 | | 3.47 | | 3.50 | | 3.67 | |
| 250 | Fold'EM | **3.22** | **3.22** | **3.39** | **3.43** | **3.75** | **3.73** | **3.99** | **3.89** |
| | cryoDRGN | 3.39 | 3.55 | 3.59 | 3.63 | 7.00 | 7.16 | 7.23 | 7.56 |
| | RELION | 7.92 | 9.24 | 6.66 | 8.76 | 10.32 | 11.03 | 13.71 | 16.70 |
| | AlphaFold3 | 8.21 | | 19.29 | | 7.29 | | 8.84 | |
| | PDB | 3.39 | | 3.47 | | 3.50 | | 3.67 | |
| 500 | Fold'EM | **3.23** | **3.26** | **3.39** | **3.39** | **3.70** | **3.70** | **3.96** | **3.93** |
| | cryoDRGN | 3.32 | 3.35 | 3.55 | 3.59 | 5.99 | 6.88 | 6.77 | 7.40 |
| | RELION | 5.94 | 6.40 | 5.94 | 6.40 | 10.00 | 10.32 | 12.00 | 12.39 |
| | AlphaFold3 | 8.21 | | 19.29 | | 7.29 | | 8.84 | |
| | PDB | 3.39 | | 3.47 | | 3.50 | | 3.67 | |
| 1000 | Fold'EM | **3.22** | **3.22** | **3.38** | **3.43** | **3.69** | **3.72** | **3.91** | **4.00** |
| | cryoDRGN | 3.25 | 3.32 | 3.55 | 3.55 | 5.81 | 5.91 | 6.06 | 6.95 |
| | RELION | 5.04 | 5.74 | 5.04 | 6.16 | 8.65 | 10.00 | 10.00 | 10.53 |
| | AlphaFold3 | 8.21 | | 19.29 | | 7.29 | | 8.84 | |
| | PDB | 3.39 | | 3.47 | | 3.50 | | 3.67 | |

Table A6: **Quantitative evaluation of C$\alpha$ RMSD (↓, Å) and TM-score (↑) on the synthetic cryo-EM datasets from Table A2.** Fold'EM structures are obtained using $100, 250, 500, \& 1000$ particles with known poses at SNR $0.01$ and $0.005$; AlphaFold3 is the unguided baseline. Scores are computed with respect to the experimentally resolved structures deposited in the PDB. Entries highlighted in **bold** indicate the best result.

| Metric | # particles | 9UGB | | 9UGC | | 8U4B | | 8U4E | |
|---|---|---|---|---|---|---|---|---|---|
| | | SNR 0.01 | SNR 0.005 | SNR 0.01 | SNR 0.005 | SNR 0.01 | SNR 0.005 | SNR 0.01 | SNR 0.005 |
| C$\alpha$ RMSD (↓, Å) | 100 | 1.05 | 1.01 | 2.78 | 1.45 | 1.681 | 1.907 | 1.769 | 2.116 |
| | 250 | 0.88 | 0.88 | 1.34 | **1.37** | 1.362 | 1.350 | 1.798 | 1.512 |
| | 500 | **0.81** | 1.12 | 1.44 | 1.41 | 1.248 | **1.227** | **1.503** | **1.479** |
| | 1000 | 0.84 | **0.86** | **1.33** | 1.44 | **1.132** | 1.228 | 1.686 | 1.711 |
| | AlphaFold3 | 8.96 | | 11.22 | | 7.780 | | 26.670 | |
| TM-score (↑) | 100 | 0.964 | 0.965 | 0.939 | 0.983 | 0.9832 | 0.9785 | 0.9810 | 0.9731 |
| | 250 | 0.967 | **0.967** | **0.987** | **0.984** | 0.9887 | 0.9890 | 0.9807 | 0.9861 |
| | 500 | **0.968** | 0.962 | 0.985 | **0.984** | 0.9906 | 0.9908 | **0.9862** | **0.9866** |
| | 1000 | 0.967 | **0.967** | **0.987** | 0.983 | **0.9923** | **0.9909** | 0.9828 | 0.9822 |
| | AlphaFold3 | 0.699 | | 0.661 | | 0.7981 | | 0.4772 | |

Table A7: **MolProbity evaluation of Fold'EM structures over synthetic targets with known poses**. We report median MolProbity score (↓), clash score (↓), and Ramachandran outlier percentage (↓), computed using `phenix.molprobity` (Williams et al., 2018) over all relaxed structures in Table A6.

| MolProbity statistic | 9UGB | | 9UGC | | 8U4B | | 8U4E | |
|---|---|---|---|---|---|---|---|---|
| | **SNR 0.01** | **SNR 0.005** | **SNR 0.01** | **SNR 0.005** | **SNR 0.01** | **SNR 0.005** | **SNR 0.01** | **SNR 0.005** |
| MolProbity score | 1.01 | 1.05 | 1.36 | 1.44 | 1.35 | 1.35 | 1.56 | 1.54 |
| Clashscore | 2.29 | 2.61 | 3.38 | 2.85 | 1.59 | 1.49 | 3.01 | 3.02 |
| Ramachandran outliers | 0.00 | 0.00 | 0.28 | 0.55 | 1.49 | 1.40 | 1.72 | 1.26 |

Table A8: **MolProbity evaluation of Fold'EM structures over real targets with known poses**. We report median MolProbity score (↓), clash score (↓), and Ramachandran outlier percentage (↓), computed using `phenix.molprobity` (Williams et al., 2018) over all structures in Table A4.

| MolProbity statistic | 9B9J | 9B77 | 8PWH | 9GXQ | 9GXR |
|---|---|---|---|---|---|
| MolProbity score | 0.84 | 0.81 | 1.71 | 1.08 | 1.07 |
| Clashscore | 0.38 | 0.60 | 5.42 | 1.21 | 0.99 |
| Ramachandran outliers | 0.10 | 0.24 | 0.43 | 0.16 | 0.40 |

### C.3 HOMOGENEOUS *ab initio* RECONSTRUCTION

Table A9: **Resolution at Fourier shell correlation (FSC)** $= 0.143$ **(↓) on the experimental cryo-EM datasets from Table A1, with poses inferred *ab initio*.** For Fold'EM *ab initio*, AlphaFold3 and PDB structures, FSC is computed using `phenix.model_map`. The number of particles used for the experiment is given below each identifier. Entries highlighted in **bold** indicate the best result.

| Method<br># particles | 9B9J<br>1000 | 9B77<br>1000 | 8PWH<br>5000 | 9GXQ<br>5000 | 9GXR<br>5000 |
|---|---|---|---|---|---|
| Fold'EM | **3.52** | **2.85** | 10.83 | **4.48** | **3.22** |
| cryoDRGN-AI | 35.90 | 31.01 | 32.80 | 24.44 | 26.88 |
| RELION (ab initio) | 11.62 | 7.95 | **9.84** | 6.25 | 4.98 |
| AlphaFold3 | 21.90 | 4.40 | 64.96 | 11.70 | 3.50 |
| PDB | 2.38 | 2.49 | 3.06 | 2.64 | 2.89 |

Table A10: **Quantitative evaluation of TM-score (↑) and $C\alpha$-RMSD (↓) on the experimental cryo-EM datasets from Table A1, with poses inferred *ab initio*.** The number of particles used is given below each identifier, and scores are computed with respect to the experimentally resolved structures deposited in the PDB. Entries highlighted in **bold** indicate the best result.

| | C$\alpha$ RMSD (↓, Å) | | | | | TM-score (↑) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Method<br># particles | 9B9J<br>1000 | 9B77<br>1000 | 8PWH<br>5000 | 9GXQ<br>5000 | 9GXR<br>5000 | 9B9J<br>1000 | 9B77<br>1000 | 8PWH<br>5000 | 9GXQ<br>5000 | 9GXR<br>5000 |
| Fold'EM | **2.49** | **2.25** | **18.91** | **2.87** | **1.54** | **0.956** | **0.967** | **0.602** | **0.857** | **0.898** |
| AlphaFold3 | 21.90 | 3.87 | 33.16 | 5.96 | 2.01 | 0.776 | 0.916 | 0.397 | 0.734 | 0.886 |

Table A11: **Resolution at Fourier shell correlation (FSC)** $= 0.143$ **(↓) on the synthetic cryo-EM datasets from Table A2, with poses inferred *ab initio*.** For Fold'EM *ab initio*, AlphaFold3 and PDB structures, FSC is computed using `phenix.model_map`. All methods use 500 particles at SNR 0.01 and 0.005. Entries highlighted in **bold** indicate the best result (excluding PDB).

| # particles | Method | 9UGB | | 9UGC | | 8U4B | | 8U4E | |
|---|---|---|---|---|---|---|---|---|---|
| | | **SNR 0.01** | **SNR 0.005** | **SNR 0.01** | **SNR 0.005** | **SNR 0.01** | **SNR 0.005** | **SNR 0.01** | **SNR 0.005** |
| 500 | Fold'EM | **3.47** | **3.25** | **3.59** | **3.91** | **5.87** | **7.23** | 11.38 | 11.53 |
| | cryoDRGN-AI | 12.80 | 27.70 | 33.30 | 33.20 | 22.59 | 38.40 | 38.40 | 38.40 |
| | RELION (ab initio) | 7.23 | 8.93 | 6.93 | 7.56 | 10.97 | 16.00 | 11.64 | 15.36 |
| | AlphaFold3 | 8.21 | | 19.29 | | 7.29 | | **8.84** | |
| | PDB | 3.39 | | 3.47 | | 3.5 | | 3.67 | |

Table A12: **Quantitative evaluation of TM-score (↑) and $C\alpha$-RMSD (↓, Å) on the synthetic cryo-EM datasets from Table A2, with poses inferred *ab initio*.** Fold'EM *ab initio* results are obtained using 500 particles at SNR 0.01 and 0.005, and scores are computed with respect to the experimentally resolved structures deposited in the PDB. Entries highlighted in **bold** indicate the best result.

| Metric | Method | 9UGB | | 9UGC | | 8U4B | | 8U4E | |
|---|---|---|---|---|---|---|---|---|---|
| | | **SNR 0.01** | **SNR 0.005** | **SNR 0.01** | **SNR 0.005** | **SNR 0.01** | **SNR 0.005** | **SNR 0.01** | **SNR 0.005** |
| C$\alpha$ RMSD (↓, Å) | Fold'EM | **1.78** | **1.57** | **2.35** | **3.07** | **3.948** | **3.941** | **7.684** | **6.885** |
| | AlphaFold3 | 8.96 | | 11.22 | | 7.780 | | 26.670 | |
| TM-score (↑) | Fold'EM | **0.969** | **0.976** | **0.948** | **0.917** | **0.9174** | **0.9182** | **0.810** | **0.809** |
| | AlphaFold3 | 0.699 | | 0.661 | | 0.7981 | | 0.4772 | |

Table A13: **MolProbity evaluation of the *ab initio* Fold'EM structures in Table A12**. For each relaxed structure, we report MolProbity score (↓), clash score (↓), and Ramachandran outlier percentage (↓), computed using `phenix.molprobity` (Williams et al., 2018).

| Statistic | 9UGB | | 9UGC | | 8U4B | | 8U4E | |
|---|---|---|---|---|---|---|---|---|
| | **SNR 0.01** | **SNR 0.005** | **SNR 0.01** | **SNR 0.005** | **SNR 0.01** | **SNR 0.005** | **SNR 0.01** | **SNR 0.005** |
| MolProbity score | 1.18 | 1.15 | 1.25 | 1.15 | 1.50 | 1.53 | 1.58 | 1.79 |
| Clashscore | 3.72 | 3.64 | 4.83 | 3.56 | 1.21 | 2.35 | 1.56 | 2.21 |
| Ramachandran outliers | 0.00 | 0.25 | 0.12 | 0.00 | 1.26 | 0.71 | 0.98 | 0.66 |

Table A14: **MolProbity evaluation of the *ab initio* Fold'EM structures in Table A10**. For each relaxed structure, we report MolProbity score (↓), clash score (↓), and Ramachandran outlier percentage (↓), computed using `phenix.molprobity` (Williams et al., 2018). Particle counts are as in Table A9.

| Statistic | 9B9J | 9B77 | 8PWH | 9GXQ | 9GXR |
|---|---|---|---|---|---|
| MolProbity score | 0.93 | 0.80 | 1.48 | 1.22 | 1.11 |
| Clashscore | 0.82 | 1.04 | 4.01 | 2.52 | 1.57 |
| Ramachandran outliers | 0.19 | 0.00 | 0.27 | 0.16 | 0.24 |

### C.4 HETEROGENEOUS RECONSTRUCTION

Table A15: **Resolution at Fourier shell correlation (FSC) $= 0.143$ (↓) on the heterogeneous experimental dataset EMPIAR-13396, with known poses.** The two conformational states `9GXQ` and `9GXR` are resolved from the same mixed particle stack, and the sweep is over the number of particles per state. For Fold'EM, AlphaFold3 and PDB structures, FSC is computed using `phenix.model_map`. Entries highlighted in **bold** indicate the best result.

| # particles per conf. | Method | 9GXQ | 9GXR |
|---|---|---|---|
| 500 | Fold'EM | 6.25 | **3.02** |
| | RELION | 8.67 | 8.40 |
| | cryoDRGN | **5.60** | 4.89 |
| 1000 | Fold'EM | **3.09** | **2.95** |
| | RELION | 6.89 | 10.75 |
| | cryoDRGN | 4.34 | 4.41 |
| 2000 | Fold'EM | **3.16** | **3.01** |
| | RELION | 5.97 | 4.80 |
| | cryoDRGN | 4.27 | 4.10 |
| 5000 | Fold'EM | **3.13** | **2.99** |
| | RELION | 7.47 | 7.07 |
| | cryoDRGN | 4.10 | 4.03 |
| 10000 | Fold'EM | **3.16** | **2.99** |
| | RELION | 6.56 | 6.40 |
| | cryoDRGN | 4.00 | 3.99 |
| AlphaFold3 | | 11.69 | 3.54 |
| PDB | | 2.64 | 2.89 |

Table A16: **Quantitative evaluation of C$\alpha$ RMSD (↓, Å) and TM-score (↑) on the heterogeneous experimental dataset EMPIAR-13396, with known poses.** Fold'EM resolves the two conformational states `9GXQ` and `9GXR` from the same mixed particle stack, and the sweep is over the number of particles per state. Entries highlighted in **bold** indicate the best result.

| | C$\alpha$ RMSD (↓, Å) | | TM-score (↑) | |
|---|---|---|---|---|
| # particles per conf. | 9GXQ | 9GXR | 9GXQ | 9GXR |
| 500 | 3.30 | 1.85 | 0.840 | 0.890 |
| 1000 | 1.63 | **1.45** | 0.891 | 0.899 |
| 2000 | 1.76 | 1.71 | 0.886 | 0.894 |
| 5000 | **1.48** | **1.45** | **0.893** | **0.900** |
| 10000 | 1.82 | 1.73 | 0.885 | 0.894 |
| AlphaFold3 | 5.96 | 2.01 | 0.734 | 0.886 |

Table A17: **Resolution at Fourier shell correlation (FSC)** $= 0.143$ **(↓) on the heterogeneous experimental dataset EMPIAR-13396, with poses inferred *ab initio*.** The two conformational states 9GXQ and 9GXR are resolved from the same mixed particle stack, and the sweep is over the number of particles per state. For Fold'EM *ab initio*, AlphaFold3 and PDB structures, FSC is computed using `phenix.model_map`. Entries highlighted in **bold** indicate the best result (excluding PDB).

| # particles per conf. | Method | 9GXQ | 9GXR |
|---|---|---|---|
| 500 | Fold'EM | **8.67** | 3.58 |
| | RELION (ab initio) | 22.40 | 12.80 |
| | cryoDRGN-AI | 38.40 | 26.90 |
| 1000 | Fold'EM | **7.47** | 4.20 |
| | RELION (ab initio) | 38.40 | 10.75 |
| | cryoDRGN-AI | 26.90 | 22.40 |
| 2000 | Fold'EM | **11.69** | **3.24** |
| | RELION (ab initio) | 19.20 | 10.34 |
| | cryoDRGN-AI | 24.40 | 26.90 |
| 5000 | Fold'EM | **8.67** | **3.02** |
| | RELION (ab initio) | 13.44 | 11.20 |
| | cryoDRGN-AI | 20.68 | 22.40 |
| 10000 | Fold'EM | **4.63** | **3.09** |
| | RELION (ab initio) | 26.88 | 8.40 |
| | cryoDRGN-AI | 17.92 | 16.80 |
| AlphaFold3 | | 11.69 | **3.54** |
| PDB | | 2.64 | 2.89 |

Table A18: **Quantitative evaluation of C$\alpha$ RMSD (↓, Å) and TM-score (↑) on the heterogeneous experimental dataset EMPIAR-13396, with poses inferred *ab initio*.** Fold'EM *ab initio* resolves the two conformational states 9GXQ and 9GXR from the same mixed particle stack, and the sweep is over the number of particles per state. Entries highlighted in **bold** indicate the best result.

| # particles per conf. | C$\alpha$ RMSD (↓, Å) | | TM-score (↑) | |
|---|---|---|---|---|
| | 9GXQ | 9GXR | 9GXQ | 9GXR |
| 500 | 4.54 | 2.61 | 0.799 | 0.874 |
| 1000 | 5.48 | 2.80 | 0.781 | 0.868 |
| 2000 | 5.84 | 2.39 | 0.737 | 0.880 |
| 5000 | 5.06 | **1.63** | 0.776 | 0.895 |
| 10000 | **3.72** | 1.66 | **0.836** | **0.896** |
| AlphaFold3 | 5.96 | 2.01 | 0.734 | 0.886 |

Table A19: **Resolution at Fourier shell correlation (FSC)** $= 0.143$ **(↓) on the heterogeneous synthetic 9UGB/9UGC mixture, with known poses.** The two conformational states are resolved from the same mixed particle stack at SNR $0.01$ and $0.005$, and the sweep is over the number of particles per state. Entries highlighted in **bold** indicate the best result.

| # particles per conf. | Method | 9UGB SNR 0.01 | 9UGB SNR 0.005 | 9UGC SNR 0.01 | 9UGC SNR 0.005 |
|---|---|---|---|---|---|
| 250 | Fold'EM | **3.25** | **3.28** | **3.43** | **3.47** |
| | RELION | 9.24 | 10.40 | 9.24 | 11.09 |
| | cryoDRGN | 4.38 | 6.16 | 6.40 | 8.76 |
| 500 | Fold'EM | **3.24** | **3.25** | **3.39** | **3.43** |
| | RELION | 7.92 | 8.76 | 8.32 | 9.24 |
| | cryoDRGN | 3.62 | 4.50 | 4.75 | 6.16 |
| 1000 | Fold'EM | **3.22** | **3.25** | **3.43** | **3.39** |
| | RELION | 6.93 | 7.93 | 7.23 | 7.92 |
| | cryoDRGN | 3.70 | 3.62 | 4.38 | 4.75 |
| AlphaFold3 | | 8.21 | | 19.29 | |
| PDB | | 3.39 | | 3.47 | |

Table A20: **Quantitative evaluation of C$\alpha$ RMSD (↓, Å) and TM-score (↑) on the heterogeneous synthetic 9UGB/9UGC mixture, with known poses.** Fold'EM resolves the two conformational states from the same mixed particle stack at SNR $0.01$ and $0.005$, and the sweep is over the number of particles per state. Entries highlighted in **bold** indicate the best result.

| | C$\alpha$ RMSD (↓, Å) | | | | TM-score (↑) | | | |
|---|---|---|---|---|---|---|---|---|
| | 9UGB | | 9UGC | | 9UGB | | 9UGC | |
| # particles per conf. | SNR 0.01 | SNR 0.005 | SNR 0.01 | SNR 0.005 | SNR 0.01 | SNR 0.005 | SNR 0.01 | SNR 0.005 |
| 250 | **0.76** | 0.99 | 1.55 | 1.38 | **0.964** | 0.960 | **0.987** | 0.982 |
| 500 | 0.99 | 0.89 | 1.36 | **1.28** | 0.960 | 0.962 | 0.982 | 0.983 |
| 1000 | 0.93 | **0.81** | **1.18** | 1.59 | 0.961 | **0.963** | 0.986 | **0.985** |
| AlphaFold3 | 8.96 | | 11.22 | | 0.699 | | 0.661 | |

Table A21: **Resolution at Fourier shell correlation (FSC)** $= 0.143$ **(↓) on the heterogeneous synthetic 9UGB/9UGC mixture, with poses inferred *ab initio*.** The two conformational states are resolved from the same mixed particle stack at SNR $0.01$ and $0.005$, and the sweep is over the number of particles per state. Entries highlighted in **bold** indicate the best result.

| # particles per conf. | Method | 9UGB SNR 0.01 | 9UGB SNR 0.005 | 9UGC SNR 0.01 | 9UGC SNR 0.005 |
|---|---|---|---|---|---|
| 250 | Fold'EM | **3.22** | **7.61** | **3.63** | **4.81** |
| | RELION (ab initio) | 8.32 | 8.32 | 23.77 | 8.76 |
| | cryoDRGN-AI | 33.28 | 33.28 | 41.60 | 18.49 |
| 500 | Fold'EM | **3.25** | **3.35** | **3.51** | **3.55** |
| | RELION (ab initio) | 6.93 | 7.92 | 7.23 | 8.32 |
| | cryoDRGN-AI | 27.73 | 20.80 | 33.28 | 27.73 |
| 1000 | Fold'EM | **3.22** | **3.25** | **3.51** | **3.51** |
| | RELION (ab initio) | 6.93 | 7.92 | 7.23 | 7.92 |
| | cryoDRGN-AI | 28.81 | 16.64 | 27.73 | 27.73 |
| | AlphaFold3 | 8.21 | | 19.29 | |
| | PDB | 3.39 | | 3.47 | |

Table A22: **Quantitative evaluation of C$\alpha$ RMSD (↓, Å) and TM-score (↑) on the heterogeneous synthetic 9UGB/9UGC mixture, with poses inferred *ab initio*.** Fold'EM resolves the two conformational states from the same mixed particle stack at SNR $0.01$ and $0.005$, and the sweep is over the number of particles per state. Entries highlighted in **bold** indicate the best result.

| | C$\alpha$ RMSD (↓, Å) | | | | TM-score (↑) | | | |
|---|---|---|---|---|---|---|---|---|
| | 9UGB | | 9UGC | | 9UGB | | 9UGC | |
| # particles per conf. | SNR 0.01 | SNR 0.005 | SNR 0.01 | SNR 0.005 | SNR 0.01 | SNR 0.005 | SNR 0.01 | SNR 0.005 |
| 250 | 1.52 | 4.13 | 2.84 | 5.09 | 0.967 | 0.862 | 0.929 | 0.876 |
| 500 | **1.50** | 1.95 | 2.71 | 2.61 | **0.968** | **0.953** | 0.936 | 0.939 |
| 1000 | 1.97 | **1.71** | **2.42** | **2.17** | 0.952 | 0.951 | **0.948** | **0.957** |
| AlphaFold3 | 8.96 | | 11.22 | | 0.699 | | 0.661 | |

### C.5 Continuous heterogeneity on IgG-1D

Table A23: **MolProbity evaluation of Fold'EM structures on IgG-1D**. For each arm of the experiment, we report MolProbity score (↓), clash score (↓), and Ramachandran outlier percentage (↓) of the relaxed structures, computed using `phenix.molprobity` (Williams et al., 2018), averaged over the eight anchors and over the 100 interpolated states.

| MolProbity statistic | cryoDRGN Anchors ($n = 8$) | cryoDRGN Interpolated ($n = 100$) | RECOVAR Anchors ($n = 8$) | RECOVAR Interpolated ($n = 100$) | B-RECOVAR Anchors ($n = 8$) | B-RECOVAR Interpolated ($n = 100$) |
|---|---|---|---|---|---|---|
| MolProbity score | 0.72 | 0.71 | 0.70 | 0.72 | 0.74 | 0.72 |
| Clashscore | 0.59 | 0.59 | 0.59 | 0.66 | 0.66 | 0.54 |
| Ramachandran outliers | 0.15 | 0.15 | 0.15 | 0.15 | 0.15 | 0.15 |

### C.6 CONFIDENCE-BASED PARTICLE SELECTION

Table A24: **Confidence-based particle selection described in Section E.4 prevents mode collapse in the ensemble.** On the heterogeneous experimental dataset EMPIAR-13396, embeddings optimized using high-confidence particles typically yield better-separated conformations, accompanied by higher TM-scores to the corresponding conformations. The "separation" column reports the difference between the RMSD of the true conformations `9GXQ` and `9GXR` and that of the predicted conformations; smaller values indicate closer agreement with the true conformational separation. **Bold** entries indicate the best result.

| # particles per conf. | Confidence method | TM-score (↑) `9GXQ` | TM-score (↑) `9GXR` | Separation (↓) |
|---|---|---|---|---|
| 7.5k | Cut by confidence | **0.835** | **0.894** | **0.58** |
| | No confidence | 0.822 | 0.888 | 1.35 |
| 10k | Cut by confidence | **0.832** | **0.894** | **0.47** |
| | No confidence | 0.823 | 0.888 | 1.48 |

## D ALGORITHMS

Algorithm 1 gives the full Fold'EM procedure described in Section 3: the coordinate ascent that alternates structure inference by guided reverse diffusion with the refinement of the nuisance variables. Algorithm 2 gives a single replicate of the residual-subspace estimator of B-RECOVAR (Appendix E.4), which provides the initial class assignment in the *ab initio* heterogeneous setting.

**Algorithm 1** Fold'EM

**Input:** Initial trunk embeddings $\boldsymbol{\mathcal{H}}$; particle images $\boldsymbol{\mathcal{Y}}$; initial nuisance variables $\boldsymbol{\eta}^{(0)}$; noise schedule $[\sigma_T, \ldots, \sigma_0]$; learning rate $\gamma_h$; ensemble size $n$; coordinate iterations $R$; diffusion runs $Q$; prior weight $\lambda_p$
**Output:** Structural ensemble $\boldsymbol{\mathcal{X}} = \{\mathbf{X}^1, \ldots, \mathbf{X}^n\}$ decoded from the optimized Pairformer states $g(\boldsymbol{\mathcal{H}})$
**for** $r = 0$ to $R - 1$ **do** ▷ Coordinate ascent
  **for** $q = 1$ to $Q$ **do** ▷ Step 1: structure inference, Eq. 3
    $\boldsymbol{\mathcal{X}} \sim \sigma_T \cdot [\mathbf{N}^1, \ldots, \mathbf{N}^n]^T$ $\mathbf{N}^i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
    **for** $t \in [T-1, \ldots, 0]$ **do**
      $\boldsymbol{\mathcal{H}} \leftarrow \boldsymbol{\mathcal{H}} + \gamma_h \nabla_{\boldsymbol{\mathcal{H}}} \Big( \log p\big(\boldsymbol{\mathcal{Y}} \mid D_\theta(\boldsymbol{\mathcal{X}}; g(\boldsymbol{\mathcal{H}}), t), \boldsymbol{\eta}^{(r)}\big) + \lambda_{\mathrm{p}} \log p(\boldsymbol{\mathcal{H}}|\mathbf{a}) \Big)$ ▷ Eq. 5
      $\boldsymbol{\mathcal{X}} \leftarrow \mathrm{ReverseSDE}(\boldsymbol{\mathcal{X}}; g(\boldsymbol{\mathcal{H}}), t, \sigma_t)$ ▷ Eq. 4
    **end for**
  **end for**
  $\boldsymbol{\eta}^{(r+1)} \leftarrow \arg\max_{\boldsymbol{\eta}} \Big( \log p(\boldsymbol{\mathcal{Y}} \mid \boldsymbol{\mathcal{X}}, \boldsymbol{\eta}) + \log p(\boldsymbol{\eta}) \Big)$ ▷ Step 2: pose and class refinement, Eq. 6
**end for**
$\boldsymbol{\mathcal{Z}} \leftarrow g(\boldsymbol{\mathcal{H}})$ ▷ Optimized conditioning variables
$\boldsymbol{\mathcal{X}} \sim p(\boldsymbol{\mathcal{X}} \mid \boldsymbol{\mathcal{Z}})$ ▷ Decode ensemble: unguided reverse diffusion, Eq. 4
**return** $\boldsymbol{\mathcal{X}}$ ▷ Ensemble from the final reverse-diffusion run, conditioned on $g(\boldsymbol{\mathcal{H}})$

**Algorithm 2** B-RECOVAR residual-subspace estimator for one replicate

**Input:** particles with poses and CTFs; band $\Omega_{\mathrm{a}}$; number of classes $n$ and of modes $q$; fitting subset $\mathcal{I}$; mask fraction $f$; $T$ alternating updates; $L$ refinement iterations
Reconstruct the consensus $\hat{\rho}$ from all particles and derive the solvent mask $\mathcal{M}$ from it
Scale every particle to the consensus and subtract its projection ▷ residuals $\mathbf{r}_k$
**for** $m = 1, \ldots, q$ **do**
    From a random start, alternate $T$ times between back-projecting the residuals of $\mathcal{I}$ into a masked mode $W_m$ and projecting them onto it ▷ alternating least squares
    Remove the fitted mode from the residuals ▷ deflation
**end for**
Fit the coefficients $\hat{\mathbf{b}}_k$ of every particle on all modes jointly ▷ joint coefficients
Cluster $\{\hat{\mathbf{b}}_k\}$ into $n$ classes by $k$-means ▷ initial labels
Refine the labels by $L$ cross-validated updates equation 31 ▷ two folds, centered rule
**return** labels $\{\hat{c}_k\}_{k=1}^{K}$

# E ADDITIONAL DETAILS

## E.1 PRE-PROCESSING

**Synthetic Datasets.** All synthetic datasets were prepared using the PDB deposited structure (Burley et al., 2017). First, noiseless particles $\mathbf{y}$ were rendered by projecting the experimental PDB structure $\mathbf{X}^{\text{expt}}$ using the image formation model in Section E.3.3. White Gaussian noise was added to the noiseless particles in the Fourier domain to match a specific signal to noise ratio (SNR).

$$\hat{\mathbf{y}}^{\text{noisy}} = \hat{\mathbf{y}} + \sigma_{\text{noise}}\hat{\mathbf{N}}, \qquad \mathbf{N} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

The noise scale $\sigma_{\text{noise}}$ was set using,

$$\sigma_{\text{noise}} = \frac{\text{std}(\hat{\mathbf{y}})}{\sqrt{\text{SNR}}}.$$

Where $\text{SNR}$ is the desired SNR for the dataset. Thus, the resulting particle satisfies

$$\text{SNR} = \frac{\text{Var}[\hat{\mathbf{y}}]}{\text{Var}[\sigma_{\text{noise}}\hat{\mathbf{N}}]}.$$

In all our synthetic experiments, we set $\text{SNR} = 0.01$ or $0.005$.

**Real datasets** All experimental datasets were downloaded from EMPIAR (Iudin et al., 2016). We preprocess the deposited particle ensemble generation. We only retain the particles listed in the depositors' final STAR files, and exclude particles identified as junk. The particles are then centered and Fourier-cropped to $128 \times 128$ pixels, followed by phase flipping. We then apply a soft circular mask with a radius equal to $0.4$ times the box size.

## E.2 BASELINES

**Known poses.** We primarily compare against RELION (Zivanov et al., 2018) and CryoDRGN (Zhong et al., 2021). Both methods reconstruct 3D density from particle images.

- **CryoDRGN.** With known poses, CryoDRGN (Zhong et al., 2021) volumes were trained using the following command:

```
cryodrgn train_vae <particles.mrcs> \
    --poses <poses.pkl> --ctf <ctf.pkl> [--ind <ind.pkl>] \
    --zdim 8 -n 50 --no-amp
```

- **RELION.** With known poses, RELION (Zivanov et al., 2018) volumes were generated using the following command:

```
relion_refine --ctf --skip_align --iter 5 \
    --particle_diameter d --tau2_fudge 1 \
    --split_random_halves --flatten_solvent \
    --solvent_mask mask.mrc
```

  Here, `mask.mrc` is a low-pass reference generated from the EMD map with a $60$ Å blob mask. The particle diameter `d` depends on the protein.

**ab initio.** We primarily compare against RELION (ab initio) and CryoDRGN-AI (Levy et al., 2025). These methods jointly estimate particle poses and 3D structure from the particle images without externally supplied orientations.

- **CryoDRGN-AI.** CryoDRGN-AI volumes were trained using the following command:

```
cryodrgn abinit <particles.mrcs> --ctf <ctf.pkl> \
    --zdim 8 --num-epochs 50 --no-amp --seed 0 --num-workers 0 \
    --lazy --checkpoint 50 --o <outdir>
```

- **RELION (ab initio).** Without poses, RELION volumes were generated in two stages. First, an initial model was constructed using VDAM without externally provided poses:

```
relion_refine --grad --denovo_3dref --i particles.star \
    --o init<SUF>/run --iter 200 --K 1 \
    --sym $SYM --random_seed $SEED --particle_diameter $DIAM \
    --flatten_solvent --zero_mask --ctf \
    --ini_high 40 --oversampling 1 --healpix_order 1 \
    --offset_range 6 --offset_step 2 --pool 30 --pad 1 \
    --skip_gridding --dont_combine_weights_via_disc --gpu --j $J
```

  We then performed 3D auto-refinement from the resulting initial model:

```
mpirun --oversubscribe --bind-to none -n 3 relion_refine_mpi \
    --auto_refine --split_random_halves --i particles.star \
    --ref <init map> --ini_high 30 --o refine<SUF>/run \
    --random_seed $SEED --ctf --particle_diameter $DIAM \
    --flatten_solvent --zero_mask --sym $SYM \
    --oversampling 1 --healpix_order 2 \
    --auto_local_healpix_order 4 --offset_range 5 \
    --offset_step 2 --low_resol_join_halves 40 \
    --norm --scale --pool 30 --pad 2 \
    --dont_combine_weights_via_disc --gpu --j $JT
```

  In all Fold'EM *ab initio* experiments, we use the poses and initial map generated in the first stage for the subsequent refinement.

**Heterogeneous reconstruction.** For heterogeneous datasets, we first clustered the particles into structural groups and then reconstructed each density separately using the corresponding particle subset. This provides a density estimate for each discrete structural class without introducing additional heterogeneity modelling assumptions.

### E.2.1 AlphaFold3 and hardware resources

For all our experiments, we used Protenix v1.0.4 (Zhang et al., 2026) – a PyTorch reimplementation of AlphaFold3 (Abramson et al., 2024). For map and pose estimation, we used RELION 5.0.0 (Zivanov et al., 2018). For structural validation, we used Phenix (Liebschner et al., 2019) v1.21.2 − 5419. All experiments were carried out on NVIDIA H100 and L40S GPUs running Debian GNU/Linux 12.

### E.2.2 Collision and bond length loss

To prevent the embedding optimization from generating physically invalid protein structures, we incorporate numerous stereochemical regularizers. Given an ensemble of conformations $\boldsymbol{\mathcal{X}} = \{\mathbf{X}^1, \ldots, \mathbf{X}^n\}$ corresponding to amino acid sequence $\mathbf{a}$, we minimize three loss functions that avoid elongated bonds and steric clashes.

**Bond length loss.** The bond-length loss for a structure with $m$ atoms is given by,

$$\mathcal{L}_{\text{bond}}(\mathbf{X}; \mathbf{a}) = \sum_{i=1}^{m} \sum_{j=i+1}^{m} B_{ij} \cdot (\max(0, |d_{ij} - d_{ij}^{\text{ideal}}| - \delta_{\text{bond}}))^2$$

Here, $\mathbf{B} \in \mathbb{Z}^{m \times m}$ is a mask with every element $b_{ij} \in \{0, 1\}$. If atoms $i$ and $j$ are covalently bonded, $b_{ij} = 1$, otherwise $b_{ij} = 0$. $d_{ij}^{\text{ideal}}$ is the sum of covalent radii of atoms $i$, $j$, and $d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|_2$. $\delta_{\text{bond}} = 0.2$ Å is a tolerance margin.

**Collision loss.** Collisions are caused when any two non-bonded atoms are close in space. To prevent this, we minimize:

$$\mathcal{L}_{\text{collision}}(\mathbf{X}; \mathbf{a}) = \sum_{i=1}^{m} \max_{j \neq i, B_{ij}=0} (\max(0, (d_{ij}^{\text{ideal}} + p^{\text{collision}}) - d_{ij})) \tag{7}$$

Here, $p^{\text{collision}} = 0.4\,\text{Å}$ is a padding distance to prevent over-penalization of near-contact atoms. Together, the resulting validity log-likelihood of an ensemble of conformations is given by

$$\log p(\mathbf{B} \mid \boldsymbol{\mathcal{X}}, \mathbf{a}) = -\sum_{k=1}^{n} (\lambda_{\text{bond}} \mathcal{L}_{\text{bond}}(\mathbf{X}^k; \mathbf{a}) + \lambda_{\text{collision}} \mathcal{L}_{\text{collision}}(\mathbf{X}^k; \mathbf{a})) \tag{8}$$

### E.2.3 Relaxation

Despite being regularized with loss functions that prevent collision and bond-violation, some minor clashes can still persist. Hence, all outputs of Fold'EM are relaxed using AMBER14 force field (Hornak et al., 2006) in vacuum, in order to relax nucleic acids too. The energy of the structures is minimized for a maximum of 2000 iterations with an energy tolerance threshold of $2.39\,\text{kcal/mol}$ and stiffness of $10.0\,\text{kcal/mol/Å}^2$

### E.2.4 Metrics

**$C\alpha$ Root mean square deviation ($C\alpha$ RMSD).** Given a structure $\mathbf{X}$ generated by Fold'EM, we measure its structural accuracy relative to the experimentally resolved structure $\mathbf{X}^{\text{expt}}$ using the RMSD over matched $C\alpha$ atoms. We first align $\mathbf{X}$ to $\mathbf{X}^{\text{expt}}$ using the optimal rotation $\mathbf{R}^{\text{kabsch}}$ and translation $\mathbf{t}^{\text{kabsch}}$ obtained by Kabsch rigid-body alignment (Kabsch, 1976). We then compute,

$$\text{RMSD}_{C\alpha} = \sqrt{\frac{1}{N_{C\alpha}} \sum_{i \in \mathcal{I}_{C\alpha}}^{m} \cdot |\mathbf{R}^{\text{kabsch}} \mathbf{x}_i + \mathbf{t}^{\text{kabsch}} - \mathbf{x}_i^{\text{expt}}|_2^2}$$

where $\mathcal{I}_{C\alpha}$ denotes the set of matched $C\alpha$ atoms and $N_{C\alpha} = |\mathcal{I}_{C\alpha}|$.

**Template Modeling score (TM-score).** For a Fold'EM generated structure $\mathbf{X}$ and experimental structure $\mathbf{X}^{\text{expt}}$, the Template Modeling score (TM-score) measures the global structural similarity (Zhang & Skolnick, 2004). In contrast to RMSD, the TM-score is less sensitive to a small number of large local deviations. It is defined as

$$\text{TMScore} = \max_{\mathbf{R},\mathbf{t}} \frac{1}{L_{\text{ref}}} \sum_{i=1}^{L_{\text{ali}}} \frac{1}{1 + \left(\frac{d_i(\mathbf{R},\mathbf{t})}{d_0(L_{\text{ref}})}\right)^2}, \tag{9}$$

where $L_{\text{ref}}$ is the number of residues in the reference structure, $L_{\text{ali}}$ is the number of aligned residues and $d_i(\mathbf{R}, \mathbf{t})$ is the distance between the $i$th pair of aligned $C\alpha$ atoms after rigid-body superposition. The length-dependent normalization is given by

$$d_0(L_{\text{ref}}) = 1.24(L_{\text{ref}} - 15)^{1/3} - 1.8\ \text{Å}. \tag{10}$$

TM-scores lie in $(0, 1]$. Larger values indicate greater structural similarity.

**MolProbity scores.** We assess stereochemical quality using `phenix.molprobity` (Williams et al., 2018). We report the MolProbity score, clashscore and percentage of Ramachandran outliers. The clashscore is the number of serious all-atom steric overlaps per 1000 atoms. The Ramachandran outlier percentage is the fraction of residues whose backbone $\phi$ and $\psi$ angles fall outside the allowed regions of the reference Ramachandran distributions. The MolProbity score is a composite measure combining all-atom clashes, side-chain rotamer outliers and Ramachandran statistics. Lower scores indicate better stereochemical quality.

**Fourier shell correlation$= 0.143$.** For atomic models generated by Fold'EM, Fold'EM ab initio, AlphaFold3, and obtained from the PDB, we report the resolution (Å) at which the Fourier shell correlation (FSC) curve wrt the experimental map $\rho$ from EMD crosses the $0.143$ threshold, following the commonly used FSC criterion for cryo-EM resolution assessment (Rosenthal & Henderson, 2003). We use `phenix.map_box` to compute the FSC curve and determine the corresponding resolution at the $0.143$ threshold.

### E.3 CRYO-EM IMAGE FORMATION MODEL AND LIKELIHOOD FUNCTION

Given a particle image $\mathbf{y}$, its log-likelihood given a conformation $\mathbf{X}$ and pose $\boldsymbol{\phi}$ can be written as $\log p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\phi}) = \log p(\mathbf{y} \mid \mathbf{y}_{\mathrm{c}}(\mathbf{X}, \boldsymbol{\phi}))$, where $\mathbf{y}_{\mathrm{c}}$ is the simulated noiseless particle image and $\log p(\mathbf{y} \mid \mathbf{y}_{\mathrm{c}})$ is the noise log-likelihood.

#### E.3.1 IMAGE FORMATION MODEL

Given an atomic structure $\mathbf{X} = \{(\mathbf{x}_i, e_i)\}_{i=1}^m$, where $\mathbf{x}_i \in \mathbb{R}^3$ is the position and $e_i$ the chemical element of atom $i$, we approximate its electrostatic potential as a superposition of atomic potentials, $\rho_{\mathbf{X}}(\mathbf{r}) = \sum_{i=1}^m \rho_{e_i}(\mathbf{r} - \mathbf{x}_i)$. Each atomic potential is specified by its electron scattering factor $f_e$, which we identify, up to the conventional proportionality constant, with its Fourier transform, $\widehat{\rho}_e(\boldsymbol{\xi}) = f_e(\|\boldsymbol{\xi}\|)$. The Fourier representation of the molecular potential is therefore

$$\widehat{\rho}_{\mathbf{X}}(\boldsymbol{\xi}) = \sum_{i=1}^{m} f_{e_i}(\|\boldsymbol{\xi}\|) \exp\bigl(-2\pi i\, \boldsymbol{\xi}^\top \mathbf{x}_i\bigr)\,, \tag{11}$$

with $\boldsymbol{\xi}$ denoting the 3D spatial frequency. For a pose $\boldsymbol{\phi} = (\mathbf{R}, \mathbf{t})$, with $\mathbf{R} \in \mathrm{SO}(3)$ and in-plane translation $\mathbf{t} \in \mathbb{R}^2$, we define the projection operator $\mathcal{P}_{\boldsymbol{\phi}}$ projecting the density onto a plane oriented at $\boldsymbol{\phi}$. We express the projection in the Fourier domain through the slice theorem: projection of a rotated 3D potential corresponds to sampling a central plane of its 3D Fourier transform. Let $\boldsymbol{\omega} = (\omega_1, \omega_2)^\top$ denote a 2D spatial frequency. We define the pose-dependent slice operator

$$(\mathcal{S}_{\boldsymbol{\phi}}\widehat{\rho})(\boldsymbol{\omega}) = e^{-2\pi i \boldsymbol{\omega}^\top \mathbf{t}} \widehat{\rho}\left(\mathbf{R}^\top \begin{bmatrix} \boldsymbol{\omega} \\ 0 \end{bmatrix}\right), \tag{12}$$

where the rotation selects the corresponding central Fourier plane and the translation appears as a phase ramp. The noiseless particle image is formed by further applying the known contrast transfer function (CTF), $G(\boldsymbol{\omega})$, in the Fourier domain $\widehat{\mathbf{y}}_{\mathrm{c}}(\mathbf{X}, \boldsymbol{\phi})(\boldsymbol{\omega}) = G(\boldsymbol{\omega})\,(\mathcal{S}_{\boldsymbol{\phi}}\widehat{\rho}_{\mathbf{X}})\,(\boldsymbol{\omega})$, or in the real space $\mathbf{y}_{\mathrm{c}}(\mathbf{X}, \boldsymbol{\phi}) = \mathcal{F}^{-1}\left[G\,\mathcal{S}_{\boldsymbol{\phi}}\widehat{\rho}_{\mathbf{X}}\right]$. Known particle-specific imaging parameters, including the CTF, are suppressed from $\boldsymbol{\phi}$ throughout the paper; they can equivalently be included among the nuisance variables introduced above.

#### E.3.2 NOISE MODEL AND LIKELIHOOD FUNCTION

The fundamental electron-detection noise is shot noise and is therefore Poisson at the level of individual electron counts. The particle images used for reconstruction, however, are processed rather than raw counting measurements: detector response, frame integration, motion correction, dose weighting, gain correction, background subtraction, normalization, and other sources of experimental uncertainty make a simple pixelwise Poisson model inappropriate for the resulting images. As is standard in likelihood-based single-particle cryo-EM reconstruction (Scheres, 2012), we therefore model the residual image noise as additive Gaussian noise, allowing its variance to depend on spatial frequency.

Specifically, let $\mathbf{y}_c$ denote a candidate noiseless particle image and assume additive Gaussian noise in the Fourier domain, $\widehat{\mathbf{y}} = \widehat{\mathbf{y}}_c + \widehat{\mathbf{N}}$. We follow the community practice and model each Fourier coefficient of $\widehat{\mathbf{N}}$ as an independent zero-mean complex Gaussian with frequency-dependent variance $\sigma^2(\boldsymbol{\omega})$. The corresponding noise log-likelihood is, up to an additive constant,

$$\log p(\mathbf{y} \mid \mathbf{y}_c) = -\frac{1}{2} \sum_{\boldsymbol{\omega} \in \Omega} \frac{|\widehat{\mathbf{y}}(\boldsymbol{\omega}) - \widehat{\mathbf{y}}_c(\boldsymbol{\omega})|^2}{\sigma^2(\boldsymbol{\omega})}. \tag{13}$$

The Gaussian model should be understood as an effective likelihood for processed cryo-EM particle images rather than as a model of the underlying electron-arrival process. In practice, we do not optimize this Gaussian log-likelihood directly. Instead, we use a band-limited normalized correlation between the observed and simulated particle images in Fourier space,

$$\mathcal{L}_{\mathrm{corr}}(\mathbf{y}, \mathbf{y}_c) = \frac{\mathrm{Re}\bigl[\sum_{\boldsymbol{\omega} \in \Omega} \widehat{\mathbf{y}}(\boldsymbol{\omega})\, \widehat{\mathbf{y}}_c(\boldsymbol{\omega})^*\bigr]}{\sqrt{\sum_{\boldsymbol{\omega} \in \Omega} |\widehat{\mathbf{y}}(\boldsymbol{\omega})|^2} \sqrt{\sum_{\boldsymbol{\omega} \in \Omega} |\widehat{\mathbf{y}}_c(\boldsymbol{\omega})|^2 + \epsilon}},$$

where $\Omega = \{\boldsymbol{\omega} : k_{\min} \le \|\boldsymbol{\omega}\|_2 < k_{\max} + 1\}$ is the frequency band used for guidance and $\epsilon$ is a small numerical stabilizer. Maximizing this objective is equivalent to minimizing squared error between

the observed and simulated Fourier images after normalizing both to unit norm. It can therefore be viewed as a scale-invariant surrogate for the Gaussian image log-likelihood.

We use this surrogate because the overall amplitude and contrast of experimental cryo-EM particles can vary for reasons that are not informative about molecular structure. The normalization removes sensitivity to such global scale differences and makes the guidance depend primarily on agreement in the spatial-frequency content and phase structure of the particle. Restricting the comparison to $\Omega$ further excludes very low frequencies dominated by background and normalization effects, as well as high frequencies with little reliable signal. Compared with the full Gaussian likelihood, this objective discards explicit frequency-dependent noise weighting in favor of a simpler and empirically more robust similarity measure for inference-time optimization.

#### E.3.3 Particle image rendering from atomic structure

We implement the image-formation model of Section E.3.1 directly from the atomic structure $\mathbf{X} = \{(\mathbf{x}_i, e_i)\}_{i=1}^m$. In the implementation, we approximate the atomic potential $\rho_{e_i}$ of atom $i$ by an isotropic Gaussian with amplitude $Z_i$, the atomic number associated with element $e_i$.

Let $a$ denote the image pixel size in Å/pixel. We set the Gaussian width according to the Nyquist resolution $2a$,

$$\sigma = \frac{2a}{\sqrt{2}\pi}, \tag{14}$$

which corresponds to

$$\sigma_{\mathrm{px}} = \frac{\sigma}{a} = \frac{\sqrt{2}}{\pi} \approx 0.45 \tag{15}$$

pixels.

Because integrating an isotropic 3D Gaussian along the viewing axis yields a 2D Gaussian with the same width, we evaluate the projection analytically without constructing an intermediate 3D voxel grid. For a pose $\phi = (\mathbf{R}, \mathbf{t})$, with $\mathbf{R} \in SO(3)$ and in-plane translation $\mathbf{t} \in \mathbb{R}^2$, we first rotate atom $i$ into the camera frame,

$$\tilde{\mathbf{x}}_i = \mathbf{R}^\top \mathbf{x}_i. \tag{16}$$

Its projected center, expressed in pixel coordinates relative to the image center, is

$$\mathbf{c}_i = \frac{1}{a} \begin{bmatrix} \tilde{x}_{i,1} \\ \tilde{x}_{i,2} \end{bmatrix} + \mathbf{t}, \tag{17}$$

where $\mathbf{t}$ is measured in pixels.

We use the projected atomic centers to define the continuous 2D projection

$$P_\phi(\mathbf{X})(\mathbf{r}) = \sum_{i=1}^m Z_i \exp\left(-\frac{\|\mathbf{r} - \mathbf{c}_i\|_2^2}{2\sigma_{\mathrm{px}}^2}\right), \tag{18}$$

where $\mathbf{r} = (u, v)^\top$ denotes image-plane coordinates in pixels.

We integrate each Gaussian over the finite pixel area. Since the isotropic 2D Gaussian is separable, this integral factorizes into two one-dimensional Gaussian integrals. For a pixel centered at coordinate $p$ and a Gaussian centered at $c$, we define

$$g(p; c) = \int_{p-\frac{1}{2}}^{p+\frac{1}{2}} \exp\left(-\frac{(u-c)^2}{2\sigma_{\mathrm{px}}^2}\right) \mathrm{d}u. \tag{19}$$

This integral has a closed-form expression in terms of the error function,

$$g(p; c) = \sigma_{\mathrm{px}} \sqrt{\frac{\pi}{2}} \left[\mathrm{erf}\left(\frac{p + \frac{1}{2} - c}{\sqrt{2}\sigma_{\mathrm{px}}}\right) - \mathrm{erf}\left(\frac{p - \frac{1}{2} - c}{\sqrt{2}\sigma_{\mathrm{px}}}\right)\right], \tag{20}$$

where $\mathrm{erf}$ denotes the error function

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} \mathrm{d}t. \tag{21}$$

The rendered value at pixel $(p, q)$ is therefore

$$I_\phi[p, q] = \sum_{i=1}^{m} Z_i \, g(p; c_{i,1}) \, g(q; c_{i,2}). \tag{22}$$

#### E.3.4 Contrast Transfer Function

We assume the CTF parameters of each particle are known. We model contrast transfer under the weak-phase approximation (Wade, 1992) and use the CTF parameterization of Rohou & Grigorieff (2015). Following the image-formation model of Section E.3, the rendered projection is multiplied by the corresponding particle-specific CTF in the Fourier domain.

For a two-dimensional spatial frequency $\boldsymbol{\omega} = (\omega_x, \omega_y)^\top$, let

$$s = \|\boldsymbol{\omega}\|_2, \qquad \theta = \operatorname{atan2}(\omega_y, \omega_x), \tag{23}$$

where $s$ is measured in $\text{Å}^{-1}$.

To account for astigmatism, the defocus is parameterized by two principal defocus values $d_u$ and $d_v$ and their orientation $\theta_d$. The effective defocus at Fourier-space angle $\theta$ is

$$d_f(\theta) = \frac{1}{2}\left[d_u + d_v + (d_u - d_v)\cos\left(2(\theta - \theta_d)\right)\right]. \tag{24}$$

When $d_u = d_v$, this reduces to a rotationally symmetric CTF with a single defocus value.

The corresponding CTF phase is

$$\gamma(\boldsymbol{\omega}) = 2\pi\left[-\frac{1}{2} d_f(\theta)\lambda s^2 + \frac{1}{4} C_s \lambda^3 s^4\right] - \varphi, \tag{25}$$

where $\lambda$ is the electron wavelength, $C_s$ is the spherical aberration coefficient and $\varphi$ is the phase shift.

The CTF is then

$$G(\boldsymbol{\omega}) = \sqrt{1 - w^2}\, \sin\left(\gamma(\boldsymbol{\omega})\right) - w\, \cos\left(\gamma(\boldsymbol{\omega})\right), \tag{26}$$

where $w$ denotes the amplitude-contrast fraction.

The electron wavelength is determined by the accelerating voltage $V$ through the relativistic de Broglie relation,

$$\lambda(V) = \frac{hc}{\sqrt{eV\,(eV + 2m_e c^2)}}, \tag{27}$$

where $h$ is Planck's constant, $e$ the elementary charge, $m_e$ the electron rest mass and $c$ the speed of light. Expressing $V$ in volts and $\lambda$ in Å, this becomes

$$\lambda(V) = \frac{12.2642598}{\sqrt{V\,(1 + 9.78475598 \times 10^{-7} V)}}. \tag{28}$$

The CTF is applied in the Fourier domain. Let $\widehat{\mathbf{I}}_\phi(\boldsymbol{\omega})$ denote the Fourier transform of the rendered projection from the previous section. The simulated noiseless particle is given by

$$\widehat{\mathbf{y}}_c(\mathbf{X}, \phi)(\boldsymbol{\omega}) = G(\boldsymbol{\omega})\, \widehat{\mathbf{I}}_\phi(\boldsymbol{\omega}), \tag{29}$$

with the real-space image obtained by inverse Fourier transformation.

For phase-flipped experimental particle stacks, we apply the magnitude $|G(\boldsymbol{\omega})|$ rather than the signed CTF $G(\boldsymbol{\omega})$ to match the preprocessing of the observed images.

### E.4 B-RECOVAR: BOOTSTRAP-AGGREGATED RECOVAR CLASS ASSIGNMENT

In the *ab initio* heterogeneous setting, Algorithm 1 requires an initial class assignment $c^{(0)} = (c_k^{(0)})_{k=1}^K$ for $\boldsymbol{\eta}^{(0)}$, inferred directly from the particle images (Section 3.1). Standard low-dimensional embeddings, including cryoDRGN latent coordinates and principal components of the RECOVAR volume covariance, can be unreliable when few particles are available per class. Nuisance variation, such as continuous motion, defocus variation, and specimen-preparation effects, can have variance comparable to the structural signal of interest, and the leading covariance directions may therefore depend strongly on the sampled particles.

We therefore use B-RECOVAR, a bootstrap-style, subsampled ensemble for class initialization. Each replicate estimates a low-dimensional residual subspace by pose-aware regularized back-projection, clusters the resulting particle coefficients, and refines the labels by cross-validated comparison with class reconstructions. Replicate labelings are then aggregated by spectral voting, and the normalized voting margin is reported as a particle-level confidence score.

**Model and notation.** Let $\boldsymbol{\mathcal{Y}} = \{\mathbf{y}_k\}_{k=1}^K$ denote $K$ particle images drawn from a mixture of $n$ discrete conformations with densities $\rho_1, \dots, \rho_n$, mixing proportions $\pi_1, \dots, \pi_n > 0$, $\sum_{j=1}^n \pi_j = 1$, and class labels $c_k \in \{1, \dots, n\}$. Poses $\boldsymbol{\phi}_k$ and CTFs $G_k$ are obtained from a preceding consensus refinement and registered to a common coordinate system. The labels are unknown, and the procedure uses no reference structures or atomic models.

We restrict volumes to the Fourier shells $\Omega_{\mathrm{a}} = \{\boldsymbol{\xi} : k_{\min} \le \|\boldsymbol{\xi}\|_2 < k_{\max}\}$ and images to the corresponding central-plane coefficients: the lowest-frequency shells are dominated by global contrast and normalization, whereas the highest-frequency shells are noise-dominated. On this band, the imaging operator for particle $k$ is $\mathbf{A}_k = G_k \mathcal{S}_{\boldsymbol{\phi}_k}$, where $\mathcal{S}_{\boldsymbol{\phi}}$ is the central-slice operator of equation 12, and we assume $\mathbf{y}_k = \alpha_k^\star \mathbf{A}_k \rho_{c_k} + \boldsymbol{\varepsilon}_k$, with $\alpha_k^\star > 0$ a particle-specific intensity scale and $\boldsymbol{\varepsilon}_k$ zero-mean noise, independent across particles and approximately white on $\Omega_{\mathrm{a}}$.

**Residual subspace.** Define the mixture mean $\bar{\rho} = \sum_{j=1}^n \pi_j \rho_j$ and the class deviations $\delta_j = \rho_j - \bar{\rho}$. Since $\sum_{j=1}^n \pi_j \delta_j = 0$, the class deviations span a space of dimension $q = \dim \operatorname{span}\{\delta_1, \dots, \delta_n\} \le n - 1$. Let $U_1, \dots, U_q$ be a basis of this space and write $\delta_j = \sum_{m=1}^q b_{jm} U_m$. After relative intensity normalization and consensus subtraction, the residual of particle $k$ satisfies

$$\mathbf{r}_k \approx \sum_{m=1}^{q} b_{c_k m} \mathbf{A}_k U_m + \boldsymbol{\varepsilon}'_k, \tag{30}$$

where any common scale is absorbed into the modes $U_m$, and $\boldsymbol{\varepsilon}'_k$ includes measurement noise and scale-estimation error. Thus, particles from class $j$ have a common mean coefficient vector $\mathbf{b}_j \in \mathbb{R}^q$, and if $q = n - 1$, these vectors form the vertices of an $(n-1)$-simplex. For $n = 2$, the subspace is spanned by $\Delta\rho = \rho_1 - \rho_2$, with $\delta_1 = \pi_2 \Delta\rho$ and $\delta_2 = -\pi_1 \Delta\rho$, and the coefficient sign distinguishes the two classes up to a global label exchange. For a balanced two-state mixture, $\|\delta_1\| = \|\delta_2\|$, so the residual energy alone carries no class information.

For a particle subset $\mathcal{K}$, define the regularized back-projection

$$\mathcal{R}\big(\{(\nu_k, \mathbf{d}_k)\}_{k \in \mathcal{K}}\big) = \Big( \sum_{k \in \mathcal{K}} \nu_k \mathbf{A}_k^* \mathbf{A}_k + \lambda_{\mathrm{w}} \mathbf{I} \Big)^{-1} \sum_{k \in \mathcal{K}} \mathbf{A}_k^* \mathbf{d}_k,$$

and let $\mathcal{M}$ denote multiplication in real space by a smooth solvent mask derived from the consensus reconstruction. Using all particles, we form the consensus $\tilde{\rho} = \mathcal{R}(\{(1, \mathbf{y}_k)\}_{k=1}^K)$, the scales $\tilde{\alpha}_k = \mathrm{Re}\langle \mathbf{y}_k, \mathbf{A}_k \tilde{\rho} \rangle / \|\mathbf{A}_k \tilde{\rho}\|^2$, the normalized images $\check{\mathbf{y}}_k = \tilde{\alpha}_k^{-1} \mathbf{y}_k$, and the residuals $\mathbf{r}_k = \check{\mathbf{y}}_k - \mathbf{A}_k \tilde{\rho}$. The scale normalization makes the class coefficients comparable across particles.

We seek masked modes $W_1, \dots, W_q$ and coefficient vectors $\tilde{\mathbf{b}}_k \in \mathbb{R}^q$ through the constrained factorization

$$\underset{\substack{\{W_m\}, \{\tilde{\mathbf{b}}_k\} \\ \|\tilde{\mathbf{b}}_{\cdot m}\|_2 = 1,\ m = 1, \dots, q}}{\text{minimize}} \quad \sum_{k=1}^{K} \Big\| \mathbf{r}_k - \sum_{m=1}^{q} \tilde{b}_{km} \mathbf{A}_k W_m \Big\|^2 + \lambda_{\mathrm{w}} \sum_{m=1}^{q} \|W_m\|^2.$$

The normalization removes the scale ambiguity between a mode and its coefficients. The modes are identifiable only up to a change of basis of the subspace, and the class labels only up to a permutation. For $n = 2$, omitting the mask yields a rank-one, one-dimensional RECOVAR fit without noise correction (Gilles & Singer, 2025). The smooth mask allows for variance reduction, as it suppresses solvent-dominated contributions, provided that it contains the class-specific density.

We estimate the factorization by greedy masked alternating least squares with deflation. Let $\mathcal{I}$ be the particle subset used to fit one replicate; for a single full-data fit, $\mathcal{I} = \{1, \dots, K\}$. Set $\mathbf{r}_k^{(0)} = \mathbf{r}_k$. For $m = 1, \dots, q$, we initialize $\mathbf{a}_{\mathcal{I}m} = (a_{km})_{k\in\mathcal{I}}$ as a random unit vector and iterate

$$W_m \leftarrow \mathcal{M}\,\mathcal{R}\big(\{(a_{km}^2, a_{km}\mathbf{r}_k^{(m-1)})\}_{k\in\mathcal{I}}\big),$$
$$a_{km} \leftarrow \frac{\langle \mathbf{r}_k^{(m-1)}, \mathbf{A}_k W_m\rangle}{\|\mathbf{A}_k W_m\|^2}, \qquad \mathbf{a}_{\mathcal{I}m} \leftarrow \frac{\mathbf{a}_{\mathcal{I}m}}{\|\mathbf{a}_{\mathcal{I}m}\|_2}, \quad k \in \mathcal{I}.$$

After the final update, we recompute the unnormalized coefficients $\check{a}_{km} = \langle \mathbf{r}_k^{(m-1)}, \mathbf{A}_k W_m\rangle / \|\mathbf{A}_k W_m\|^2$ and deflate, $\mathbf{r}_k^{(m)} = \mathbf{r}_k^{(m-1)} - \check{a}_{km}\mathbf{A}_k W_m$, for $k \in \mathcal{I}$. Once all modes are fitted, the coefficients of every particle are estimated jointly from the original residual, $\tilde{\mathbf{b}}_k = \arg\min_{\mathbf{b}\in\mathbb{R}^q} \|\mathbf{r}_k^{(0)} - \sum_{m=1}^q b_m \mathbf{A}_k W_m\|^2$, $k = 1, \dots, K$. Under equation 30, the expected coefficient vector depends on the class but not on the pose, up to the fitted change of basis. We cluster $\{\tilde{\mathbf{b}}_k\}_{k=1}^K$ by $k$-means with $n$ clusters and multiple initializations. For $n = 2$, we cluster the standardized scalar $\tilde{b}_{k1}\|\mathbf{A}_k W_1\|$, whose noise variance is approximately constant across particles.

**Cross-validated refinement.** At each refinement iteration, we partition the particles into two folds $\mathcal{J}$ and $\mathcal{J}'$ and, given the current labels $\{\tilde{c}_k\}$, reconstruct one masked volume per class from each fold, $\tilde{\rho}_j^{\mathcal{F}} = \mathcal{M}\,\mathcal{R}(\{(1, \check{\mathbf{y}}_k)\}_{k\in\mathcal{F},\ \tilde{c}_k = j})$ for $\mathcal{F} \in \{\mathcal{J}, \mathcal{J}'\}$. For $k \in \mathcal{F}$, we score class $j$ against the reconstruction from the opposite fold $\mathcal{F}^{\mathrm{c}}$, $s_{kj} = \mathcal{L}_{\mathrm{corr}}(\check{\mathbf{y}}_k, \mathbf{A}_k \tilde{\rho}_j^{\mathcal{F}^{\mathrm{c}}})$, where $\mathcal{L}_{\mathrm{corr}}$ is the band-limited normalized correlation of Section E.3 evaluated on $\Omega_{\mathrm{a}}$, and center the scores of each class within the fold, $\bar{s}_j^{\mathcal{F}} = |\mathcal{F}|^{-1} \sum_{k'\in\mathcal{F}} s_{k'j}$. We then update

$$\tilde{c}_k \leftarrow \arg\max_{j\in\{1,\dots,n\}} \big(s_{kj} - \bar{s}_j^{\mathcal{F}}\big), \qquad k \in \mathcal{F}. \tag{31}$$

Cross-validation avoids direct self-correlation between a particle and a reconstruction containing its noise, and the fold-specific centering compensates for class-dependent score offsets caused by unequal reconstruction precision. The refinement requires sufficient particles from every current class in both folds.

**Ensemble voting.** With few particles per conformation, a single residual-subspace fit can be unstable even for $n = 2$: different initializations and particle samples may recover a nuisance direction with explained energy comparable to that of the class mode. The class subspace, however, should recur across particle subsamples and frequency bands containing structural signal, whereas sample-specific nuisance directions are less stable. We therefore fit $B$ randomized replicates of Algorithm 2, each with a particle subset $\mathcal{I}_b$, a selected frequency band, and an independent initialization. Modes are fitted on $\mathcal{I}_b$, but coefficients, clustering, and refinement are evaluated for all particles. Each replicate produces labels $\{\tilde{c}_k^{(b)}\}$, defined only up to permutation.

To align the labelings, we select as reference the replicate with the highest mean pairwise agreement and, for every replicate $b$, choose the permutation $P_b$ that maximizes the trace of its contingency matrix with the reference, using the Hungarian algorithm (Kuhn, 1955); the aligned labels are $\check{c}_k^{(b)} = P_b(\tilde{c}_k^{(b)})$.

For $n = 2$, let $v_k^{(b)} \in \{-1, +1\}$ encode the aligned vote of replicate $b$, and let $\psi_b$ denote its balanced accuracy relative to the unknown labels. Under conditional independence of replicate errors given the class,

$$\mathrm{Cov}\big(v^{(b)}, v^{(b')}\big) = (1 - \beta^2)(2\psi_b - 1)(2\psi_{b'} - 1), \qquad b \neq b', \tag{32}$$

where $\beta = \pi_1 - \pi_2$. Thus, the off-diagonal covariance entries admit a rank-one factorization. Following the spectral meta-learner of Parisi et al. (2014), we obtain $\mathbf{u} \in \mathbb{R}^B$ from a spectral rank-one fit to these entries, orienting it relative to the reference replicate. The weighted vote

$\tilde{v}_k = \sum_{b=1}^{B} u_b v_k^{(b)}$ defines the binary assignment by its sign, and $\kappa_k = |\tilde{v}_k| / \sum_{b=1}^{B} |u_b|$ is the confidence score. A reversed replicate receives a negative weight, and a chance-level replicate is downweighted. The relative gap between the two leading eigenvalues of the vote covariance indicates whether the replicates agree on a common split.

For $n > 2$, we apply the same meta-learner separately to the aligned one-versus-rest indicators $v_{kj}^{(b)} = +1$ if $\check{c}_k^{(b)} = j$ and $-1$ otherwise. Let $\mathbf{u}^{(j)}$ denote the resulting class-specific weights and define the class scores $\bar{v}_{kj} = \sum_{b=1}^{B} u_b^{(j)} v_{kj}^{(b)} / \sum_{b=1}^{B} |u_b^{(j)}|$. We assign $\tilde{c}_k = \arg\max_j \bar{v}_{kj}$ and report $\kappa_k = \frac{1}{2}(\bar{v}_{k,(1)} - \bar{v}_{k,(2)})$, where $\bar{v}_{k,(1)} \geq \bar{v}_{k,(2)}$ are the two largest class scores. Class-specific weights allow a replicate that resolves only a subset of classes to contribute to those classes. Because replicates share particles, consensus maps, and masks, conditional independence is a working approximation; the margins and eigengaps are therefore diagnostic confidence measures rather than calibrated probabilities.

**Comparison with RECOVAR.** RECOVAR and B-RECOVAR both use consensus poses and CTFs, require no learned network, and infer structural variability from pose-aware volume statistics. RECOVAR estimates a regularized, noise-whitened volume covariance and embeds particles in a high-dimensional PCA space; B-RECOVAR instead estimates only the $n-1$ masked, band-limited class-deviation modes from consensus-subtracted residuals. For $n = 2$, this is a rank-one, uncorrected analogue of one-dimensional RECOVAR. Restricting the fit to the class subspace avoids relying on the leading direction of the full covariance, which can be dominated by sample-dependent nuisance variation at low particle counts.

B-RECOVAR further stabilizes the assignment by aggregating fits over particle subsamples, frequency bands, and initializations with the spectral meta-learner of equation 32; the normalized margin $\kappa_k$ provides a particle-level confidence score. It then applies the cross-fitted refinement of equation 31, avoiding direct self-correlation between a particle and its comparison volume. For a matched comparison, both methods use the same label-free readout and spectral aggregation; only B-RECOVAR uses cross-fitted refinement. RECOVAR randomizes its half-set split and downsampled box, whereas B-RECOVAR randomizes the particle subset, frequency band, and initialization.

**Implementation details.** We use $k_{\min} = 3$ and $k_{\max} \in \{24, 32, 40\}$ Fourier pixels, with $q = n-1$. Although the shells below 16 pixels contribute little directly to class discrimination in our data, they are retained to stabilize the reconstructions. The regularization is $\lambda_{\mathrm{w}} = 0.1$ times the mean diagonal entry of $\sum_k \mathbf{A}_k^* \mathbf{A}_k$. The solvent mask retains the $8\%$ of voxels with the largest values of $|\mathcal{F}^{-1}\hat{\rho}|$ and is smoothed with a Gaussian of standard deviation two voxels; mask fractions between $4\%$ and $15\%$ yield comparable results.

We use $T = 8$ alternating updates, $B = 24$ replicates, and independent $80\%$ particle subsamples; the three frequency bands are allocated evenly among replicates. The consensus, normalized residuals, and mask are computed once per band using all particles. Each replicate uses $L = 4$ refinement iterations. In the experimental data, with fewer than approximately $5{,}000$ particles per conformation, both the class direction and competing nuisance directions account for about $1.5\%$ of the residual energy, which motivates the ensemble aggregation. The dominant computational cost is $Bq$ single-mode fits, approximately 10 CPU-minutes for $10^4$ particles on a $D = 128$ grid with $n = 2$. Replicates can be evaluated in parallel; modes within a replicate are fitted sequentially because of deflation.

### E.5 Hyperparameters and implementation details

We use the following hyperparameters for all experiments.

- $k_{min} = 3$ and $k_{max} = 48$.
- $D = 128$ pixels.
- Due to memory constraints, at each optimization step, we guide using $\min(|\boldsymbol{\mathcal{Y}}|, 128)$ particles.
- Optimizer: Adam (Kingma & Ba, 2015)
- Learning rate: $0.06$

- Max gradient norm: $0.6$
- $\lambda_{\text{bond}} = 10^{-3}$ and $\lambda_{\text{collision}} = 0.09$
- $R$ (number outer iterations) was 3 for real examples and 5 for synthetic examples.
- **On the prior term.** In the formulation from Maddipatla et al. (2026a), the prior $\log p(\boldsymbol{\mathcal{Z}} \mid \mathbf{a})$ was essential: with $\boldsymbol{\mathcal{Z}}$ treated as a free variable, nothing else prevented the optimization from drifting into regions of embedding space that the Pairformer never produces. In the present parameterization this role is partly assumed by the reparameterization itself, since displacements in $\boldsymbol{\mathcal{H}}$ reach the conditioning variable only through $g$ and are therefore expressed in the basis of directions the trunk natively realizes.

# F Additional experiments

## F.1 Comparison with other experiment-guided AlphaFold methods

We compare Fold'EM with recent experiment-guided AlphaFold approaches that either optimize the final Pairformer embedding $\boldsymbol{\mathcal{Z}}$ (IT-Opt, Maddipatla et al., 2026a; see also Li et al., 2026) or steer the reverse diffusion process in coordinate space (coordinate-space guidance, Maddipatla et al., 2026b). To evaluate whether optimizing $\boldsymbol{\mathcal{H}}$ rather than $\boldsymbol{\mathcal{Z}}$ (Eq. 2) improves the ability to correct structural errors and accommodate larger conformational changes, we construct synthetic cryo-EM datasets from `9UGB` and `8U4B` at SNR $0.005$, and report C$\alpha$-RMSD and TM-score with respect to the experimental structure, as well as Fourier shell correlation (FSC) against the experimental density map at the $0.143$ threshold, across different numbers of particles (Figure A1). For `9UGB`, all three approaches correct the errors in the AlphaFold prediction, but Fold'EM performs better across all metrics. For `8U4B`, however, Fold'EM recovers the experimental structure, whereas IT-Opt and coordinate-space guidance degrade the initial AlphaFold prediction.

## F.2 Two-state heterogeneity on UvrA

**Label-free class assignment at low particle counts.** Figure A4 reports the accuracy of B-RECOVAR on the UvrA protein–DNA complex of EMPIAR-13396, a balanced mixture of the states `9GXQ` and `9GXR`, at 250 to $5{,}000$ particles per state, which corresponds to between $0.13\%$ and $2.6\%$ of the particles deposited for each state. At these sample sizes the leading direction of the RECOVAR covariance depends on the particular particles drawn as much as on the structural difference between the states. Across ten draws of the same size, the accuracy of RECOVAR ranges from chance to near-perfect, so that the median over draws is not representative of a single run (Figure A4A). B-RECOVAR is less accurate at $250$ particles per state, but from $500$ particles onward it is both more accurate and far less variable across draws, and at $5{,}000$ particles per state it reaches $92.5\%$ with a range across draws of under one percentage point. cryoDRGN, reported as the median over ten seeds, remains below B-RECOVAR at every sample size.

**Confidence of the assignment.** In addition to a label, the ensemble vote returns for every particle the normalized margin $\kappa_k$ (Appendix E.4), which we use as a confidence score. Figure A4B stratifies the particles by this score. Accuracy increases monotonically with confidence, from near chance in the lowest bins to $80\%$ in the highest bin, which contains $44\%$ of the particles. The score is therefore informative about which labels can be trusted, although it is not a calibrated probability. Fold'EM uses this ordering during heterogeneous structure inference, where each ensemble member is guided only by the more confident half of the particles assigned to it. Table A24 shows that discarding the less confident half yields a more diverse pair of modelled conformations, measured by the C$\alpha$ RMSD between the two members, and a higher TM-score of the inferred structures against the deposited models. This is consistent with the least confident particles being the most likely to be mislabelled; mislabelled particles pull the two members towards a common average rather than towards their respective states.

**Ablation of B-RECOVAR.** Figures A4C and D isolate the contribution of each component of the estimator. The base estimator is a rank-one fit to the consensus-subtracted residuals, equivalent to a one-dimensional RECOVAR fit without noise correction and without a solvent mask. It reaches $63.8\%$ accuracy, and the class scores of the two states overlap heavily ($d' = 0.54$). Restricting the fitted mode to the interior of a solvent mask removes the solvent-dominated part of the residual variance and raises the accuracy to $69.1\%$. Refining the labels by cross-validated comparison against per-class reconstructions, so that no particle is correlated with a volume that contains its own noise, raises it further to $72.1\%$. The largest gain comes from the ensemble. Fitting replicates on random particle subsets, frequency bands and initializations, and aggregating their labels by spectral voting, yields $80.6\%$ and doubles the separation between the states, from $d' = 1.13$ to $d' = 2.33$; the class direction recurs across replicates, whereas the nuisance directions that compete with it at these particle counts do not (Appendix E.4). The score distributions of Figure A4D show the same progression. Each component narrows the two distributions and moves them apart, and the ensemble turns a broad overlap into two separated modes.

### F.3 Continuous heterogeneity on IgG-1D

This section gives the protocol and the full results of the continuous-heterogeneity experiment of Section 4.

**Data.** The IgG-1D dataset of CryoBench (Jeon et al., 2024) consists of $100$ ground-truth atomic models, generated from PDB entry `1HZH` by rotating one Fab domain about the hinge through a closed $360^\circ$ trajectory, and $25{,}000$ synthetic particles at SNR $0.01$, $250$ per state, with ground-truth poses and per-particle CTFs.

**Conformational coordinates.** We embed each particle in three label-free conformational coordinates, namely the latent of a cryoDRGN (Zhong et al., 2021) VAE trained on the stack, the leading principal components of RECOVAR (Gilles & Singer, 2025), and the residual-subspace embedding of B-RECOVAR (Appendix E.4). Each coordinate defines one arm of the experiment; the arms differ only in the coordinate used to select the particles of each state.

**Anchors and interpolation.** We fix eight *anchor* conformations at equispaced positions along the ring and, for each anchor, optimize one structure against the $250$ particles nearest to it in the conformational coordinate of that arm. Each arm thus yields eight Pairformer states $\mathbf{H}_1, \ldots, \mathbf{H}_8$, optimized independently against disjoint particle sets, with no term coupling them. An intermediate conformation at position $\alpha$ along the ring is obtained as the convex combination $\mathbf{H}(\alpha) = \sum_k c_k(\alpha)\, \mathbf{H}_k$ of the two neighbouring anchors, with piecewise-linear coefficients $c_k$. For each of the $100$ ground-truth states we decode $\mathbf{H}(\alpha)$ through a full reverse-diffusion trajectory, with a light coordinate-space refinement against the $250$ particles that the same coordinate places nearest to that state, and relax the result (Appendix E.2.3).

**Evaluation.** We report the TM-score against the ground-truth model of the requested state (Appendix E.2.4) and the Fourier NCC between the rendered structure and the particles of that state, expressed as a fraction of the NCC of the ground-truth model against the same particles. Over the full ring, the $100$ structures of the B-RECOVAR arm, all decoded from interpolated states, attain a median NCC of $92\%$ of the ground-truth value and a median TM-score of $0.93$, on par with the eight anchors themselves ($89\%$ and $0.91$). The cryoDRGN coordinate reaches $85\%$ and $0.89$, and the RECOVAR coordinate $90\%$ and $0.92$ (Figure 4). Figure A5 shows the latent geometry of the optimized states and the decoded structures for each coordinate, and Table A23 the stereochemical quality of the anchors and of the interpolated structures.

## G Related work

**Map reconstruction.** Reconstructing an electrostatic potential (ESP) map at a resolution sufficient for atomic model building typically requires tens to hundreds of thousands of particles. RELION (Zivanov et al., 2018) and cryoSPARC (Punjani et al., 2017) estimate the particle poses and the map jointly within a Bayesian framework. For heterogeneous data, these pipelines partition the

particles into $K$ discrete classes and reconstruct one map per class. cryoDRGN (Zhong et al., 2021) instead learns a continuous latent space over the particles with a variational autoencoder and decodes each latent into a three-dimensional density, and RECOVAR (Gilles & Singer, 2025) estimates the covariance of the underlying volume and embeds the particles in its leading principal components. Both take the poses from a prior consensus reconstruction. CryoDRGN-AI (Levy et al., 2025) extends this approach to the *ab initio* setting by estimating the poses and the heterogeneous structural distribution jointly from the images.

**Model building and experimental guidance.** Building an atomic model from a reconstructed map is a separate downstream problem. ModelAngelo (Jamali et al., 2024) combines sequence and density information in a graph-based neural architecture to locate residues, infer their connectivity and assemble an atomic model. More recent approaches use a protein structure prediction model (Abramson et al., 2024) as a structural prior and guide it with the experimental density. CryoBoltz (Raghu et al., 2026) steers Boltz (Passaro et al., 2025) toward conformations consistent with an ESP map, but requires an input structure roughly aligned to the map to establish the correspondence used during guidance. Maddipatla et al. (2026b) instead guide AlphaFold3 *ab initio*, from sequence alone, toward structures consistent with experimental ESP maps. Embed-Opt (Li et al., 2026) and IT-Opt (Maddipatla et al., 2026a) take a complementary inference-time approach and optimize the latent representations of AlphaFold to produce structures consistent with synthetic ESP maps.

The closest prior work is CoCoFold (Liao et al., 2026), which constrains AlphaFold directly with individual cryo-EM particle images. CoCoFold performs target-specific fine-tuning of the AlphaFold-Multimer structure module, with the Evoformer frozen, using an FRC loss between experimental particles and differentiable projections; it assumes upstream pose/CTF estimates and uses a coarse reconstruction to establish a fixed coordinate frame.