CryoGeo: Latent Pose Equivariance for Amortized Ab Initio Cryo-EM Reconstruction

Zhongyang Li[1] **Diedong Feng**[1] **Shen Cheng**[2] **Bing Zeng**[1] **Shuaicheng Liu**[1*]
[1]University of Electronic Science and Technology of China
[2]Independent Researcher
{lizhongyang,fengdiedong}@std.uestc.edu.cn
sumching51@gmail.com
{eezeng,liushuaicheng}@uestc.edu.cn

Abstract

Cryo-electron microscopy (cryo-EM) reconstructs 3D macromolecular structures from noisy 2D particle images by jointly estimating the density and projection pose of each particle. Amortized inference can greatly accelerate this process, but remains vulnerable to pose collapse because known image-space symmetries are often learned only implicitly through reconstruction losses. We ask whether this instability can be mitigated by enforcing cryo-EM transformation laws directly in latent projection-pose space. We propose CryoGeo, a geometry-constrained framework for amortized *ab initio* cryo-EM reconstruction. CryoGeo enforces latent pose equivariance by requiring in-plane transformations to induce valid group actions on the predicted projection pose: rotations act as right multiplications on the predicted $SO(3)$ orientation, while shifts transform frame-consistently. CryoGeo uses an output-level group-action objective to couple transformed observations to renderer queries constructed by applying the physical pose action to each predicted base pose. Across synthetic and real cryo-EM datasets, CryoGeo improves pose consistency, mitigates collapse-prone behavior, and yields more reliable amortized *ab initio* reconstructions than unconstrained variants and neural baselines. These results show that latent pose equivariance turns in-plane symmetry from an implicitly learned nuisance factor into a self-supervised geometric constraint for stable, efficient amortized structural reconstruction. Code will be released at `https://github.com/Jacker-Li/CryoGeo`.

1 Introduction

Cryo-electron microscopy (cryo-EM) has become central for determining the three-dimensional structures of macromolecular complexes from large collections of noisy two-dimensional particle images [1, 2, 3]. The computational core of single-particle cryo-EM is an inverse problem requiring recovery of the unknown 3D electrostatic potential of a molecule while estimating the projection pose of each particle image [4, 5]. This problem is especially challenging in the *ab initio* setting because neither an initial 3D template nor ground-truth poses are available. Severe image noise, unknown contrast transfer functions, and the non-convex pose-density coupling make reliable initialization a major bottleneck in high-resolution structural determination [6].

Classical cryo-EM pipelines handle this difficulty through iterative pose search and refinement [7, 8, 9]. While these search-based procedures are robust, they are computationally expensive because pose inference is performed separately for each image [10]. As dataset sizes grow [11], this per-particle optimization cost has motivated amortized inference methods that learn a feed-forward neural

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

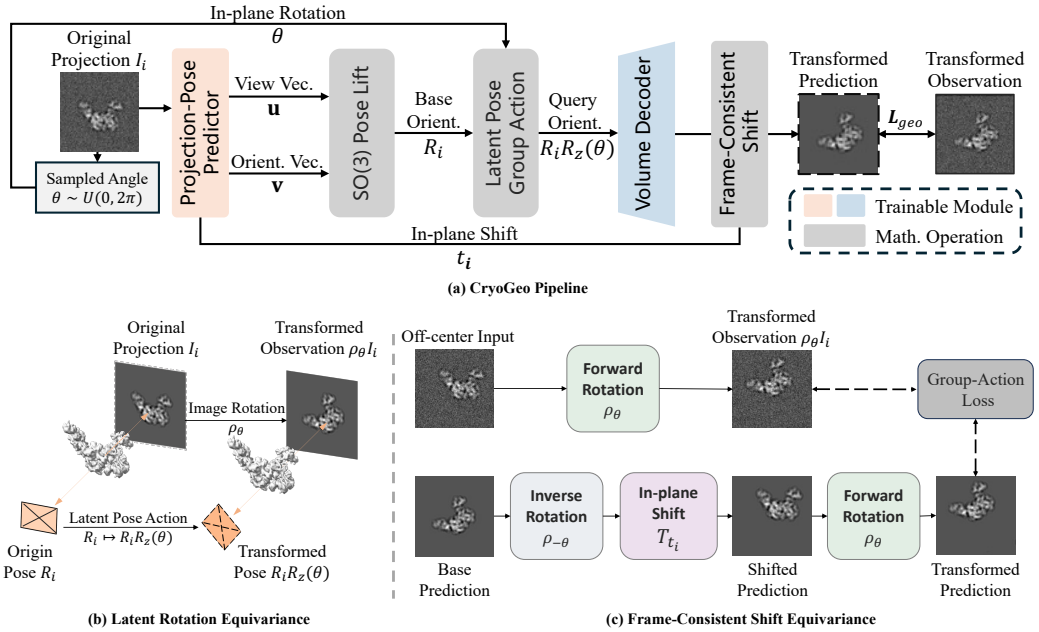


Figure 1: **Overview of CryoGeo.** (a) CryoGeo constrains amortized reconstruction by enforcing latent pose equivariance. The projection-pose predictor maps an input projection I_i to a base orientation $R_i \in SO(3)$ and an in-plane shift t_i. A sampled in-plane rotation θ induces the latent group action $R_i \mapsto R_i R_z(\theta)$, yielding the query orientation used by the volume decoder. The predicted shift is applied through the frame-consistent operator $\rho_\theta T_{t_i} \rho_{-\theta}$, and the transformed prediction is matched to the transformed observation $\rho_\theta I_i$ using the group-action loss $\mathcal{L}_{geo}$. (b) Image-plane rotation induces a right action on the latent orientation, mapping R_i to $R_i R_z(\theta)$. (c) Image-plane shift correction is performed in the appropriate rotated coordinate frame through the conjugated shift operator $\rho_\theta T_{t_i} \rho_{-\theta}$.

mapping from particle images to latent variables including pose and shift to significantly accelerate the reconstruction process [12, 13, 14].

However, fully amortized cryo-EM reconstruction remains highly vulnerable to pose collapse. In the cold-start *ab initio* regime, the neural encoder receives supervision only through reconstruction losses induced by an initially unreliable 3D volume. When the target structure contains pseudo-symmetry or the signal-to-noise ratio is severely low, unconstrained encoders often assign inconsistent or degenerate poses to physically related projections. Existing neural methods attempt to mitigate this instability using data augmentation, semi-amortized refinement, or pose search [13, 15, 16], but these strategies fail to directly constrain the geometry of the predicted pose manifold.

We argue that pose collapse in amortized cryo-EM is fundamentally a *latent geometry failure*. Known physical transformations in image space are not guaranteed to induce corresponding transformations in the predicted projection-pose space. Specifically, if a projection image is rotated by an angle θ in the image plane, its underlying 3D viewing orientation must transform via a deterministic right action around the camera optical axis, and its in-plane shift must transform within the rotated coordinate frame. Standard reconstruction losses do not guarantee that a feed-forward encoder respects these physical transformation laws, leaving the pose manifold under-constrained.

To address this gap, we introduce **CryoGeo**, a geometry-constrained amortized framework for *ab initio* cryo-EM reconstruction. CryoGeo tackles pose collapse by enforcing *latent pose equivariance*. Instead of treating in-plane symmetries as mere data augmentation or imposing equivariance only on intermediate image features, CryoGeo requires known image-plane transformations to act directly on the predicted output pose. During training, we explicitly sample an in-plane rotation, apply the corresponding right action to the predicted $SO(3)$ orientation, and shift the coordinates in a frame-consistent manner. The model is then self-supervised to render a projection from this transformed latent pose that matches the rotated observation.

We instantiate this principle using a valid $SO(3)$ pose parameterization via differentiable Gram-Schmidt orthogonalization [17] and an efficient real-valued Hartley-domain voxel grid for rendering [18]. This design avoids complex-valued Fourier arithmetic while allowing the geometric constraint to be evaluated efficiently at scale. These components enable practical optimization, while the central constraint acts on the rendered projection pose itself. Extensive experiments on synthetic benchmarks including Spike and Spliceosome and experimental EMPIAR datasets demonstrate that CryoGeo consistently improves pose accuracy and global density recovery. These improvements are particularly evident under severe pseudo-symmetric ambiguity, all while preserving the computational efficiency of feed-forward amortized inference.

Our main contributions are summarized below.

- We identify pose collapse in amortized cryo-EM reconstruction as a latent geometry failure, highlighting the necessity of preserving image-space transformations within the predicted projection-pose manifold.
- We formulate latent pose equivariance for cryo-EM, mathematically linking in-plane image rotations to right actions on the predicted $SO(3)$ orientation and frame-consistent shift transformations.
- We introduce an output-level self-supervised group-action objective that constrains the final predicted projection pose used for rendering, distinguishing CryoGeo from methods that impose equivariance only on encoder features or use rotations as ordinary data augmentation.
- We demonstrate that CryoGeo achieves strong pose accuracy and reconstruction robustness on challenging synthetic and real datasets without sacrificing inference speed.

2 Related work

CryoGeo builds on three lines of work encompassing classical cryo-EM reconstruction, amortized neural inference, and equivariant representation learning. We review these directions with an emphasis on how existing methods handle pose ambiguity and image-plane symmetries.

Iterative pose search versus amortized inference. Classical cryo-EM methods reconstruct structures using iterative pose search and refinement [19, 20, 7, 8] or common-line approaches [21, 22, 23]. Neural approaches include adversarial reconstruction and amortized pose inference [24, 25, 26, 27], with the latter learning a feed-forward mapping from images to poses. Researchers have further extended these approaches to end-to-end orientation estimation [12, 28], fast neural *ab initio* reconstruction [14, 29], and continuous heterogeneous reconstructions [30, 31, 32, 33, 34]. While highly effective for clean data, fully amortized models such as CryoAI [13] remain vulnerable to pose collapse under severe noise or pseudo-symmetry. Semi-amortized methods draw inspiration from multiple-choice learning [35, 36] to mitigate this instability. CryoSPIN [15] combines amortized prediction with local gradient-based pose refinement, while CryoDRGN2 [16] uses hierarchical pose search. Both approaches retain the computational cost of per-particle pose search or refinement. CryoGeo differs by remaining fully amortized while mitigating collapse-prone behavior through an explicit output-level geometric constraint on the predicted projection pose.

Equivariant neural networks. Spatial transformer networks [37] learn spatial transformations, while equivariant architectures [38, 39] constrain how features transform under known symmetry groups. Feature-level equivariance alone does not determine the physical transformation law of the pose readout. Prior cryo-EM methods incorporate equivariance into pose inference [40, 41]. CryoGeo applies the physical group action to each predicted base pose and couples the resulting renderer queries to transformed observations through its self-supervised objective.

Neural volume representations. Neural cryo-EM methods represent 3D volumes using either coordinate-based implicit networks [42, 43, 44, 30, 13] or explicit frequency grids [45]. Implicit networks offer flexibility but suffer from spectral bias in high-frequency recovery [46, 47] and demand computationally expensive projection rendering. Standard explicit Fourier grids store complex-valued coefficients, with operations involving both real and imaginary components [48]. CryoGeo adopts an explicit real-valued Hartley-domain voxel grid to avoid complex-valued volume parameters. This specific design enables fast and differentiable projection-slice rendering to efficiently evaluate the latent pose equivariance objective at scale.

3 Methodology

CryoGeo is designed around the principle that known image-plane transformations should induce valid, predictable transformations on the latent cryo-EM projection pose. Rather than relying on reconstruction loss to discover this relation implicitly, CryoGeo enforces it directly on the predicted pose manifold. We first formulate the projection model, derive latent pose equivariance, and instantiate it through a self-supervised group-action objective with frame-consistent shift handling.

3.1 Cryo-EM projection model

The physical image formation process in single-particle cryo-EM dictates the relationship between a 3D macromolecular structure and its 2D observations [4, 49]. Let $V : \mathbb{R}^3 \to \mathbb{R}$ denote the unknown electrostatic potential of the target molecule. An observed particle image I_i is generated by projecting V under a latent projection pose consisting of a 3D orientation $R_i \in SO(3)$ and an in-plane shift $t_i \in \mathbb{R}^2$. The standard generative model is expressed as

$$I_i = \mathcal{A}_{C_i,t_i}\big(\Pi(V, R_i)\big) + \eta_i$$

where Π is the projection operator, $\mathcal{A}_{C_i,t_i}$ applies the particle-specific CTF C_i and shift t_i, and η_i denotes noise. CryoGeo jointly optimizes the density V and an amortized pose predictor that estimates the latent projection pose from each input image.

3.2 Latent pose equivariance

The key geometric fact used by CryoGeo is that an in-plane image rotation has a deterministic action on the latent camera pose. Let ρ_θ denote a 2D in-plane rotation and let $R_z(\theta) \in SO(3)$ denote the corresponding rotation around the camera optical axis. For the ideal continuous projection operator, the relationship is defined as

$$\rho_\theta \Pi(V, R) = \Pi(V, RR_z(\theta)).$$

Thus, rotating the observed projection should induce a right multiplication on the predicted $SO(3)$ orientation. If $f_\phi(I) = (R, t)$ is the predicted projection pose, the desired output-level equivariance requires

$$f_\phi(\rho_\theta I) \approx f_\phi(I) \circ g_\theta, \qquad R(\rho_\theta I) \approx R(I)R_z(\theta),$$

where g_θ denotes the valid group action on the projection pose. CryoGeo enforces this physical transformation law at the renderer-query level by analytically coupling transformed queries to the same predicted base pose. The projection identity is exact for the continuous pure projection model. In the practical pipeline with CTF modulation, finite-grid interpolation, shifts, and noise, CryoGeo enforces an equivariance-consistent approximation.

3.3 Frame-consistent shift transformation

Latent pose equivariance must explicitly account for in-plane translations to maintain geometric consistency. The predicted shift t_i is defined in the coordinate frame of the original image. Rotating a non-centered particle image also rotates the particle's displacement from the image center. For example, a horizontal displacement becomes vertical after a quarter-turn, so the shift action on the rendered query must transform consistently with the observation. Applying the original shift unchanged to a projection rotated by θ would therefore translate it along incorrect axes. CryoGeo resolves this by applying the shift through a conjugated operator formulated as

$$T_{t_i}^{(\theta)} = \rho_\theta T_{t_i} \rho_{-\theta},$$

where T_{t_i} is the translation operator induced by t_i. This operator first maps the query projection back to the original image frame, applies the predicted shift, and then rotates the result forward. The full projection-pose action is therefore derived as

$$R_i^{(\theta)} = R_i R_z(\theta), \qquad T_{t_i}^{(\theta)} = \rho_\theta T_{t_i} \rho_{-\theta}.$$

This formulation extends latent pose equivariance from orientation alone to the full projection pose.

3.4 Self-supervised group-action objective

CryoGeo enforces the defined transformation law without relying on ground-truth poses. Given an input image I_i, the pose predictor estimates a base orientation R_i and an in-plane shift t_i. During training, we sample $\theta \sim \mathcal{U}(0, 2\pi)$ and construct the query orientation $R_i R_z(\theta)$. Let $\mathcal{D}_h^0$ denote the discrete projection renderer before CTF modulation and shift application, and let $\rho_{\theta,h}$ denote digital image rotation on the finite sampling grid h. The transformed prediction is

$$\widehat{I}_{i,\theta} = \rho_{\theta,h} T_{t_i} C_i \rho_{-\theta,h} \mathcal{D}_h^0(V, R_i R_z(\theta)).$$

The corresponding self-supervised group-action loss is

$$\mathcal{L}_{\text{geo}} = \mathbb{E}_{I_i,\theta}\left[\left\|\widehat{I}_{i,\theta} - \rho_{\theta,h} I_i\right\|_2^2\right].$$

Here, the particle-specific CTF C_i and the predicted shift operator T_{t_i} act in the original detector frame. The renderer first produces a projection at the query orientation $R_i R_z(\theta)$, before CTF modulation or shift application. We apply $\rho_{-\theta,h}$ to rotate this projection back to the original detector frame. In that frame, C_i modulates the projection and T_{t_i} applies the translation predicted from the original image. We then apply $\rho_{\theta,h}$ to rotate the result forward, so that the rendered prediction and the target $\rho_{\theta,h} I_i$ are compared in the same rotated frame. For a fixed base prediction, changing θ changes the query orientation and the two digital rotations while retaining the same C_i and t_i. Appendix B.3 details this operator sequence. This objective ties the image-space transformation to the corresponding physical pose transformation, with all transformed renderer queries analytically coupled to the same predicted base pose. Appendix C.3 analyzes the relation to reconstruction and rotation augmentation, together with the resulting local mechanism for escaping pose collapse.

3.5 Pose parameterization and differentiable rendering

CryoGeo represents predicted orientations using a differentiable $SO(3)$ parameterization. The amortized pose predictor outputs a view vector $u_i \in \mathbb{R}^3$, an orientation vector $v_i \in \mathbb{R}^3$, and an in-plane shift t_i from each input image. The view vector defines the optical axis, while the orientation vector determines the in-plane basis. A differentiable Gram-Schmidt layer maps nonzero, nonparallel u_i and v_i to a valid rotation matrix [17]. This orthogonalization process is defined as

$$r_3 = \frac{u_i}{\|u_i\|}, \quad r_1 = \frac{v_i - (v_i^\top r_3) r_3}{\|v_i - (v_i^\top r_3) r_3\|}, \quad r_2 = r_3 \times r_1, \quad R_i = [r_1, r_2, r_3] \in SO(3).$$

This parameterization ensures that the right action $R_i \mapsto R_i R_z(\theta)$ remains well-defined during end-to-end training. We employ a steerable convolutional backbone as an inductive bias, but the core equivariance constraint in CryoGeo is imposed at the output projection-pose level rather than strictly on intermediate features. Detailed architecture and shift-head implementation are provided in the supplementary material.

CryoGeo utilizes an explicit real-valued Hartley-domain voxel grid $V_{\mathcal{H}}$ [18] to achieve computationally efficient differentiable rendering. Projections are synthesized by extracting central slices according to the Hartley slice theorem. The frequency representation is computed as

$$\mathcal{H}(\Pi(V, R))(k) = V_{\mathcal{H}}\left(R[k_x, k_y, 0]^\top\right),$$

where $k = (k_x, k_y)$ represents the 2D spatial frequency [50, 45]. The rendered slice is modulated by the particle-specific CTF before being compared with the observed image. Full details regarding Hartley-domain shift modulation, CTF implementation, interpolation, and preprocessing are provided in the supplementary material.

4 Experiments

We evaluate CryoGeo on synthetic and experimental cryo-EM reconstruction tasks. The experiments are designed to answer three questions. First, does latent pose equivariance improve amortized *ab initio* reconstruction quality under noisy and ambiguous projection settings? Second, does the proposed output-level geometric constraint improve pose accuracy compared with feature-level equivariance alone? Third, does CryoGeo preserve the computational advantage of amortized inference? We report reconstruction quality, pose accuracy, ablation results, and computational efficiency using standard benchmarks and evaluation protocols detailed in the supplementary material.

Table 1: Main reconstruction results on synthetic and real cryo-EM datasets. Synthetic results are evaluated by map-to-map FSC against ground truth at FSC $= 0.5$, and real-data results by gold-standard half-map FSC at FSC $= 0.143$. Resolutions are reported in Å; lower is better. Low-frequency threshold crossings indicate collapse cases.

Dataset / Setting	CryoAI [13]	CryoDRGN2 [16]	CryoSPIN [15]	CryoSPARC [8]	**CryoGeo (Ours)**
Synthetic map-to-map FSC against ground truth					
Spike, Clean	9.401	6.650	7.175	11.360	**6.491**
Spike, SNR $= 0.1$	15.147	10.487	11.854	11.854	**9.737**
Spliceosome, Clean	10.457	11.546	10.457	12.048	**9.724**
Spliceosome, SNR $= 0.1$	92.336	21.317	20.525	15.396	**14.979**
Real-data gold-standard FSC from half maps					
EMPIAR-10028	7.660	**7.540**	7.660	**7.540**	**7.540**
EMPIAR-11233	7.374	6.523	7.374	9.692	**5.561**

4.1 Experimental setup

We evaluate CryoGeo on both synthetic and real cryo-EM datasets. For synthetic experiments, we use the SARS-CoV-2 Spike protein [51] and the Spliceosome complex [52] sourced from the Protein Data Bank [53]. These two structures test complementary reconstruction regimes. Spike contains strong pseudo-symmetric ambiguity, while Spliceosome provides a highly asymmetric target. For each structure, we generate 50,000 projection images at a spatial size of 128×128 using randomly sampled orientations and particle-specific contrast transfer functions [54]. We evaluate both clean projections and noisy projections at $\mathrm{SNR} = 0.1$.

We additionally use a synthetic heat shock protein (HSP) dataset based on Hsp90 [55] to evaluate shift prediction and frame-consistent shift handling.

We further evaluate CryoGeo on two experimental cryo-EM datasets from the EMPIAR public archive [56]. EMPIAR-10028 contains 80S ribosome particles [57], and EMPIAR-11233 contains TRPM8 ion-channel particles [58]. Following standard preprocessing, all particles are downsampled to 128×128 before reconstruction. For real data, we split particles into two independent halves and report gold-standard FSC from independently reconstructed half maps.

We compare CryoGeo against representative iterative, amortized, and semi-amortized baselines. CryoSPARC [8] represents a widely used iterative optimization method. CryoDRGN2 [16] represents a neural reconstruction method with pose search. CryoAI [13] serves as a fully amortized neural baseline, and CryoSPIN [15] represents a semi-amortized method that combines amortized prediction with subsequent pose refinement. All methods are evaluated under matched image resolution and preprocessing settings. Additional baseline configurations, preprocessing details, and FSC protocols are provided in the supplementary material.

4.2 Reconstruction quality

Table 1 summarizes the FSC-based reconstruction quality across all datasets. CryoGeo achieves the highest resolution across all synthetic settings and the EMPIAR-11233 dataset, matching the strongest baseline on EMPIAR-10028. The performance gap widens significantly under extreme noise conditions. For instance, on the noisy Spliceosome dataset, the unconstrained CryoAI baseline experiences severe pose collapse, degrading to 92.336 Å. In contrast, CryoGeo maintains a stable FSC crossing at 14.979 Å.

Figures 2 and 3 visually corroborate these quantitative trends on synthetic datasets, demonstrating that CryoGeo preserves clearer structural boundaries and finer local density under noisy conditions. Conversely, fully amortized baselines like CryoAI [13] lose high-frequency details or collapse into blurred structures. This degradation is particularly evident on the pseudo-symmetric Spike benchmark [51], where ambiguous views induce inconsistent pose assignments.

Figures 4 and 5 present the reconstruction results on experimental cryo-EM data. On EMPIAR-11233, CryoGeo achieves the best gold-standard FSC resolution and recovers well-defined structural features.

Table 2: High-precision view-direction ablation on synthetic datasets. Latent pose equivariance remains effective with a standard encoder, showing that the main stabilizing signal comes from constraining the predicted pose manifold rather than from a specialized equivariant feature extractor.

Model	Med. (°) ↓	RMSE (°) ↓	Acc@10° (%) ↑	Acc@5° (%) ↑
Spliceosome				
Standard Enc.	3.94	4.91	97.8	66.2
Equivariant Enc. only	2.64	3.27	**100.0**	90.2
Standard Enc. + Latent Pose Eq.	1.61	2.31	**100.0**	97.1
CryoGeo (Ours)	**1.52**	**2.22**	**100.0**	**97.5**
Spike				
Standard Enc.	6.82	11.78	70.2	33.6
Equivariant Enc. only	3.18	4.37	98.0	77.2
Standard Enc. + Latent Pose Eq.	2.29	3.95	97.9	81.8
CryoGeo (Ours)	**2.10**	**3.68**	**98.5**	**84.6**

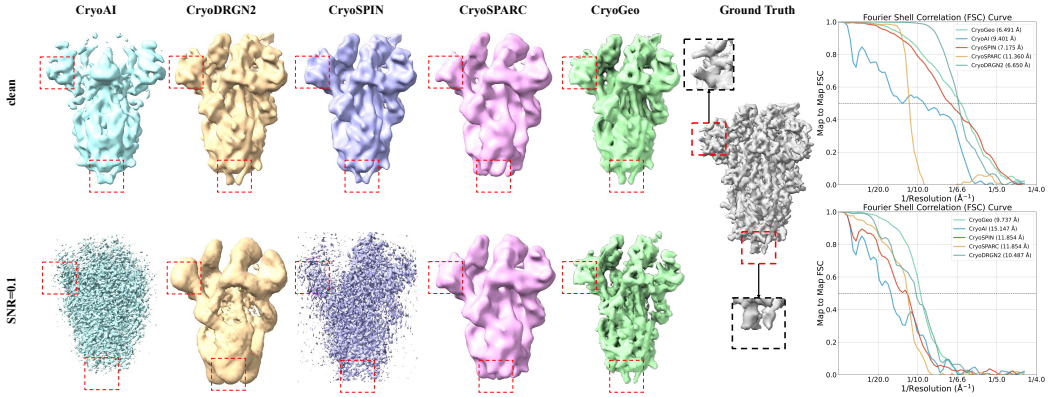


Figure 2: **Reconstruction results on the simulated Spike dataset.** The figure compares reconstructed 3D density maps and map-to-map FSC curves under clean and SNR $= 0.1$ conditions. Since Spike is a synthetic benchmark, reconstruction quality is evaluated by computing map-to-map FSC against the ground-truth density, with resolution reported at the FSC $= 0.5$ threshold. Under clean projections, most methods recover the global Spike morphology. Under the noisy setting, unconstrained amortized baselines show stronger degradation in structural details and FSC profiles, while CryoGeo maintains more stable density recovery and a more favorable FSC curve. The result highlights the benefit of enforcing latent pose equivariance on a pseudo-symmetric structure where pose ambiguity is severe.

On EMPIAR-10028, CryoGeo produces a coherent global reconstruction with local density continuity comparable to that of the strongest baselines. These results indicate that latent pose equivariance remains effective on experimental cryo-EM data with microscope noise, imperfect particle centering, and particle-specific CTF variation.

4.3 Ablation study: latent pose equivariance

Table 2 details the high-precision view-direction ablation results. Across both datasets, an equivariant encoder improves pose accuracy over the standard baseline, while explicitly enforcing latent pose equivariance provides the largest improvement. For example, on the highly ambiguous Spike dataset, the standard encoder combined with latent pose equivariance reduces the root mean square error to 3.95°, outperforming the equivariant-only encoder at 4.37°. The full CryoGeo model further reduces the error to 3.68°, with 84.6% of predictions having an angular error below 5°.

These empirical results demonstrate that feature-level equivariance and latent pose equivariance act in a complementary manner. These results suggest that the primary stabilizing effect comes from constraining the output projection-pose manifold rather than solely relying on an equivariant feature extractor.

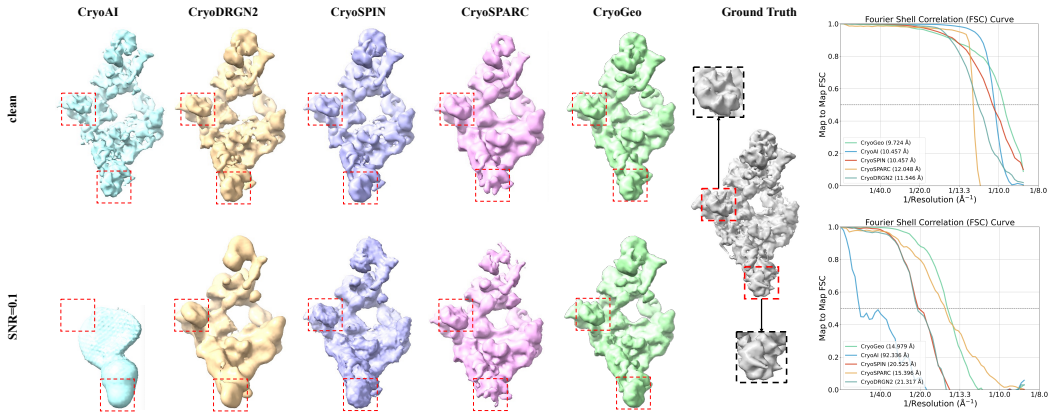


Figure 3: **Reconstruction results on the simulated Spliceosome dataset.** The figure compares reconstructed 3D density maps and map-to-map FSC curves under clean and SNR $= 0.1$ conditions. Because Spliceosome is a synthetic benchmark, reconstruction quality is evaluated by computing map-to-map FSC against the ground-truth density, with resolution reported at the FSC $= 0.5$ threshold. The asymmetric structure provides a complementary setting to the pseudo-symmetric Spike benchmark. Under the noisy setting, several baselines show degraded structural details and weaker FSC profiles, while CryoGeo maintains more stable recovery of the global asymmetric architecture and local density features.

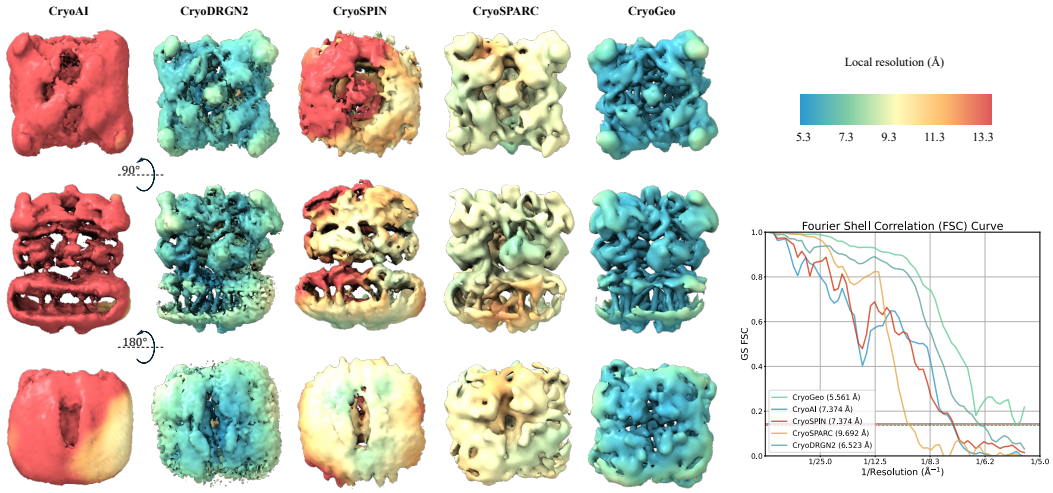


Figure 4: **Ab initio reconstruction of detergent-solubilized TRPM8 (EMPIAR-11233).** Left: reconstructed 3D density maps shown from three viewing directions and colored by local resolution. Right: gold-standard FSC curves computed from independently reconstructed half maps and evaluated using the FSC $= 0.143$ criterion. CryoGeo yields a stable global reconstruction with well-defined structural features and achieves the highest FSC-based resolution among the compared methods.

Frame-consistent shift handling. We evaluate shift handling on HSP using two variants of CryoGeo. Zero shift fixes in-plane translations to zero, while Direct shift predicts translations without applying the frame-consistent shift transformation. Table 3 reports shift RMSE, the mean angular error of the full $SO(3)$ orientation, and map-to-map FSC resolution at FSC $= 0.5$. Zero shift and Direct shift yield mean orientation errors of $28.80°$ and $28.86°$, respectively, compared with $0.78°$ for CryoGeo. Direct shift yields a shift RMSE of 2.83 pixels, compared with 0.70 pixels for CryoGeo.

Figure 6 presents the reconstructed maps and their map-to-map FSC curves against the ground-truth density. At the FSC $= 0.5$ threshold, CryoGeo achieves a resolution of 3.765 Å, while Zero shift and

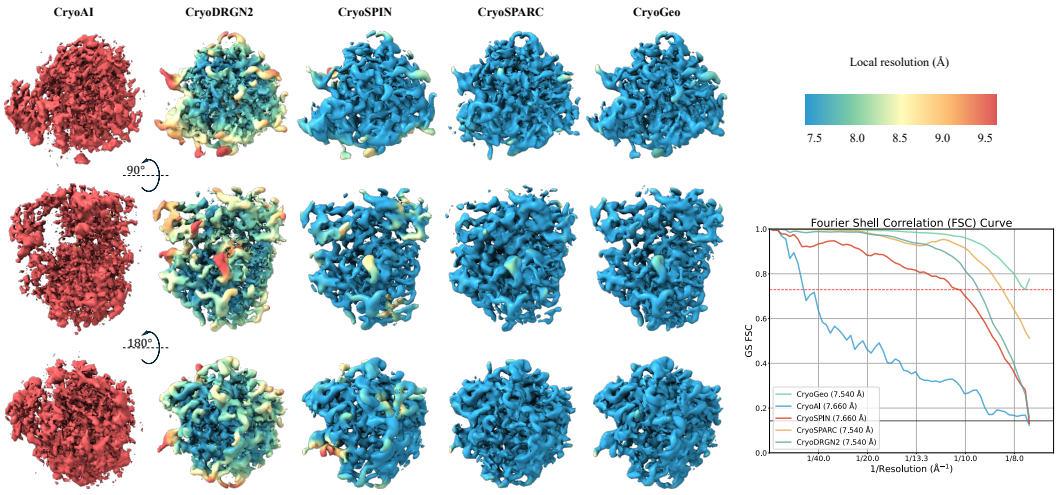


Figure 5: **Ab initio reconstruction of the 80S ribosome (EMPIAR-10028).** Left: reconstructed 3D density maps colored by local resolution. Right: gold-standard FSC curves computed from independently reconstructed half maps and evaluated using the FSC $= 0.143$ criterion. CryoGeo yields a stable global reconstruction with clear local density continuity across major ribosomal domains.

Table 3: Ablation of shift prediction and frame-consistent shift handling on HSP. We report shift RMSE, the mean angular error of the full $SO(3)$ orientation, and map-to-map FSC resolution against the ground-truth density at FSC $= 0.5$. Lower is better for all metrics. A dash indicates that shift RMSE is not reported.

Variant	Shift RMSE (px) ↓	Mean full $SO(3)$ error (°) ↓	FSC resolution (Å) ↓
Zero shift	–	28.80	4.683
Direct shift	2.83	28.86	5.053
CryoGeo	**0.70**	**0.78**	**3.765**

Direct shift achieve 4.683 Å and 5.053 Å, respectively. These results show that frame-consistent shift handling improves both pose estimation and reconstruction quality on HSP.

In Direct shift, each particle's predicted translation is applied unchanged to all its rotated queries, although the target translation rotates with the image. Under uniformly sampled in-plane rotations, a local quadratic approximation favors a near-zero shared translation estimate, providing an explanation for the similar orientation errors of Direct shift and Zero shift.

4.4 Computational efficiency

Table 4 reports runtime and peak GPU memory on 50,000 non-centered particle images of size 128×128, with joint orientation and in-plane translation estimation. All measurements use a batch size of 64 on a single NVIDIA RTX 4090. We evaluate CryoSPIN and CryoGeo with matched configurations of 15, 24, and 32 pose heads. For both runtime and peak GPU memory, we report the mean of three measurements, with peak memory measured over the complete training pipeline.

Across the matched head configurations, CryoGeo's amortized epoch time ranges from 1.96 to 2.83 minutes, with peak GPU memory between 7.18 and 9.97 GB. For CryoSPIN, the corresponding ranges are 2.74 to 3.03 minutes and 7.77 to 10.47 GB. Its subsequent unamortized pose-refinement stage takes an average of 2.45 minutes per epoch. CryoGeo performs reconstruction with amortized training alone.

Qualitative training evolution, baseline convergence details, and pose-evaluation procedures are provided in the supplementary material.

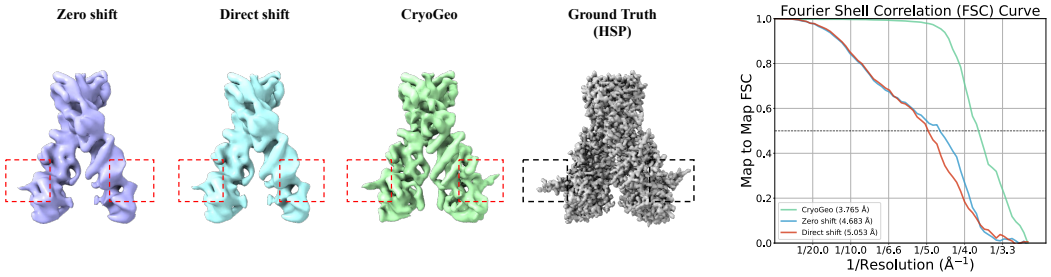


Figure 6: **Effect of shift handling on HSP reconstruction.** Left: reconstructed density maps for Zero shift, Direct shift, and CryoGeo, alongside the ground-truth map. Dashed boxes mark corresponding local regions. Right: map-to-map FSC curves against the ground-truth density, with the FSC $= 0.5$ threshold indicated by the horizontal dashed line.

Table 4: Runtime and peak GPU memory on 50,000 non-centered particle images of size 128×128, using a batch size of 64 on a single NVIDIA RTX 4090. All methods estimate both orientations and in-plane translations. All runtime and memory values are means of three measurements. For each measurement, peak GPU memory is the maximum across all training stages. The table reports amortized training time per epoch. The subsequent unamortized pose-refinement stage of CryoSPIN takes an average of 2.45 minutes per epoch. CryoGeo requires no subsequent unamortized refinement stage.

Model	Heads	Peak VRAM (GB) ↓	Amortized epoch time (min) ↓
CryoAI [13]	–	23.50	6.01
CryoSPIN [15]	15	7.77	2.74
	24	8.74	2.97
	32	10.47	3.03
CryoGeo (Ours)	15	7.18	1.96
	24	8.38	2.42
	32	9.97	2.83

5 Conclusion

In this paper, we presented CryoGeo to stabilize amortized *ab initio* cryo-EM reconstruction. We frame pose collapse as a latent geometry failure where known image-plane transformations are not preserved in predicted projection-pose space. To mitigate this, CryoGeo enforces latent pose equivariance, using in-plane symmetries as frame-consistent geometric constraints on the predicted pose manifold. Experiments show that this output-level constraint improves reconstruction quality and pose accuracy on ambiguous synthetic and real-world datasets. It also preserves the speed of fully amortized inference by avoiding costly per-particle refinement. These findings suggest that stabilizing amortized inverse problems benefits from enforcing known observational symmetries on latent physical variables.

Limitations. CryoGeo targets homogeneous reconstruction without explicitly modeling structural heterogeneity. The latent equivariance identity is exact for continuous pure projection; the practical pipeline enforces an equivariance-consistent approximation with CTF modulation, finite-grid interpolation, sub-pixel shifts, and noise. Despite improved stability under pose ambiguity, CryoGeo cannot recover viewing directions absent under severe preferred orientation. Future work could extend the latent equivariance constraint to continuous heterogeneous modeling to better disentangle viewing direction ambiguity from molecular flexibility. It could also combine equivariant geometric priors with regularization to recover density from limited or missing viewing angles.

Acknowledgments and Disclosure of Funding

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62372091 and in part by the Hainan Province Key R&D Program under Grant No. ZDYF2024(LALH)001. The authors declare no competing interests.

References

[1] Werner Kühlbrandt. The resolution revolution. *Science*, 343(6178):1443–1444, 2014.

[2] Eva Nogales. The development of cryo-EM into a mainstream structural biology technique. *Nature Methods*, 13(1):24–27, 2016.

[3] Jean-Paul Renaud, Ashwin Chari, Claudio Ciferri, Wen-ti Liu, Hervé-William Rémigy, Holger Stark, and Christian Wiesmann. Cryo-EM in drug discovery: achievements, limitations and prospects. *Nature Reviews Drug Discovery*, 17(7):471–492, 2018.

[4] Joachim Frank. *Three-Dimensional Electron Microscopy of Macromolecular Assemblies: Visualization of Biological Molecules in Their Native State*. Oxford University Press, 2nd edition, 2006.

[5] Amit Singer and Fred J Sigworth. Computational methods for single-particle electron cryomicroscopy. *Annual Review of Biomedical Data Science*, 3(1):163–190, 2020.

[6] Tamir Bendory, Alberto Bartesaghi, and Amit Singer. Single-particle cryo-electron microscopy: Mathematical theory, computational challenges, and opportunities. *IEEE Signal Processing Magazine*, 37(2):58–76, 2020.

[7] Sjors HW Scheres. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *Journal of Structural Biology*, 180(3):519–530, 2012.

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

[9] 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.

[10] Timothy S Baker and R Holland Cheng. A model-based approach for determining orientations of biological macromolecules imaged by cryoelectron microscopy. *Journal of Structural Biology*, 116(1):120–130, 1996.

[11] Philip R Baldwin, Yong Zi Tan, Edward T Eng, William J Rice, Alex J Noble, Carl J Negro, Michael A Cianfrocco, Clinton S Potter, and Bridget Carragher. Big data in cryoEM: automated collection, processing and accessibility of EM data. *Current Opinion in Microbiology*, 43:1–8, 2018.

[12] Youssef SG Nashed, Frédéric Poitevin, Harshit Gupta, Geoffrey Woollard, Michael Kagan, Chun Hong Yoon, and Daniel Ratner. CryoPoseNet: End-to-end simultaneous learning of single-particle orientation and 3D map reconstruction from cryo-electron microscopy data. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 4049–4059. IEEE, 2021.

[13] Axel Levy, Frédéric Poitevin, Julien Martel, Youssef Nashed, Ariana Peck, Nina Miolane, Daniel Ratner, Mike Dunne, and Gordon Wetzstein. CryoAI: Amortized inference of poses for ab initio reconstruction of 3D molecular volumes from real cryo-EM images. In *European Conference on Computer Vision*, pages 540–557. Springer, 2022.

[14] Jiakai Zhang, Shouchen Zhou, Haizhao Dai, Xinhang Liu, Peihao Wang, Zhiwen Fan, Yuan Pei, and Jingyi Yu. CryoFastAR: Fast cryo-EM ab initio reconstruction made easy. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8462–8471. IEEE, 2025.

[15] Shayan Shekarforoush, David B Lindell, Marcus A Brubaker, and David J Fleet. CryoSPIN: Improving ab-initio cryo-EM reconstruction with semi-amortized pose inference. In *Advances in Neural Information Processing Systems*, volume 37, pages 55785–55809, 2024.

[16] Ellen D Zhong, Adam Lerer, Joseph H Davis, and Bonnie Berger. CryoDRGN2: Ab initio neural reconstruction of 3D protein structures from real cryo-EM images. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4046–4055. IEEE, 2021.

[17] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5738–5746. IEEE, 2019.

[18] Ronald N Bracewell. Discrete Hartley transform. *Journal of the Optical Society of America*, 73(12):1832–1835, 1983.

[19] Fred J Sigworth. A maximum-likelihood approach to single-particle image refinement. *Journal of Structural Biology*, 122(3):328–339, 1998.

[20] 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.

[21] Amit Singer, Ronald R Coifman, Fred J Sigworth, David W Chester, and Yoel Shkolnisky. Detecting consistent common lines in cryo-EM by voting. *Journal of Structural Biology*, 169(3):312–322, 2010.

[22] Ido Greenberg and Yoel Shkolnisky. Common lines modeling for reference free ab-initio reconstruction in cryo-EM. *Journal of Structural Biology*, 200(2):106–117, 2017.

[23] Gabi Pragier and Yoel Shkolnisky. A common lines approach for ab initio modeling of cyclically symmetric molecules. *Inverse Problems*, 35(12):124005, 2019.

[24] Nina Miolane, Frédéric Poitevin, Yee-Ting Li, and Susan Holmes. Estimation of orientation and camera parameters from cryo-electron microscopy images with variational autoencoders and generative adversarial networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4174–4183. IEEE, 2020.

[25] Harshit Gupta, Michael T McCann, Laurène Donati, and Michael Unser. CryoGAN: A new reconstruction paradigm for single-particle cryo-EM via deep adversarial learning. *IEEE Transactions on Computational Imaging*, 7:759–774, 2021.

[26] Dan Rosenbaum, Marta Garnelo, Michal Zielinski, Charlie Beattie, Ellen Clancy, Andrea Huber, Pushmeet Kohli, Andrew W Senior, John Jumper, Carl Doersch, et al. Inferring a continuous distribution of atom coordinates from cryo-EM images using VAEs. *arXiv preprint arXiv:2106.14108*, 2021.

[27] Claire Donnat, Axel Levy, Frédéric Poitevin, Ellen D Zhong, and Nina Miolane. Deep generative modeling for volume reconstruction in cryo-electron microscopy. *Journal of Structural Biology*, 214(4):107920, 2022.

[28] Ruyi Lian, Bingyao Huang, Liguo Wang, Qun Liu, Yuewei Lin, and Haibin Ling. End-to-end orientation estimation from 2D cryo-EM images. *Acta Crystallographica Section D: Structural Biology*, 78(2):174–186, 2022.

[29] 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.

[30] 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.

[31] Axel Levy, Gordon Wetzstein, Julien NP Martel, Frédéric Poitevin, and Ellen D Zhong. Amortized inference for heterogeneous reconstruction in cryo-EM. In *Advances in Neural Information Processing Systems*, volume 35, pages 13038–13049, 2022.

[32] Muyuan Chen and Steven J Ludtke. Deep learning-based mixed-dimensional Gaussian mixture model for characterizing variability in cryo-EM. *Nature Methods*, 18(8):930–936, 2021.

[33] Ali Punjani and David J Fleet. 3DFlex: determining structure and motion of flexible proteins from cryo-EM. *Nature Methods*, 20(6):860–870, 2023.

[34] Ellen D Zhong, Adam Lerer, Joseph H Davis, and Bonnie Berger. Exploring generative atomic models in cryo-EM reconstruction. *arXiv preprint arXiv:2107.01331*, 2021.

[35] Abner Guzman-Rivera, Dhruv Batra, and Pushmeet Kohli. Multiple choice learning: Learning to produce multiple structured outputs. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

[36] Christian Rupprecht, Iro Laina, Robert DiPietro, Maximilian Baust, Federico Tombari, Nassir Navab, and Gregory D Hager. Learning in an uncertain world: Representing ambiguity through multiple hypotheses. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3611–3620. IEEE, 2017.

[37] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

[38] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International Conference on Machine Learning*, pages 2990–2999. PMLR, 2016.

[39] Maurice Weiler and Gabriele Cesa. General E(2)-equivariant steerable CNNs. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[40] Gabriele Cesa, Kumar Pratik, and Arash Behboodi. Equivariant self-supervised deep pose estimation for Cryo EM. In *Proceedings of 2nd Annual Workshop on Topology, Algebra, and Geometry in Machine Learning (TAG-ML)*, pages 21–36. PMLR, 2023.

[41] Larissa de Ruijter and Gabriele Cesa. Equivariant amortized inference of poses for cryo-EM. *arXiv preprint arXiv:2406.01630*, 2024.

[42] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[43] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3D-structure-aware neural scene representations. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[44] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Advances in Neural Information Processing Systems*, volume 33, pages 7462–7473, 2020.

[45] Karen Ullrich, Rianne van den Berg, Marcus Brubaker, David Fleet, and Max Welling. Differentiable probabilistic models of scientific imaging with the Fourier slice theorem. *arXiv preprint arXiv:1906.07582*, 2019.

[46] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning*, pages 5301–5310. PMLR, 2019.

[47] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *Advances in Neural Information Processing Systems*, volume 33, pages 7537–7547, 2020.

[48] Chiheb Trabelsi, Olexa Bilaniuk, Ying Zhang, Dmitriy Serdyuk, Sandeep Subramanian, João Felipe Santos, Soroush Mehri, Negar Rostamzadeh, Yoshua Bengio, and Christopher J Pal. Deep complex networks. *arXiv preprint arXiv:1705.09792*, 2017.

[49] Miloš Vulović, Raimond BG Ravelli, Lucas J van Vliet, Abraham J Koster, Ivan Lazić, Uwe Lücken, Hans Rullgård, Ozan Öktem, and Bernd Rieger. Image formation modeling in cryo-electron microscopy. *Journal of Structural Biology*, 183(1):19–32, 2013.

[50] Ronald N Bracewell. Strip integration in radio astronomy. *Australian Journal of Physics*, 9(2):198–217, 1956.

[51] Alexandra C Walls, Young-Jun Park, M Alejandra Tortorici, Abigail Wall, Andrew T McGuire, and David Veesler. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell*, 181(2):281–292.e6, 2020.

[52] Clemens Plaschka, Pei-Chun Lin, and Kiyoshi Nagai. Structure of a pre-catalytic spliceosome. *Nature*, 546(7660):617–621, 2017.

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

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

[55] David Goodsell. Molecule of the Month: Hsp90. RCSB PDB-101, 2008.

[56] 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.

[57] Wilson Wong, Xiao-chen Bai, Alan Brown, Israel S Fernandez, Eric Hanssen, Melanie Condron, Yan Hong Tan, Jake Baum, and Sjors HW Scheres. Cryo-EM structure of the Plasmodium falciparum 80S ribosome bound to the anti-protozoan drug emetine. *eLife*, 3:e03080, 2014.

[58] Melinda M Diver, Yifan Cheng, and David Julius. Structural insights into TRPM8 inhibition and desensitization. *Science*, 365(6460):1434–1440, 2019.

[59] Guang Tang, Liwei Peng, Philip R Baldwin, Deepinder S Mann, Wen Jiang, Ian Rees, and Steven J Ludtke. EMAN2: an extensible image processing suite for electron microscopy. *Journal of Structural Biology*, 157(1):38–46, 2007.

[60] Thomas D Goddard, Conrad C Huang, Elaine C Meng, Eric F Pettersen, Gregory S Couch, John H Morris, and Thomas E Ferrin. UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Science*, 27(1):14–25, 2018.

[61] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[62] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[63] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

A Dataset preparation and preprocessing

Figure 7 provides representative particle images from the synthetic and experimental datasets used in our evaluation.

It illustrates the visual gap between clean simulated projections, controlled low-SNR synthetic projections, and real cryo-EM particle images with acquisition-specific noise and background artifacts.

A.1 Synthetic datasets

We generate synthetic datasets from atomic models obtained from the Protein Data Bank [53]. We use the SARS-CoV-2 Spike protein [51] and the Spliceosome complex [52] as two complementary targets. Spike contains strong pseudo-symmetric ambiguity, while Spliceosome provides a highly asymmetric structure. For each target, we compute a ground-truth density volume at a spatial size of 128^3 and generate 50,000 projection images with image size 128×128 by sampling random viewing orientations. Each projection is modulated by a particle-specific contrast transfer function [54] in the frequency domain. We construct two versions of each synthetic dataset: a clean version without additive noise and a noisy version with additive Gaussian noise at $\mathrm{SNR} = 0.1$.

A.2 Experimental datasets

We evaluate real-data reconstruction on EMPIAR-10028 and EMPIAR-11233 from the EMPIAR public archive [56]. EMPIAR-10028 contains 105,247 80S ribosome particle images [57] originally extracted at a box size of 360 with pixel size 1.34 Å. We downsample these particles to 128×128, corresponding to a pixel size of 3.77 Å. EMPIAR-11233 contains 42,040 TRPM8 ion-channel particles [58] originally extracted at a box size of 320 with pixel size 1.06 Å. We downsample these particles to 128×128, corresponding to a pixel size of 2.65 Å. The experimental particles contain non-negligible in-plane shifts, so CryoGeo predicts a 2D shift correction jointly with the orientation.

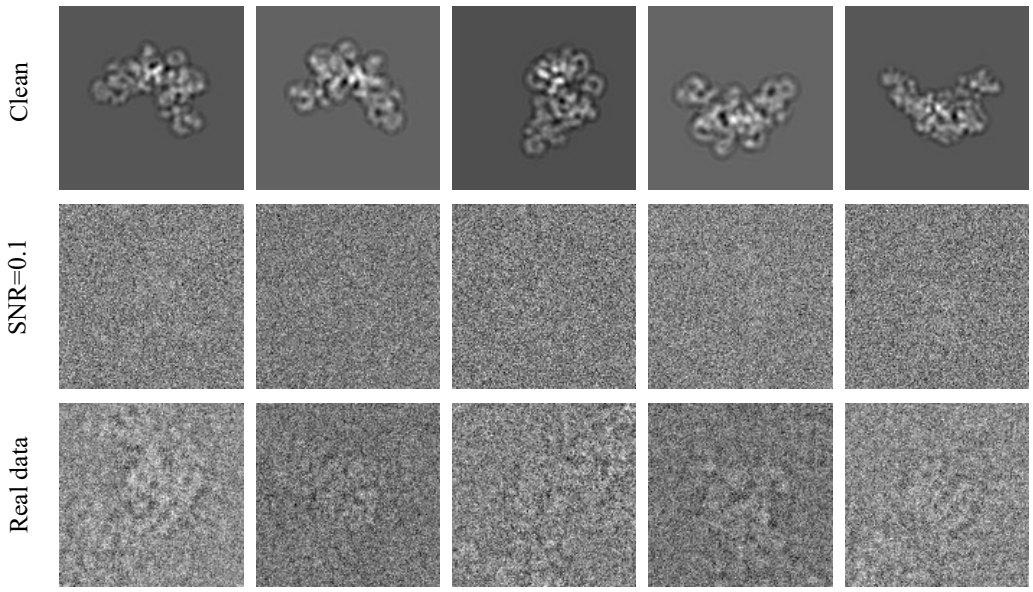


Figure 7: Representative particle images from the evaluated synthetic and experimental datasets. The figure shows clean simulated projections, noisy simulated projections at $\mathrm{SNR} = 0.1$, and experimental particle images from the real EMPIAR datasets. The visualization illustrates the difference between idealized synthetic inputs, controlled low-SNR simulations, and real cryo-EM particles with background noise and acquisition-specific artifacts.

A.3 FSC evaluation protocol

For synthetic datasets, we evaluate reconstruction quality using map-to-map FSC against the known ground-truth density and report the resolution at the FSC $= 0.5$ threshold. For experimental datasets, we randomly split the particles into two independent halves, reconstruct two half maps, and compute gold-standard FSC at the FSC $= 0.143$ threshold utilizing the EMAN2 software suite [59]. All methods are evaluated at the same image resolution and under the same preprocessing settings.

The Fourier Shell Correlation (FSC) between two 3D maps is computed over spherical frequency shells. For two Fourier-transformed volumes $F_1(\mathbf{k})$ and $F_2(\mathbf{k})$, the FSC at spatial frequency shell q is defined as

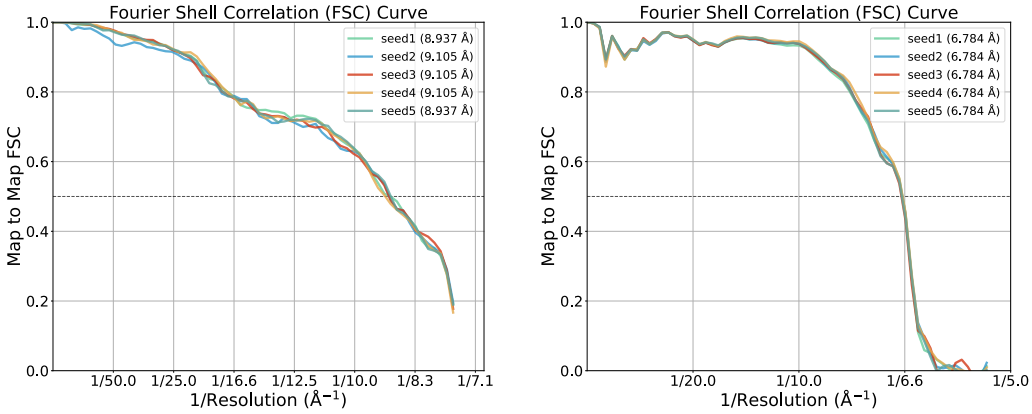
$$\mathrm{FSC}(q) = \frac{\sum_{\mathbf{k}\in\Omega_q} F_1(\mathbf{k})F_2(\mathbf{k})^*}{\sqrt{\sum_{\mathbf{k}\in\Omega_q} |F_1(\mathbf{k})|^2 \sum_{\mathbf{k}\in\Omega_q} |F_2(\mathbf{k})|^2}},$$

where Ω_q denotes the set of Fourier coefficients within the shell centered at frequency q, and $(\cdot)^*$ denotes complex conjugation. For synthetic datasets, F_1 and F_2 correspond to the reconstructed map and the ground-truth density. For experimental datasets, F_1 and F_2 correspond to the independently reconstructed half maps. The reported resolution is the reciprocal spatial frequency at which the FSC curve first falls below the corresponding threshold. If the FSC remains above the threshold up to the Nyquist frequency, we report the Nyquist-limited resolution.

A.4 Random-seed stability

To assess sensitivity to random initialization, we train CryoGeo five times on each experimental dataset, EMPIAR-10028 [57, 56] and EMPIAR-11233 [58, 56], using different random seeds. Since ground-truth poses are unavailable for real cryo-EM data, we evaluate stability by computing map-to-reference FSC between each reconstructed full map and the corresponding deposited reference map: EMD-2660 for EMPIAR-10028 and EMD-0639 for EMPIAR-11233.

For fair comparison, both reconstructed maps and reference maps are resampled to a 128^3 grid using UCSF ChimeraX [60] and aligned to a common coordinate frame before evaluation. This



EMPIAR-10028 / EMD-2660 EMPIAR-11233 / EMD-0639

Figure 8: Fixed-budget random-seed stability on EMPIAR-10028 [57, 56] and EMPIAR-11233 [58, 56]. CryoGeo is trained with five random initializations per dataset. Full maps are resampled to 128^3, aligned to the corresponding deposited references, and evaluated by map-to-reference FSC. The downsampled Nyquist limits are 7.54 Å and 5.30 Å, respectively. The highly consistent curves indicate that the results are not driven by a single favorable initialization.

Table 5: Random-seed stability on experimental cryo-EM datasets measured by map-to-reference FSC against deposited reference maps. CryoGeo is trained with five random initializations per dataset. All maps are resampled to a 128^3 grid and aligned before evaluation. Resolution is reported at FSC $= 0.5$; lower is better.

Run	EMPIAR-10028 / EMD-2660 (Å) ↓	EMPIAR-11233 / EMD-0639 (Å) ↓
Seed 1	8.937 (H^-)	6.784 (H^-)
Seed 2	9.105 (H^-)	6.784 (H^-)
Seed 3	9.105 (H^+)	6.784 (H^-)
Seed 4	9.105 (H^+)	6.784 (H^-)
Seed 5	8.937 (H^-)	6.784 (H^+)
Mean $\pm$ Std.	9.038 $\pm$ 0.092	6.784 $\pm$ 0.000

analysis complements the gold-standard half-map FSC reported in Table 1, which measures internal consistency between independently reconstructed half maps.

Figure 8 shows the full map-to-reference FSC curves across all five seeds. The curves remain closely aligned on both experimental datasets, suggesting limited sensitivity to random initialization under the fixed 100-epoch training budget. Table 5 reports the corresponding resolutions at the FSC $= 0.5$ threshold, together with the mean and standard deviation across five runs. For EMPIAR-11233, all five FSC curves cross the 0.5 threshold at the same discrete frequency shell.

In Table 5, H^+ denotes the same handedness as the deposited reference after alignment, while H^- denotes its chirality-inverted counterpart. Before FSC evaluation, H^- maps are chirality-corrected following Section D.2. All reconstructed volumes from the five random-seed runs are included in the supplementary zip file for visual inspection.

B Architecture and implementation details

B.1 Projection-pose predictor

The projection-pose predictor maps each input particle image to a view vector $u_i \in \mathbb{R}^3$, an orientation vector $v_i \in \mathbb{R}^3$, and an in-plane shift $t_i \in \mathbb{R}^2$. The vectors u_i and v_i are lifted to a valid rotation matrix through the differentiable Gram-Schmidt layer [17] described in the main paper. The shift

head predicts a bounded 2D translation, implemented using a scaled hyperbolic tangent output to avoid unstable translations during early training. The pose predictor can be configured with multiple heads to represent alternative pose hypotheses for each particle.

We instantiate the encoder with a VGG-style [61] steerable convolutional backbone [39] as an image-level inductive bias for in-plane transformations. The input image is represented as a trivial representation, while intermediate feature maps use regular representations of a discrete rotation group C_N. The backbone consists of double and triple convolutional blocks with kernel size 3 and padding 1. The base channel progression is $64, 128, 256, 256, 256$, scaled according to the group size. Spatial downsampling is performed using equivariant pooling layers with pooling kernel size 2.

The steerable encoder provides a feature-level inductive bias, while the group-action objective couples transformed observations to the pose queries used by the renderer.

B.2 Hartley-domain volume renderer

CryoGeo represents the 3D density using a trainable voxel grid in the Hartley domain [18]. For an image size of $N \times N$, we use a frequency grid of size $(N+1)^3$ for central-slice extraction. Given a query orientation R, the renderer rotates a canonical 2D frequency plane and samples the corresponding central slice from the 3D Hartley volume using differentiable grid sampling.

The extracted Hartley slice is converted into the frequency-domain projection used for CTF modulation. The particle-specific CTF is applied before comparing the rendered projection with the observed image. Sub-pixel shifts are implemented through the corresponding Hartley-domain modulation, avoiding direct resampling of the spatial-domain image during shift correction. This real-valued representation avoids explicit complex-valued Fourier arithmetic while remaining differentiable with respect to both the volume grid and the predicted pose.

B.3 CTF handling under in-plane rotations

CryoGeo applies each particle's CTF and predicted in-plane shift in the coordinate frame of the original particle image. Let $\mathcal{D}_h^0(V, R_i)$ denote the discrete projection renderer before CTF modulation and shift application, where h denotes the finite sampling grid. Let C_i be the particle-specific CTF operator determined by the acquisition metadata, and let T_{t_i} denote the translation operator associated with the predicted shift t_i. The original image prediction is

$$\widehat{I}_i = T_{t_i} C_i \, \mathcal{D}_h^0(V, R_i).$$

For an in-plane rotation θ, the query orientation is $R_i R_z(\theta)$. Writing $\rho_{\theta,h}$ for the digital image rotation, the transformed prediction is

$$\widehat{I}_{i,\theta} = \rho_{\theta,h} T_{t_i} C_i \rho_{-\theta,h} \mathcal{D}_h^0\big(V, R_i R_z(\theta)\big).$$

Specifically, we first rotate the query projection back to the original detector frame, apply the original particle-specific CTF and predicted shift in that frame, and then rotate the result forward. The resulting prediction is compared with the transformed observation $\rho_{\theta,h} I_i$.

At the continuous-operator level, the corresponding transformations are

$$C_i^{(\theta)} = \rho_\theta C_i \rho_{-\theta}, \qquad T_{t_i}^{(\theta)} = \rho_\theta T_{t_i} \rho_{-\theta}.$$

An isotropic CTF remains unchanged under this conjugation, whereas an anisotropic CTF has its astigmatic axes rotated together with the observation. The original-frame implementation therefore handles both cases through the same operator sequence.

Together with the continuous projection identity $\Pi(V, R_i R_z(\theta)) = \rho_\theta \Pi(V, R_i)$, these transformations give

$$\rho_\theta T_{t_i} C_i \rho_{-\theta} \Pi\big(V, R_i R_z(\theta)\big) = \rho_\theta T_{t_i} C_i \Pi(V, R_i).$$

The complete image formation model thus transforms consistently with the observation. On the finite grid, interpolation and resampling make this covariance approximate.

Table 6: Residual statistics for digital rotation and the complete renderer. The digital-rotation row reports $\varepsilon^{\mathrm{rot}}$, while the renderer rows report $\varepsilon^{\mathrm{cov}}$. SD denotes the sample standard deviation, and p95 denotes the 95th percentile. The mean, SD, median, and p95 are expressed as percentages; the final column reports the maximum relative residual at $\theta = 0$ as a dimensionless ratio.

Setting	N	Mean $\pm$ SD (%)	Median (%)	p95 (%)	Max at $\theta = 0$
Digital rotation only	1000	5.33 ± 0.74	5.25	6.62	1.33×10^{-6}
Complete renderer, isotropic CTF	1000	6.19 ± 1.32	6.35	8.04	1.10×10^{-6}
Complete renderer, anisotropic CTF	1000	6.19 ± 1.32	6.35	8.03	1.12×10^{-6}

Numerical covariance evaluation. We quantify the covariance residual of the implemented renderer using a fixed ground-truth volume and known poses, without training. For each sampled pose-angle combination, we compute

$$\varepsilon_{i,\theta}^{\mathrm{cov}} = \frac{\left\| \widehat{I}_{i,\theta} - \rho_{\theta,h} \widehat{I}_i \right\|_2}{\left\| \rho_{\theta,h} \widehat{I}_i \right\|_2},$$

where $\widehat{I}_i$ and $\widehat{I}_{i,\theta}$ are the predictions defined above. We evaluate isotropic and anisotropic particle-specific CTFs separately over $N = 1000$ pose-angle combinations.

We also measure the round-trip residual of digital rotation:

$$\varepsilon_{i,\theta}^{\mathrm{rot}} = \frac{\left\| \rho_{-\theta,h} \rho_{\theta,h} x_i - x_i \right\|_2}{\left\| x_i \right\|_2},$$

where x_i denotes an image used in the digital-rotation control.

Table 6 summarizes the measurements. Digital rotation alone yields a mean round-trip residual of 5.33%. The complete renderer yields similar covariance residuals under isotropic and anisotropic CTFs, with a mean of 6.19% in both settings and 95th percentiles below 8.1%. At $\theta = 0$, the maximum residual is on the order of 10^{-6} for all three settings.

B.4 Training protocol

We implement CryoGeo in PyTorch [62] and optimize the pose predictor and Hartley-domain volume jointly using AdamW [63]. Unless otherwise stated, we train for 200 epochs with a batch size of 32, using learning rates of 10^{-4} for the pose predictor and 5×10^{-3} for the volume grid. The learning rates are separated because the explicit volume grid benefits from faster optimization, while the pose predictor requires a smaller learning rate for stable pose estimation.

C Baseline configurations

We evaluate all baselines under matched preprocessing, image resolution, and FSC evaluation protocols. For each baseline, we follow the experimental settings described in the corresponding paper and the default hyperparameters provided in the official open-source implementation whenever available. All methods use the same downsampled particle images, the same CTF information when applicable, and the same synthetic or real-data splits as CryoGeo. For synthetic datasets, map-to-map FSC values are computed against the same ground-truth density at the FSC $= 0.5$ threshold. For experimental datasets, all methods are evaluated using the same half-map split and gold-standard FSC protocol at the FSC $= 0.143$ threshold.

For each baseline, we report the best stable checkpoint after convergence under the same evaluation protocol. For stochastic neural methods, we perform at least five independent runs when convergence varies noticeably across random initializations. For each run, we monitor the training loss and FSC-based reconstruction quality, and select the best stable checkpoint after the reconstruction quality reaches a plateau. When the default training schedule does not reach a stable plateau, we extend the training duration while keeping the remaining hyperparameters unchanged. For neural baselines, we train for at least 500 epochs or until both the training loss and reconstruction quality stabilize.

CryoAI. We run CryoAI [13] following the settings described in the original paper and the default hyperparameters in the official implementation. The input image resolution, CTF information, and preprocessing are matched to those used by CryoGeo. When the default training schedule does not converge to a stable reconstruction, we extend the training duration while keeping the remaining hyperparameters unchanged. We repeat training with at least five random initializations when convergence varies across runs and report the best converged checkpoint according to the same FSC-based evaluation protocol.

CryoDRGN2. We run CryoDRGN2 [16] following the standard *ab initio* pose-search reconstruction pipeline described in the original paper and official implementation. All inputs are generated from the same preprocessed particle stacks used by the other methods. We use the official default settings unless a longer training schedule is required for convergence. For synthetic datasets, the final reconstructed map is aligned to the ground-truth density before map-to-map FSC evaluation. For experimental datasets, the reconstructed half maps are evaluated using the same gold-standard FSC protocol as the other methods.

CryoSPIN. We run CryoSPIN [15] following the semi-amortized pipeline described in the original paper, consisting of an amortized pose-prediction stage followed by an unamortized pose-refinement stage. The official implementation of CryoSPIN does not include in-plane translation prediction. Direct application to our real datasets yielded limited reconstruction quality. We therefore implemented the in-plane translation predictor described in the CryoSPIN paper. The reported real-data results use this updated implementation, with shifts predicted directly from particle images. Because CryoSPIN uses a multi-hypothesis pose representation, we additionally evaluate variants with 24 and 32 pose heads to account for sensitivity to the number of hypotheses. We train the amortized stage for 500 epochs and then run 100 epochs of unamortized pose refinement. For each dataset, we report the best converged checkpoint after the full semi-amortized refinement procedure. The runtime reported in Table 4 separates the amortized stage and the subsequent unamortized refinement stage.

CryoSPARC. We run CryoSPARC [8] using its standard *ab initio* reconstruction workflow with default settings unless otherwise specified. For synthetic datasets, the final reconstructed map is aligned to the ground-truth density before map-to-map FSC evaluation. For experimental datasets, particles are split into two independent halves before reconstruction, and the resulting half maps are evaluated using the same gold-standard FSC protocol as all other methods. Since CryoSPARC is an iterative optimization pipeline rather than an epoch-based neural method, we use its standard convergence criteria and inspect the resulting reconstructions before FSC evaluation.

C.1 Baseline selection rationale

We selected baselines to cover four representative reconstruction paradigms: iterative optimization, neural pose-search reconstruction, fully amortized inference, and semi-amortized pose refinement. CryoSPARC [8] represents a widely used iterative *ab initio* reconstruction pipeline, CryoDRGN2 [16] represents neural reconstruction with explicit pose search, CryoAI [13] represents fully amortized pose inference, and CryoSPIN [15] represents semi-amortized inference with subsequent local pose refinement.

CryoFastAR [14] was not included because reproducing its full training pipeline under our setting would require additional implementation and configuration details that were not available to us at the time of our experiments.

We did not include recent equivariant pose-inference methods [40, 41] as direct reconstruction baselines because they are closely related in motivation but differ from CryoGeo in scope and in where equivariance is imposed. Cesa et al. [40] train an equivariant graph-based pose estimator using a common-line self-supervised loss without performing intermediate 3D reconstruction. De Ruijter and Cesa [41] incorporate equivariant encoders into the cryoAI pipeline to improve convergence and pose consistency in amortized inference. In contrast, CryoGeo enforces equivariance at the output projection-pose level through a renderer-consistent group-action objective, coupling the transformed latent pose, frame-consistent shift correction, and transformed observation within the same amortized *ab initio* reconstruction framework.

Direct evaluation under our matched FSC protocols would require additional implementation choices. To our knowledge, complete official code and configuration files for reproducing their full training and reconstruction procedures under our setting were not publicly released at the time of our experiments. Instead, we isolate the relevant design question in Table 2: the "Equivariant Enc. only" variant provides a controlled within-framework comparison to the pose-estimator/encoder-level equivariance direction explored in these works, evaluated under the same reconstruction pipeline, encoder backbone, and training protocol as CryoGeo. This isolates the contribution of the latent projection-pose equivariance objective from confounding implementation differences.

We evaluated the standalone `v0.3.3-beta` release of CryoDRGN-AI [29] (package version `0.3.3b0`) in preliminary experiments using our 128×128 reconstruction setting. Under this configuration, we did not obtain consistent convergence to the correct global structure. We report CryoDRGN2 [16] as the representative neural pose-search baseline in our comparisons.

C.2 Feature-level equivariance versus latent projection-pose equivariance

Let $G = SO(2)$ denote the in-plane image rotation group acting on images by ρ_θ. A feature encoder E is feature-level equivariant if there exists a feature-space representation Γ_θ such that

$$E(\rho_\theta I) = \Gamma_\theta E(I).$$

This property constrains intermediate features, but it does not by itself determine how the final predicted cryo-EM projection pose transforms.

Let the pose predictor be decomposed as

$$f_\phi = h \circ E,$$

where h maps features to a projection pose (R, t). CryoGeo instead targets output-level projection-pose equivariance:

$$R(\rho_\theta I) \approx R(I) R_z(\theta),$$

together with the frame-consistent shift action

$$T_t^{(\theta)} = \rho_\theta T_t \rho_{-\theta}.$$

CryoGeo enforces this physical transformation law at the renderer-query level by analytically coupling transformed queries to the same predicted base pose.

Feature-level equivariance does not imply this output-level property in general. From $E(\rho_\theta I) = \Gamma_\theta E(I)$, we only obtain

$$f_\phi(\rho_\theta I) = h(E(\rho_\theta I)) = h(\Gamma_\theta E(I)).$$

For this expression to match the desired physical pose action for all feature vectors z and all in-plane rotations θ, the readout head h would additionally need to satisfy

$$h(\Gamma_\theta z) = h(z) \circ g_\theta.$$

Here g_θ denotes the physical projection-pose action:

$$(R, t) \circ g_\theta = (R R_z(\theta), Q_\theta t),$$

where $Q_\theta \in SO(2)$ is the in-plane rotation matrix satisfying $T_{Q_\theta t} = \rho_\theta T_t \rho_{-\theta}$. This additional condition is not guaranteed by feature-level equivariance alone. An arbitrary readout head can destroy or distort the group action even when the encoder E is exactly equivariant.

A simple counterexample illustrates the gap. For every feature vector z, let the pose head be constant:

$$h(z) = (R_0, t_0).$$

Then

$$f_\phi(\rho_\theta I) = h(\Gamma_\theta E(I)) = (R_0, t_0).$$

However, output-level latent pose equivariance would require

$$f_\phi(I) \circ g_\theta = (R_0 R_z(\theta), Q_\theta t_0).$$

For a generic rotation angle θ, these two poses are not equal. Thus, even a perfectly equivariant feature encoder does not guarantee that the final predicted projection pose follows the physical cryo-EM transformation law.

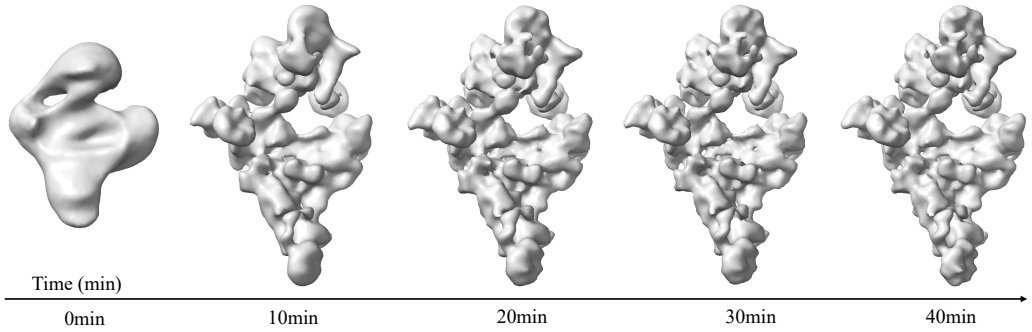


Figure 9: Qualitative evolution of the reconstructed 3D volume during training. The figure shows intermediate reconstructions at different training times, illustrating that CryoGeo recovers the global molecular shape early in optimization and gradually refines local density details in later stages. This visualization complements the runtime analysis by showing the progression of the amortized reconstruction process.

CryoGeo couples transformed observations to renderer queries generated from the same predicted base pose through the group-action objective in Section 3.4. For each query, the particle-specific CTF and predicted shift are applied in the original detector frame following the operator sequence in Appendix B.3. This objective couples the image-space transformation to the physical action on the projection pose used for rendering.

Recent equivariant pose-inference methods [40, 41] are closely related in motivation, but the distinction above explains why feature-level or pose-estimation equivariance does not automatically imply the renderer-consistent latent projection-pose equivariance enforced by CryoGeo.

C.3 Relation to reconstruction and rotation augmentation

Analytic coupling of pose queries. Given an input image I_i, let $P_i = f_\phi(I_i) = (R_i, t_i)$ denote its predicted pose. The prescribed pose action is

$$A_\theta(P_i) = \big(R_i R_z(\theta), Q_\theta t_i\big),$$

where Q_θ is the in-plane rotation satisfying $T_{Q_\theta t_i} = \rho_\theta T_{t_i} \rho_{-\theta}$. Rotation augmentation obtains a pose from the transformed image, while CryoGeo constructs the transformed query from the base prediction:

$$P^{\mathrm{aug}}_{i,\theta} = f_\phi(\rho_{\theta,h} I_i), \qquad P^{\mathrm{geo}}_{i,\theta} = A_\theta(P_i).$$

CryoGeo therefore parameterizes all transformed renderer queries through the same base pose. The equality $P^{\mathrm{geo}}_{i,\theta} = A_\theta(P_i)$ is enforced directly at the renderer-query level. The pose gradients of these queries are pulled back through $DA_\theta(P_i)^*$ to update the shared base prediction, where DA_θ denotes the differential of the pose action and * denotes its adjoint.

Relation to reconstruction. We use the original and transformed image predictions defined in Appendix B.3, including the specified CTF and shift operations. Suppressing the particle index, define the reconstruction residual and the finite-renderer covariance discrepancy as

$$r = \widehat{I} - I, \qquad \Delta_{\theta,h} = \widehat{I}_\theta - \rho_{\theta,h}\widehat{I}.$$

The transformed residual then satisfies

$$r_{\mathrm{geo},\theta} = \widehat{I}_\theta - \rho_{\theta,h} I = \rho_{\theta,h} r + \Delta_{\theta,h}.$$

Let $\ell_{\mathrm{rec}} = \|r\|_2^2$ and $\ell_{\mathrm{geo},\theta} = \|r_{\mathrm{geo},\theta}\|_2^2$. Averaging over images and sampled angles gives the corresponding objectives $\mathcal{L}_{\mathrm{rec}}$ and $\mathcal{L}_{\mathrm{geo}}$. With $B_{\theta,h} = \rho^*_{\theta,h}\rho_{\theta,h} - \mathrm{Id}$, an exact expansion gives

$$\ell_{\mathrm{geo},\theta} - \ell_{\mathrm{rec}} = \langle r, B_{\theta,h} r\rangle + 2\mathrm{Re}\langle \rho_{\theta,h} r, \Delta_{\theta,h}\rangle + \|\Delta_{\theta,h}\|_2^2,$$

where Id is the identity operator.

Under exact continuous covariance and norm-preserving rotation, the discrepancy and non-isometry terms vanish, giving $\ell_{\mathrm{geo},\theta} = \ell_{\mathrm{rec}}$. The corresponding ideal objectives and their gradients coincide. For the finite implementation, $B_{\theta,h}$ captures departures of digital rotation from isometry, while $\Delta_{\theta,h}$ captures the renderer covariance discrepancy. These terms define the correction to the reconstruction objective, and their derivatives determine the corresponding gradient correction. The analytic query construction couples these corrections to the same base pose. Table 6 provides numerical measurements of the normalized covariance discrepancy under isotropic and anisotropic CTFs.

Connection to pose collapse. A collapsed predictor can assign the same pose P_c to an image and its rotated versions. Rotation augmentation can then repeatedly query P_c, whereas CryoGeo constructs the queries $A_\theta(P_c)$. The gradients from these in-plane-rotated queries are coupled through the shared base pose.

Let w collect the predictor and volume parameters, and consider a smooth local manifold $\mathcal{C}$ of collapsed configurations in parameter space. Define the gradient correction

$$\delta g(w) = \nabla_w \mathcal{L}_{\mathrm{geo}}(w) - \nabla_w \mathcal{L}_{\mathrm{rec}}(w).$$

At a reconstruction stationary point $w_c \in \mathcal{C}$, we have $\nabla_w \mathcal{L}_{\mathrm{rec}}(w_c) = 0$. If

$$\Pi_{N_{w_c}\mathcal{C}}\, \delta g(w_c) \neq 0,$$

where $\Pi_{N_{w_c}\mathcal{C}}$ is the orthogonal projection onto the normal space of $\mathcal{C}$ at w_c, then a gradient-descent step $w^+ = w_c - \eta\, \delta g(w_c)$ leaves $\mathcal{C}$ to first order in the step size η. This gives a local mechanism by which the coupled gradient correction moves optimization away from a collapsed configuration. The same condition applies when $\mathcal{C}$ represents viewing-direction collapse.

Local curvature also determines sensitivity to this correction. After fixing the global pose-volume gauge, a smooth objective perturbation around an isolated stationary point with a nonsingular Hessian produces a first-order displacement governed by the inverse Hessian. Corrections along shallow-curvature directions can therefore produce substantial changes in the stationary pose configuration.

D Additional evaluation details

D.1 Qualitative training evolution

Figure 9 shows how the reconstructed volume evolves during training, from early recovery of the global molecular shape to later refinement of local density details. This qualitative trajectory complements the quantitative runtime analysis in Table 4.

D.2 Global alignment and chirality correction

For synthetic datasets, ground-truth particle orientations are available, but an *ab initio* reconstruction is only identifiable up to a global coordinate transformation and possible handedness ambiguity. Before computing pose errors, we therefore align the predicted density map to the ground-truth density and correct the handedness when necessary. Figure 10 illustrates this chirality correction and global coordinate alignment pipeline.

We align the predicted density map to the ground-truth density using volumetric fitting in UCSF ChimeraX [60]. When chirality correction is required, we first reflect the predicted density and then align the reflected map. Let $R_{\mathrm{align}} \in SO(3)$ denote the rotational component of this alignment, mapping coordinates of the predicted volume, after chirality correction when needed, into the ground-truth coordinate frame.

We use the column-vector convention, in which R_{pred} maps camera coordinates to predicted-volume coordinates. For chirality correction, we define

$$F_x = \mathrm{diag}(-1, 1, 1), \qquad F_z = \mathrm{diag}(1, 1, -1).$$

Here, F_x reflects the predicted volume in its coordinate frame, while F_z reverses the camera depth axis so that the aligned orientation remains a proper rotation. The aligned orientation is

$$R_{\mathrm{aligned}} = R_{\mathrm{align}} F_x R_{\mathrm{pred}} F_z.$$

Without chirality correction, it is

$$R_{\mathrm{aligned}} = R_{\mathrm{align}} R_{\mathrm{pred}}.$$

We compute pose errors using these aligned orientations in the ground-truth coordinate frame.

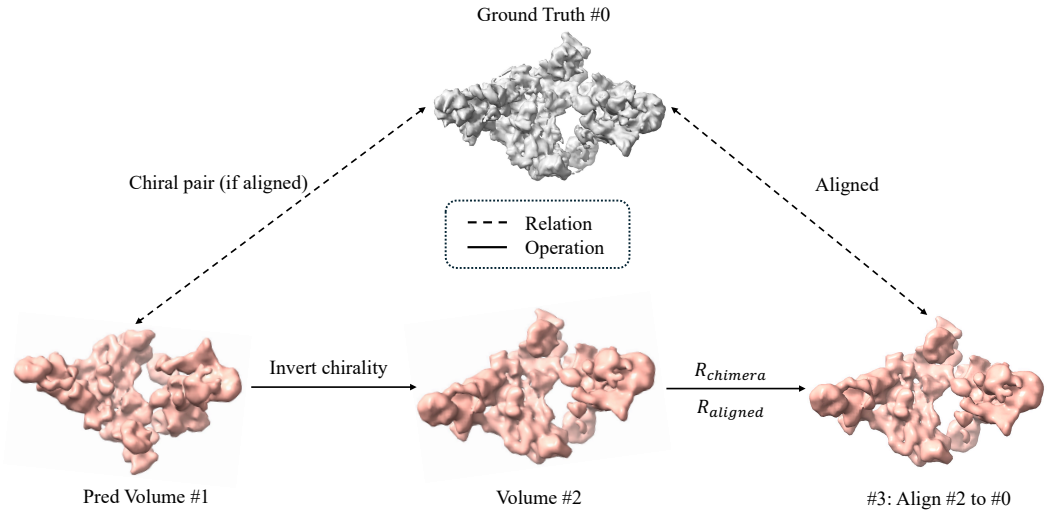


Figure 10: Illustration of the chirality correction and global coordinate alignment pipeline used for pose evaluation. Since *ab initio* cryo-EM reconstructions are identifiable only up to a global rotation and possible handedness ambiguity, the predicted density map is first aligned to the ground-truth reference frame before computing pose errors. This procedure ensures that the reported angular errors reflect per-particle pose accuracy rather than arbitrary global coordinate-frame differences.

D.3 View-direction error and high-precision accuracy

We evaluate pose accuracy using the view direction, defined as the third column of the rotation matrix. Let $u_{\mathrm{gt}}^{(i)}$ denote the ground-truth view direction and let $\hat{u}_{\mathrm{pred}}^{(i)}$ denote the view direction extracted from $R_{\mathrm{aligned}}^{(i)}$ for particle i. The view-direction error is

$$\theta_{\mathrm{view}}^{(i)} = \arccos\left(\left\langle u_{\mathrm{gt}}^{(i)}, \hat{u}_{\mathrm{pred}}^{(i)} \right\rangle\right) \times \frac{180}{\pi}.$$

We report four metrics over the dataset. The median angular error provides a robust measure of central tendency:

$$\mathrm{Med.} = \mathrm{median}\left(\left\{\theta_{\mathrm{view}}^{(i)}\right\}_{i=1}^{N}\right).$$

The root mean square error penalizes large angular deviations and is sensitive to collapse-prone outliers:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\theta_{\mathrm{view}}^{(i)}\right)^2}.$$

We also compute the cosine distance on the view sphere:

$$D_{\cos} = \frac{1}{N}\sum_{i=1}^{N}\left(1 - \left\langle u_{\mathrm{gt}}^{(i)}, \hat{u}_{\mathrm{pred}}^{(i)} \right\rangle\right).$$

Finally, high-precision accuracy measures the fraction of particles whose view-direction error falls below a threshold $\theta_{\mathrm{th}} \in \{5^\circ, 10^\circ\}$:

$$\mathrm{Acc@}\theta_{\mathrm{th}} = \frac{100}{N}\sum_{i=1}^{N}\mathbb{I}\left(\theta_{\mathrm{view}}^{(i)} \le \theta_{\mathrm{th}}\right).$$

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction describe CryoGeo within the evaluated scope of amortized *ab initio* cryo-EM reconstruction with latent pose equivariance. The claims are supported by the method in Section 3, the experiments in Section 4, the ablation and efficiency results, and the limitations discussed in Section 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5 discusses limitations, including the focus on homogeneous reconstruction, the approximate nature of the practical equivariance objective, and failure under severe missing-view or preferred-orientation settings. Computational efficiency is reported in Section 4.4 and Table 4.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The mathematical material in this paper serves to formulate and explain the method. Section 3 and Appendices B.3, C.2, and C.3 provide the pose transformation laws, covariance and objective derivations, and local analysis of the optimization behavior, with the corresponding assumptions stated explicitly.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper and supplementary material describe the datasets, preprocessing, FSC computation, model architecture, training protocol, baseline configurations, and evaluation metrics needed to reproduce the main experimental results. These details are provided in Sections 3 and 4 and Appendices A–D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The public EMPIAR and PDB datasets used in the experiments are cited and the data generation, preprocessing, FSC computation, training, and baseline settings are described in the supplementary material. The code is not released at submission time, but the implementation details in Appendices A–C provide a reproducibility path.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4 describes the datasets, baselines, evaluation settings, and reported metrics. Appendices A, B, and C provide preprocessing details, FSC computation, architecture details, optimizer settings, training schedule, and baseline configurations.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: The supplementary material reports fixed-budget five-seed random-initialization stability analyses for CryoGeo on two experimental datasets, EMPIAR-10028 and EMPIAR-11233. The analysis uses map-to-reference FSC against deposited reference maps and reports the mean and standard deviation across runs to quantify initialization-induced variability and assess global convergence stability under a matched 100-epoch training budget, while the main tables report matched evaluation results under the same preprocessing and FSC protocols.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 4.4 and Table 4 report runtime and peak GPU memory on a single NVIDIA RTX 4090. Appendix B provides the training protocol, including optimizer, learning rates, batch size, and number of epochs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses public cryo-EM datasets and synthetic projections for computational structural reconstruction and does not involve human subjects, private personal data, or deployed decision-making systems. We are not aware of ethical concerns beyond standard scientific data-use and reproducibility considerations.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses the potential scientific impact of more stable and efficient amortized cryo-EM reconstruction for structural biology in the abstract and conclusion. The work is primarily methodological and uses public scientific datasets; we do not identify direct negative societal impacts beyond standard risks associated with computational tools for biological structure determination.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a system intended for deployment or user-facing decision making, and it does not introduce new sensitive data. The experiments use public scientific datasets and synthetic projections.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper cites the original sources for public datasets, molecular structures, baseline methods, and software tools, including PDB, EMPIAR, EMAN2, PyTorch, CryoSPARC, CryoAI, CryoDRGN2, and CryoSPIN. We use these assets for academic evaluation under their intended research use and provide dataset and baseline details in Appendices A and C.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce a new public dataset or benchmark asset. Synthetic projections are generated for evaluation from public molecular structures, and the generation protocol is described in Appendix A.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

Justification: The paper does not involve crowdsourcing, user studies, or human-subject experiments.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve human subjects or private personal data, and therefore IRB approval is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method does not use LLMs as a methodological component. Any LLM use, if applicable, was limited to writing, editing, or formatting assistance and did not affect the scientific content or originality of the work.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.