CRYOFATHOM: FLOW-MATCHED AMORTIZED TOMOGRAPHY FOR HELD-OUT MOLECULES

Anonymous authors
Paper under double-blind review

ABSTRACT

Single-particle cryo-electron microscopy (cryo-EM) is a key technique for estimating the 3D density of proteins and other large molecules from randomly oriented 2D projection images. Standard pipelines use many thousands of noisy images to simultaneously estimate a high-resolution molecular density and the projection operator that corresponds to each image. Emboldened by the recent progress in generative AI, we ask if one could learn a strong prior on the distribution of protein shapes to produce useful reconstructions from a tiny number of input images. To answer this question we built CryoFATHOM, a conditional flow matching model that was trained on 850 proteins from the Protein Data Bank. The model samples a molecular density conditioned on its projection images, where for simplicity we assume known orientations. Reconstruction is done by integrating the ODE of the trained conditional velocity field, with no inference-time optimization. By picking the number of Euler steps and incorporating classifier-free guidance we can trade sharpness against fidelity. We release our code and a benchmark for evaluating trainable single-particle reconstruction methods on held-out molecules. In our experiments, CryoFATHOM produced better resolution reconstructions than RELION up to about 100 projection images. Potential applications of our approach include the reconstruction of molecules with continuous heterogeneity, cryo-ET, and medical imaging methods such as CT and MRI.

1 INTRODUCTION

Single-particle cryo-EM determines the 3D density of a molecule from 2D projection images of many copies of it, frozen in ice at unknown orientations. Established pipelines such as RELION and cryoSPARC (Scheres, 2012; Punjani et al., 2017) use many thousands of such images. They estimate a single volume and an orientation for every image at once, so that each image matches the volume's projection at that orientation. That scale is required for obtaining atomic resolution structures when the orientations must be inferred from the same noisy images (Henderson, 1995). We ask what can be recovered when the orientations are known from upstream processing and the images are few: tens rather than tens of thousands. In the regime of few projection images, the 3D reconstruction problem is ill-posed since many volumes agree with the projection images and some prior is needed to select among them. Classical pipelines for cryo-EM reconstruction incorporate a prior by including generic regularization terms (Scheres, 2012). Some modern methods fit a representation to the particle stack at hand (Zhong et al., 2020; 2021b), trained anew for every molecule, without using information from related structures. These methods perform point estimation and offer no fine-grained control over the reconstructed volume during inference time.

In order to address the challenge of 3D reconstruction from very few projection images, we wish to learn a rich prior from a training set of molecules that are from the same family as the target. To this end we have developed CryoFATHOM, a conditional flow matching model trained once across many molecules to sample a volume given a set of posed projections. Reconstructing a molecule it has never seen needs no retraining and no inference-time optimization. To test the generalization ability of our model we assembled a large dataset of proteins to use as a benchmark. As opposed to a standard feed-forward network, the flow matching framework allows control during inference time over the reconstructed volume. A feed-forward network with the same architecture would

Code and benchmark: submitted as an anonymized zip; a public repository upon publication.

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

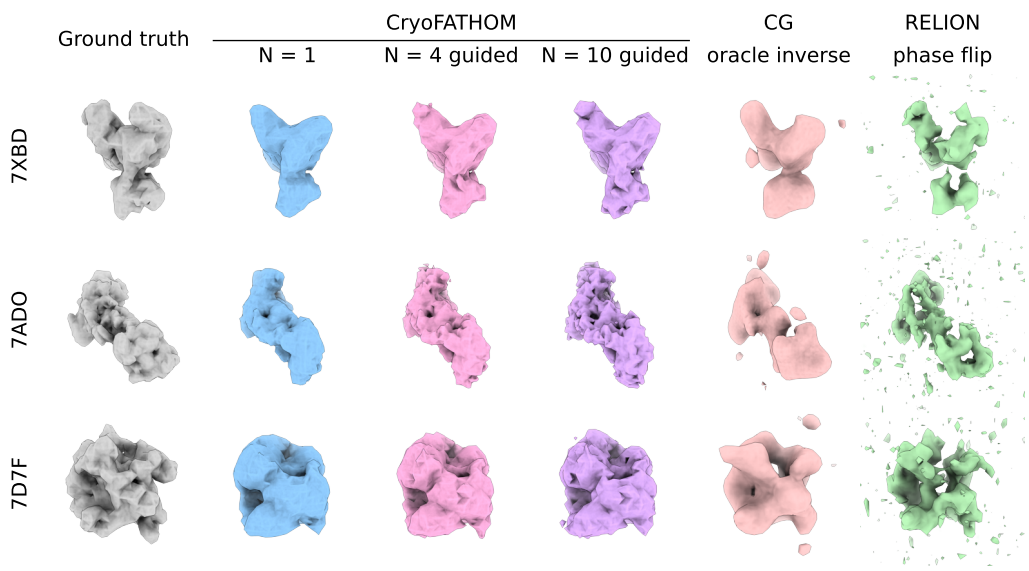


Figure 1: Reconstructions of three held-out test molecules from 16 projection images, shown as isosurface plots. Columns show the ground truth, three CryoFATHOM reconstructions with different inference-time settings, Conjugate Gradient and RELION. Appendix J provides more reconstructions with different projection counts and the technical details of how they are rendered.

transfer across molecules equally, but its optimum is the conditional mean, resulting in blurred reconstructions. In flow matching, the balance between sharpness and fidelity can be controlled during inference by choosing the number of Euler steps and incorporating classifier-free guidance (CFG) (Ho & Salimans, 2021), in a form that needs no change to training (Section 4.4). In fact, flow matching with a single Euler step returns the same conditional mean that feed-forward networks estimate (Appendix A). Figure 1 shows CryoFATHOM reconstructions on three held-out molecules using different inference time settings. The second column with $N = 1$ Euler steps corresponds to the blurry conditional mean whereas the $N = 4$ and $N = 10$ columns show reconstructions with more Euler steps and CFG.

Potential applications of our approach include single-particle reconstruction of flexible molecules with continuous heterogeneity (Moscovich et al., 2020; Gilles & Singer, 2025), where a volume is reconstructed from projection images in a small region of the latent space. Such reconstructions are weakest in sparsely populated states, and a specialized model for reconstruction from very few images could be helpful. Cryo-electron tomography (cryo-ET), which records a few tens of images at known tilt angles over a restricted range (Wiedemann & Heckel, 2024), is a further prospective application. The same problem arises outside cryo-EM. Fewer projections lower the radiation dose in computed tomography (Bian et al., 2010), and fewer measurements shorten the scan in magnetic resonance imaging (Lustig et al., 2007).

Contributions. We present CryoFATHOM, a conditional flow matching model that amortizes reconstruction across molecules. It includes a geometry-aware projection encoder that is permutation invariant and defined for any projection count. This encoder conditions a DiT velocity network. A molecular volume is reconstructed by posterior sampling in a single ODE solve, with solver steps and CFG as inference-time controls (Section 4). To train and test our model, we assembled a benchmark for held-out 3D molecular reconstruction. It contains one thousand proteins, deduplicated at 30% sequence identity to guarantee diversity and avoid potential overfitting. This dataset is split into training, validation and test sets (Section 5).

2 Related work

2.1 3D reconstruction in single-particle cryo-EM

Established methods estimate a single volume by iterative refinement against the projection images, regularized by a prior chosen in advance. RELION formulates this as maximum a posteriori estimation under a Gaussian prior on the Fourier coefficients (Scheres, 2012). CryoSPARC introduces stochastic gradient descent and branch-and-bound alignment to accelerate the same estimation problem (Punjani et al., 2017). Their optimization process is dataset-specific, the prior carries no knowledge of other molecules.

In recent years, many methods for 3D reconstruction with continuous heterogeneity have been developed, where the molecule under study can flex and stretch in various ways. CryoDRGN represents a continuous conformational space, trained on a single particle stack (Zhong et al., 2020; 2021a), later extended to ab initio settings (Zhong et al., 2021b; Levy et al., 2025). RECOVAR models the same variability by covariance estimation (Gilles & Singer, 2025). CryoAI, cryoSPIN and cryoFIRE predict per-image poses with an encoder; cryoFIRE estimates conformations as well (Levy et al., 2022a; Shekarforoush et al., 2024; Levy et al., 2022b). These networks are amortized over the images of one dataset and are trained for each molecule separately. CryoHype trains one network on a single stack of a thousand distinct molecules at known poses, with no structures held out (Gu et al., 2026). In contrast to these methods, our method models a distribution of volumes conditioned on a set of projection images.

2.2 Learning molecular priors

In this section we describe methods for 3D reconstruction that learn a prior on a set of different molecules. CryoFM learns an unconditional flow matching prior over densities from deposited maps and applies it to downstream tasks (Zhou et al., 2025; Li et al., 2025). It follows a general recipe for inverse problems, not necessarily related to molecular reconstruction: train an unconditional generative model, then steer it toward measurement consistency during sampling. The steering is done using a likelihood gradient at each denoising step of a diffusion model (Chung et al., 2023), or a flow-based analog (Martin et al., 2025; Kim et al., 2025; Pourya et al., 2026; Erbach et al., 2025). This general recipe for steering unconditional generative models pays the inference time optimization cost at every reconstruction. A different approach for 3D reconstruction with a learned prior is implemented in CryoFastAR. This model predicts poses using a feed-forward network, after which the volume is reconstructed by back-projection (Zhang et al., 2025). Like ours, it is trained across many molecules, but its learned component predicts the orientations rather than the 3D densities.

Two axes separate these approaches: whether reconstruction requires optimization specific to the dataset being processed, and whether it returns a point estimate or mechanism for sampling from a distribution. CryoFATHOM sits in the one combination none of them occupies, moving the conditioning into training so that an unseen molecule can be reconstructed simply by sampling via numerical integration of the learned flow.

	Point estimate	**Distribution over volumes**
Per-dataset optimization	RELION, CryoSPARC, CG, CryoAI	CryoDRGN, RECOVAR, CryoHype
Prior, inference-time steering	—	CryoFM
Prior, no steering	CryoFastAR	**CryoFATHOM**

3 Preliminaries

Cryo-EM image formation. Let $\mathcal{P}$ be the tomographic projection operator (aka X-ray transform) that takes a 3D scalar field $f(u_x, u_y, u_z)$ and integrates it along the u_z axis to produce an image,

$$(\mathcal{P}f)(u_x, u_y) = \int_{\mathbb{R}} f(u_x, u_y, u_z)\, du_z. \tag{1}$$

Let $\boldsymbol{x} : \mathbb{R}^3 \to \mathbb{R}$ be the electric potential of a rigid molecule that we wish to recover. In cryo-EM, each instance of the molecule in the sample is at a random orientation $R_i \in SO(3)$ and the images

captured by the microscope are tomographic projections of those instances (Bendory et al., 2020). For a centered molecule, an ideal imaging device would produce the clean image $\mathcal{P}R_i\,\boldsymbol{x}$, however, real electron microscope images include additive shot noise $\boldsymbol{\varepsilon}_i$ and a point spread function h_i. The resulting images are modeled as follows,

$$\boldsymbol{y}_i \;=\; h_i * \mathcal{P}R_i\,\boldsymbol{x} \;+\; \boldsymbol{\varepsilon}_i, \qquad i = 1, \ldots, k. \tag{2}$$

In the Fourier domain the point spread function is called the contrast transfer function $\mathrm{CTF}_i := \mathcal{F}h_i$. It acts by element-wise multiplication in the frequency domain. We assume that the orientations are given by an upstream stage of the pipeline and that the molecules have been centered. All computations are performed on a discrete grid with side length D. The information available for 3D reconstruction is therefore the conditioning set

$$S_c = \{(\boldsymbol{y}_i, R_i)\}_{i=1}^{k} \quad \text{where} \quad \boldsymbol{y}_i \in \mathbb{R}^{D^2},\; R_i \in SO(3). \tag{3}$$

Section 5 provides the simulation constants and Appendix B the implementation details.

Conditional flow matching. Flow matching learns a time-dependent velocity field that transports a source distribution (e.g. Gaussian) to the target data distribution (Lipman et al., 2023; Liu et al., 2023; Albergo & Vanden-Eijnden, 2023). It trains a neural network to output a time-dependent velocity field $v_\theta(\boldsymbol{x}, t, S_c)$, where θ are the weights of the network and S_c is the conditioning signal. This is done by minimizing the following loss function:

$$\mathcal{L}(\theta) \;=\; \mathbb{E}_{t,\,\boldsymbol{x}_0,\,(\boldsymbol{x}_1, S_c)} \Big\| \underbrace{v_\theta(\boldsymbol{x}_t, t, S_c)}_{\text{predicted velocity}} \;-\; \underbrace{(\boldsymbol{x}_1 - \boldsymbol{x}_0)}_{\text{target velocity}} \Big\|_2^2, \tag{4}$$

where $\boldsymbol{x}_t = (1-t)\,\boldsymbol{x}_0 + t\,\boldsymbol{x}_1$ has constant velocity $\boldsymbol{x}_1 - \boldsymbol{x}_0$. The expectation is over all $t \in [0, 1]$, $\boldsymbol{x}_0$ drawn from the source distribution, and $\boldsymbol{x}_1$ drawn from the target distribution together with its conditioning signal S_c. In our case, $\boldsymbol{x}_1$ is a molecule drawn from a training set of proteins, represented as a 3D array, and S_c is a set of k random tomographic projections of it, following the forward model described in Equation 2, along with their orientations. Given an input set S_c, sampling from the conditional target distribution is simple: you draw $\boldsymbol{x}_0$ from the source distribution and integrate it along the velocity field via $\frac{\mathrm{d}\boldsymbol{x}_t}{\mathrm{d}t} = v_\theta(\boldsymbol{x}_t, t, S_c)$. We integrate this ODE numerically from $t = 0$ to $t = 1$ with N Euler steps. For a comprehensive introduction on flow matching see Lipman et al. (2024).

4 Method

CryoFATHOM performs 3D reconstruction by posterior sampling: it draws volumes from $p(\boldsymbol{x} \mid S_c)$, a conditional distribution learned across many molecules. Sections 4.1 and 4.2 describe the network that parameterizes its velocity field, Section 4.3 how that network is trained, and Section 4.4 describes our sampling methods.

4.1 Model architecture

Our model has two main components: a projection encoder that turns the projection images and their orientations into conditioning tokens (Section 4.2), and the velocity field v_θ, implemented using a 3D diffusion transformer (DiT) model (Peebles & Xie, 2023). The volumes are represented as 3D arrays of dimension $64 \times 64 \times 64$. The *patchify* block cuts them into non-intersecting $4 \times 4 \times 4$ voxel blocks, and then linearly embeds them, forming 4096 *volume tokens*. The DiT has a depth of six blocks, each block applies self-attention over the volume tokens, then cross-attention with the conditioning tokens, and finally each token is processed by an MLP. The flow time t enters every block through adaptive layer normalization. A final linear layer *unpatchify* maps each token back to its original $4\times4\times4$ block of voxels. Cross-attention is the only route by which the projection images influence the reconstructed volume. Each block has its own key and value maps W_K and W_V, so the conditioning tokens are computed once and each block can extract different information from them. Self-attention uses 3D rotary position embeddings (Su et al., 2024; Heo et al., 2024), so volume tokens interact through their relative positions. Cross-attention does not, since a projection token describes a ray through the volume rather than a point on its grid; instead, volume and projection tokens both carry an absolute 3D coordinate (Section 4.2).

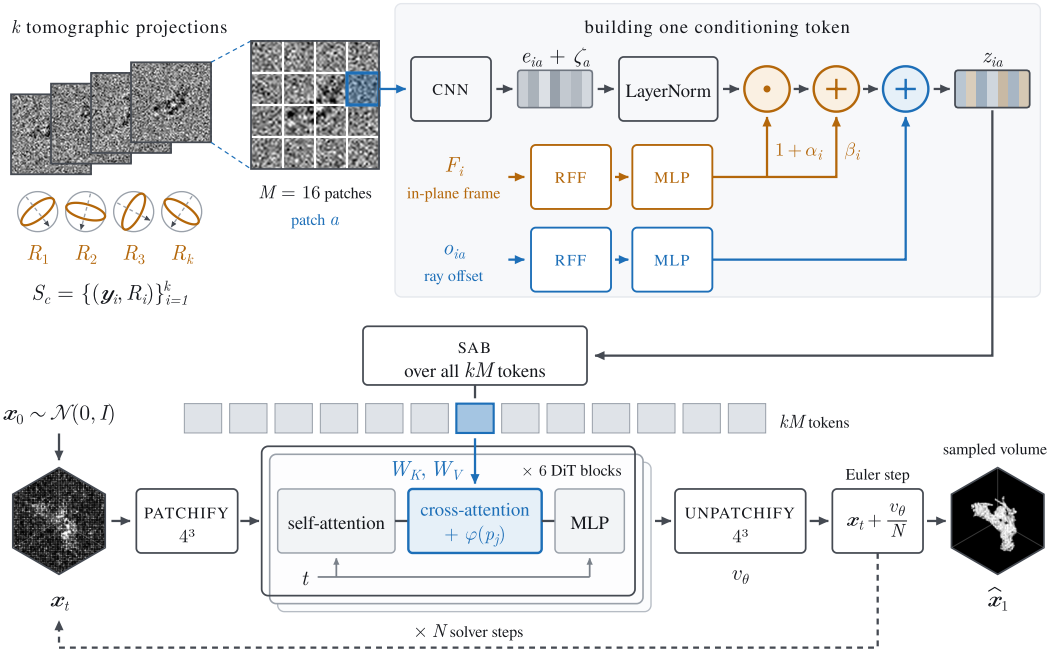


Figure 2: The projection encoder and the DiT velocity network. Each of k posed projections is cut into $M = 16$ patches. A patch token carries image content, the in-plane frame F_i scales and shifts the image content embedding as in FiLM, and o_{ia} is per-patch offset. A set attention block (SAB) mixes all kM tokens derived from the full set of projection images. Each DiT block re-projects the set of tokens into its own keys and values.

4.2 Geometry-aware projection encoder and conditioning

The conditioning signal S_c is a set of k projection images and their orientations $R_1, \ldots, R_k$. The encoder turns it into a set of tokens that the velocity network v_θ uses through cross-attention. Each projection is cut into $M = 16$ patches. Each patch becomes one token carrying the patch pixels, the orientation of its projection image, and where its ray passes through the volume. A set attention block then lets tokens from different projection images interact. Figure 2 illustrates the full architecture including the conditioning mechanism described in this section.

Each projection passes through a small convolutional network and is then cut into square patches a quarter of the image side, each linearly embedded as one token, so the conditioning set holds kM tokens. We empirically demonstrated that patchifying the projection images to multiple tokens is beneficial: a single token can carry the orientation but not where in the image its content lies, and so cannot express which parts of the volume it measures. Section 6.2 ablates the number of patches M per projection image. We denote by e_{ia} the image content of patch a of projection i, by ζ_a a learned embedding of its location in the patch grid, and by $u_a \in \mathbb{R}^2$ its center in the image plane.

We represent the orientation of projection i by its in-plane frame $F_i = \begin{bmatrix} r_i^{(1)} & r_i^{(2)} \end{bmatrix} \in \mathbb{R}^{3\times 2}$, the first two columns of R_i. The frame also places each patch in the volume: $o_{ia} = F_i\, u_a$ is the point where the ray through patch a of projection i crosses the plane through the origin perpendicular to the beam. The frame and the offset are both embedded with random Fourier features (RFF) γ_σ (Tancik et al., 2020), with separate bandwidths σ_p for positions and σ_r for orientations, each matched to the scale of its input (Appendix D.1). They enter the token as

$$z_{ia} \;=\; \mathrm{Norm}(e_{ia} + \zeta_a) \odot \big(1 + \alpha_i\big) \;+\; \beta_i \;+\; \psi\big(\gamma_{\sigma_p}(o_{ia})\big), \qquad (\alpha_i, \beta_i) \;=\; \chi\big(\gamma_{\sigma_r}(F_i)\big), \quad (5)$$

where ψ and χ are MLPs. The frame scales and shifts all M tokens of its view by the scalars (α_i, β_i) as in FiLM (Perez et al., 2018), and the per-patch offset is added to each token as $\psi\big(\gamma_{\sigma_p}(o_{ia})\big)$.

A single set attention block (Lee et al., 2019) runs over all kM tokens jointly. Attention is permutation equivariant and the cross-attention below sums over the whole token set, so by construction v_θ is invariant to the order of the projections and defined for any k.

In each block, the velocity network uses the conditioning tokens by cross-attention (Rombach et al., 2022), computed separately in each attention head, of dimension d_h. Let H be a matrix whose rows are the volume tokens h_j, let $\tilde{Z}$ be a matrix whose rows are the conditioning tokens $\tilde{z}_{ia}$ after SAB, and Φ a matrix whose row j is $\phi(\gamma_{\sigma_p}(p_j))$, an embedding of the position p_j of volume token j in the box, with ϕ a third MLP. Then cross attention is defined as,

$$\mathrm{softmax}\big(QK^\top/\sqrt{d_h}\big)\,V, \qquad Q = \big(\mathrm{Norm}(H) + \Phi\big)\,W_Q, \quad K = \tilde{Z}\,W_K, \quad V = \tilde{Z}\,W_V, \tag{6}$$

The query of volume token j thus carries $\gamma_{\sigma_p}(p_j)$ and the key of token (i, a) carries $\gamma_{\sigma_p}(o_{ia})$, in the same embedding, so their inner product can be large when volume token j lies near o_{ia}, a point on the ray through patch a of projection i (Appendix D.1).

4.3 Model training and pose resampling

We train v_θ on a training set of 850 proteins by minimizing the loss in Equation 4. Each optimizer step uses 32 examples in 16 micro-batches of two. Each micro-batch draws one projection count k for both of its examples, from a six-value ladder $\mathcal{K} = \{25, 50, 100, 300, 600, 1000\}$ with corresponding weights proportional to $(1, 1, 1, 1, 2, 2)$. Therefore, no optimizer step is specialized to one value of k: with 16 draws of k, a simple calculation shows a step covers 5.51 of the 6 values in expectation. For each example, we then pick a random source $\boldsymbol{x}_0$, a time $t \in [0, 1]$ and a random target molecule $\boldsymbol{x}_1$ with a conditioning set S_c that contains k random projections of $\boldsymbol{x}_1$, computed via the cryo-EM forward model in Equation 2. To save computation time during training, we precompute a large set of projections for each molecule and sample from it. The CTF and noise are applied afresh at every draw. A molecule is thus almost never seen twice with the same conditioning set. Interestingly, the smallest counts we evaluate lie below the training ladder's minimum, since extending it below $k = 25$ performed worse in our experiments. See Table 6 for the full optimization details.

4.4 Inference-time control

Solver steps. The number of solver steps N is chosen at inference time, and it moves the reconstruction between two regimes. A single Euler step returns the conditional mean $\mathbb{E}[\boldsymbol{x}_1 \mid S_c]$ (Appendix A), the minimum-mean-squared-error (MMSE) estimate, which is blurred because it averages over all volumes consistent with the projections. As you increase N the reconstructions appear sharper, like real molecules, but they correlate less well with the ground truth. This is another instance of the perception-distortion tradeoff (Blau & Michaeli, 2018). Section 6.1 provides an analysis on how one can mitigate that tradeoff using CFG when $N > 1$.

Classifier free guidance. CFG replaces the velocity at each step by $v_\emptyset + w\,(v_\theta(\boldsymbol{x}_t, t, S_c) - v_\emptyset)$, where $v_\emptyset$ is an unconditional velocity prediction and w, a scalar that controls the guidance strength, is held constant over the trajectory (Ho & Salimans, 2021). The unconditional velocity $v_\emptyset$ is typically learned by dropping the conditioning on a fraction of the examples during training. Ours needs no training: we bypass the encoder and run the same network with all-zero conditioning tokens, which switches its cross-attention off (Appendix G.1). We do not claim that $v_\emptyset$ approximates the unconditional velocity field. It works without any training, and in our experiments it outperformed a learned unconditional velocity field.

Guidance recovers part of the fidelity that extra solver steps lose (Section 6.1). Two settings control it: the strength w, and a hard gate applying CFG only while $t < t_{\mathrm{hi}}$, which skips the evaluations of the unconditional velocity outside that interval. Limiting guidance to part of the trajectory has precedent in image models (Kynkäänniemi et al., 2024), which guide only at intermediate noise levels; our gate instead keeps the first, noisiest steps. Both are tuned per step count on the validation split. Appendix G.1 gives the sweep and the selected values. At $N = 1$ guidance degrades the reconstruction and is not used. Guided steps evaluate the network twice while the gate is open, so due to our choices of t_{hi} the number of function evaluations (NFE) is 8 at $N = 4$ and 15 at $N = 10$. We also evaluated CFG++ (Chung et al., 2025) and found that in our setting, whose null is not a trained unconditional model, it at best matches the plain rollout on validation and falls short of our recipe (Appendix G.2).

Amplitude calibration. CFG tends to create output values that are too high, which is typically corrected by clipping or rescaling (Saharia et al., 2022; Lin et al., 2024). In our setting guidance inflates different Fourier shells non-uniformly. We therefore match the guided volume's radial power per Fourier shell, after masking, to that of the plain rollout at the same N and $\boldsymbol{x}_0$, using no ground truth (Appendix G.3). Since our metric of interest, Fourier shell correlation (Section 6), is invariant to radial reweighting, it leaves every resolution we report unchanged.

5 A BENCHMARK FOR HELD-OUT-MOLECULE RECONSTRUCTION

Our benchmark holds $1{,}000$ molecules from the Protein Data Bank (PDB), each rendered to a density map from its deposited atomic model. Candidates are asymmetric (C_1) PDB assemblies of 100 to 400 kDa with a deposited resolution better than 4Å. We then deduplicate by composition. Each entry is identified by the set of 30% sequence-identity clusters of its polymer entities, and we keep one entry per set to guarantee diversity. The result is divided $850/75/75$ into training, validation and test, stratified by molecular weight. See Appendix C for more details. The closest existing benchmark is CryoBench (Jeon et al., 2024), which standardizes simulated data and metrics but focuses on continuous and compositional heterogeneity. Its largest compositional set, Tomotwin-100, holds 100 complexes in one unsplit stack. Training a prior across molecules and testing it on unseen ones needs many molecules and a split between them.

Headline experiments use a 64^3 grid at 6Å per voxel, which puts the Nyquist limit at 12Å. The released code also builds two other resolution rungs: a cheaper 32^3 one, and a 128^3 one matching the working resolution of CryoBench. Noise is white and added after the CTF, at a signal-to-noise ratio (SNR) of 0.1 measured as a variance ratio over the disc inscribed in the image. Appendix B describes the SNR as we use it, its relation to the whole-image SNR convention used in other works, and the CTF parameters used. Orientations come from a fixed regular grid of $SO(3)$ with $36{,}864$ orientations. Every projection is precomputed, so a conditioning set can be redrawn at every training step with minimal computational cost (Section 4.3).

6 EXPERIMENTAL RESULTS

Every number below is measured on the held-out test set of the benchmark described in the last section. This is done after the training budget, the guidance recipe and the tokenization had been selected on the validation set. Each model is trained with three random seeds with the same optimization settings (Table 6) and every method is scored on the same projections of the 75 test molecules. We evaluate on a k-ladder of powers of two, $k \in \{4, 8, 16, \ldots, 1024\}$, none of which appear in the training ladder. The headline metric is the reconstruction resolution, based on where the Fourier shell correlation (FSC) against the ground truth falls below 0.5, as is standard in cryo-EM (Rosenthal & Henderson, 2003). We define one additional metric, the high-band FSC: the mean FSC over shells above 0.65 of the Nyquist frequency, finer than about 18Å for our data. We use this metric to compare our own inference-time settings and analyze how they trade sharpness against fidelity: resolution at 0.5 reports where correlation is lost, while the high-band FSC reports how much of it survives in the finest shells. The high-band FSC is not used as a basis for claiming any advantage over the classical baselines. Appendix F reports it for every method and count.

We compare against two classical baselines. Conjugate gradient (CG) and RELION. CG performs linear MMSE estimation of the molecule under the forward model of Section 3, given the exact projection operator, the exact noise level and a CTF regularizer fitted to the training split. We call it the oracle inverse, since it has oracle access to the exact forward model. RELION 5.1.0 (Scheres, 2012) is run under three treatments of the CTF: full correction, phase flipping, and none. Appendix E gives the settings and compares the three. All methods receive ground truth orientations. Both baselines are given the CTF parameters but our model is not.

6.1 MAIN RESULTS

Resolution gains from the prior. As shown in Table 1, when the number of projection images is 64 or less, CryoFATHOM outperforms the baselines, with $N = 10$ plain steps at $k = 64$ the single exception. However, the classical methods overtake it when the number of projections is large. At

Table 1: Spherical FSC resolution at 0.5 (Å, lower is better) on the 75 held-out test molecules; best per row in bold. CryoFATHOM entries are the mean over three training seeds with their standard deviation, the classical baselines have no seed axis. Our classical baselines are conjugate gradient (CG), RELION with ground truth CTF and RELION with phase flip. CryoFATHOM columns are our five samplers. Full row set in Appendix F.

	CryoFATHOM					Classical		
k	$N = 1$	$N = 4$	$N = 10$	$N = 4$ guided	$N = 10$ guided	CG oracle inv.	RELION full CTF	RELION ph. flip
4	**44.4**±0.6	45.0±0.7	46.9±1.0	44.9±0.6	45.5±0.3	108.2	98.6	97.2
8	**35.5**±0.4	36.5±0.4	38.1±0.7	36.0±0.4	36.6±0.3	65.6	63.7	60.4
16	**29.0**±0.4	29.9±0.3	31.7±0.5	29.4±0.4	29.9±0.3	39.0	42.9	38.6
32	**25.8**±0.2	26.5±0.1	28.0±0.2	26.0±0.2	26.5±0.2	29.4	32.4	29.3
64	**24.0**±0.2	24.5±0.2	25.9±0.1	24.1±0.2	24.5±0.2	25.0	27.3	25.4
128	23.0±0.2	23.3±0.1	24.4±0.1	22.9±0.2	23.2±0.2	**22.6**	24.4	23.2
256	22.4±0.1	22.6±0.1	23.7±0.1	22.2±0.1	22.6±0.2	**20.0**	22.0	21.6
512	22.0±0.1	22.3±0.1	23.5±0.1	21.9±0.1	22.3±0.2	**15.4**	18.3	18.8
1024	21.8±0.0	22.2±0.1	23.3±0.1	21.7±0.1	22.1±0.2	**12.9**	13.9	14.5

$k = 4$ we reach 44.4Å, about seven voxels across in a 384Å box, so even just four projections suffice to reconstruct the coarse shape. Figure 3 shows the $k = 16$ comparison by frequency: our method reconstructs the low and middle Fourier shells substantially better than the classical baselines. At $k \leq 32$ each of our samplers beats every classical baseline on every training seed. At $k = 64$ only the advantage over RELION with full CTF holds on every seed, and averaged over seeds it runs to about $k = 100$ against the oracle inverse and to $k = 128$ against RELION (Appendix F). From $k = 128$ to $k = 1024$ the classical reconstructions improve by 38% to 43% and ours by only 5%. This suggests the additional projections carry information our model fails to use.

Sharpness and fidelity. Solver steps set how much fine detail a reconstruction carries. A single step returns the conditional mean (Appendix A), producing a smoothed reconstruction. More steps produce samples that look sharper, and Figure 1 shows that difference as surface detail. With no guidance, this sharpness incurs a cost: in Table 1 the plain columns lose resolution as the step count rises, and ten plain steps are the worst setting at every count. Guidance recovers about two-thirds of that loss for both step counts. The high-band FSC moves with it: at $k = 16$ ten plain steps fall from 0.093 to 0.074 against the single step, and guidance brings them back to 0.086 (Table 8). Across the

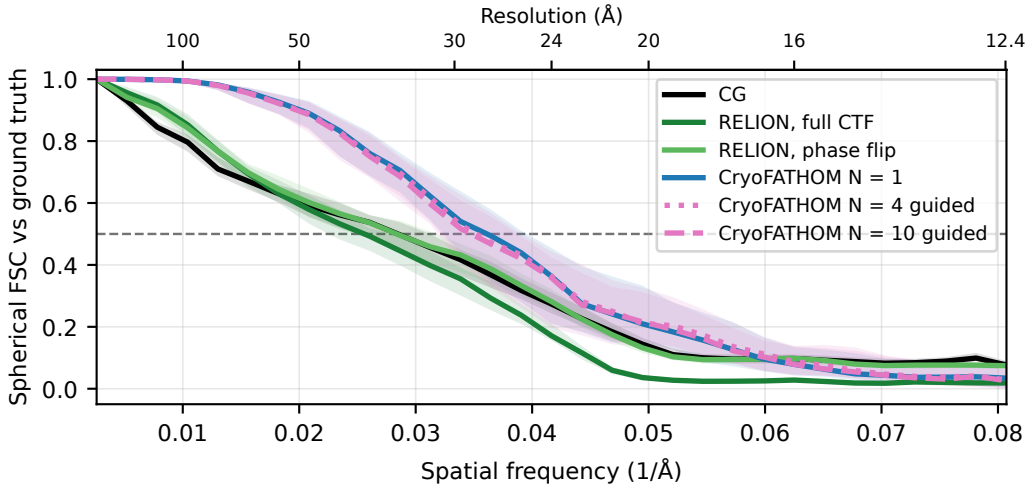


Figure 3: Shell-resolved correlation with ground truth at $k = 16$, median and interquartile range over the 75 test molecules; the upper axis is resolution in Å. Figure 5 repeats the comparison at every count on the evaluation ladder.

plain and guided samplers, guided four-step sampling is best on the high-band FSC from $k = 16$ upward, on all three seeds, and ten plain steps worst (Figure 6). Four guided steps therefore buy the surface detail of a multi-step sample at no measurable cost in the high-band FSC.

6.2 Ablation studies

The encoder turns each projection image into $M = 16$ patch tokens. Each token carries the pixels in the patch, the orientation of its projection, and where its ray crosses the volume (Section 4.2). We remove those ingredients one at a time, with three seeds per arm, and score every arm on the plain single-step sampler (Appendix H).

Withholding orientations. By removing the orientations the tokens carry image content alone. When this is done, the model stops improving as measurements accumulate, sitting at 66Å to 63Å resolution across the whole ladder with a mere 5% gain in resolution from $k = 4$ to $k = 1024$ against 51% for the full model (Table 11, Figure 7). We conclude that orientation information is a key ingredient in CryoFATHOM.

Tokens per projection. Coarsening the tokenization fails differently. At one token per projection the model reaches only 94Å to 74Å, while carrying more than twice the full model's parameters. The problem is not capacity. A single token cannot say which part of an image constrains which part of the volume. Between $M = 16$ and $M = 4$ the gap is small and confined to the smallest counts of projection images. We keep $M = 16$ because it is ahead at the small counts our claim concerns, and because its across-seed spread is smaller at almost every count (Table 12, Figures 8 and 9).

7 Conclusion

CryoFATHOM is a generative model for 3D molecular reconstruction that is based on conditional flow matching and trained on a dataset of known molecules. It takes a set of tomographic cryo-EM projection images as its input and produces a (discretized) 3D density of the molecule. The model is trained once, and reconstruction is done by a single ODE solve without any optimization or other expensive test-time computation. Thus it shows promise for high-throughput applications. The model assumes known orientations and does not take the CTF parameters, since in our experiments this did not improve the results.

In this paper we have demonstrated that CryoFATHOM outperforms established pipelines such as RELION on held-out protein reconstruction when the number of projections is small (up to around 100 images). Using just four projection images it can find the coarse shape of the molecule, and with a few dozen it recovers the low and middle Fourier shells substantially better than the classical baselines. Solver steps and classifier-free guidance act as inference-time controls that trade sharpness for fidelity. Our work has clear limitations. The resolution of reconstructions produced by the model flattens once hundreds of projections are available and our experiments were run at a 64^3 rung whose Nyquist limit is 12Å on simulated data, so neither experimental data nor near-atomic detail has been tested.

In future work we intend to raise the working resolution to the 128^3 rung, evaluate the sensitivity of the model to errors in the given orientations and test it on real experimental data. Improving the performance of the model in the high-k regime is another important goal that we intend to pursue, with many possible directions to take. For example, one could express the volumes and the velocity field in the Fourier space instead of the spatial domain, where projection images can be given as central slices thanks to the Fourier slice theorem. The setting of tomographic reconstruction from few measurements is not unique to cryo-EM and learning a prior across a family of related structures using conditional flow matching could apply in other domains as well. Cryo-electron tomography and medical imaging methods such as CT or MRI are natural candidates.

AI USE STATEMENT

We used generative AI tools for ideation and design discussions, code development, literature search, and drafting parts of the manuscript. The models are from Anthropic's Claude Opus family, accessed through Claude Code and the Claude chat interface. We have reviewed all AI-assisted work and take full responsibility for the final content, including all text, claims and artifacts.

REPRODUCIBILITY STATEMENT

The full code for generating our results and the benchmarks is submitted as an anonymized zip for review. It will be released in a public repository upon publication. Our code includes benchmark preparation, training and evaluation, together with the exact configurations used for the reported models, including ablation arms. To reproduce the results in this paper, follow the guidelines in the README file. Due to its size, the benchmark is regenerated rather than distributed as data. The benchmark generation code fetches atomic coordinates from the PDB, prepares them, and regenerates the projection stacks, applying the contrast transfer function and adding noise as described in Appendix B. The list of entries and their split assignment ship with the code, since the PDB's sequence-identity clusters are updated over time. Regenerating the benchmark takes a few hours on a single GPU and about 300 GB of disk at the 64^3 rung. Training compute is reported in Appendix I.

The appendices describe the technical details and parameters used to produce the results in this paper: the forward model, the contrast transfer function parameterization and the noise convention (Appendix B); the selection, deduplication, splitting, voxelization and pose-grid protocol used to assemble the benchmark (Appendix C); the architecture and every optimization setting (Appendix D); baseline configurations including RELION invocation parameters and the conjugate-gradient formulation (Appendix E); and the guidance sweep with its selection protocol, which states which measurements are validation and which are test (Appendix G).

REFERENCES

Michael S. Albergo and Eric Vanden-Eijnden. Building Normalizing Flows with Stochastic Interpolants. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=li7qeBbCR1t.

Alexander H. Barnett, Jeremy Magland, and Ludvig af Klinteberg. A Parallel Nonuniform Fast Fourier Transform Library Based on an "Exponential of Semicircle" Kernel. *SIAM Journal on Scientific Computing*, 41(5):C479–C504, 2019. doi:10.1137/18M120885X.

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. doi:10.1109/MSP.2019.2957822.

Junguo Bian, Jeffrey H. Siewerdsen, Xiao Han, Emil Y. Sidky, Jerry L. Prince, Charles A. Pelizzari, and Xiaochuan Pan. Evaluation of sparse-view reconstruction from flat-panel-detector cone-beam CT. *Physics in Medicine and Biology*, 55(22):6575–6599, 2010. doi:10.1088/0031-9155/55/22/001.

Yochai Blau and Tomer Michaeli. The Perception-Distortion Tradeoff. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6228–6237, 2018. doi:10.1109/CVPR.2018.00652.

Stephen K. Burley, Rusham Bhatt, Charmi Bhikadiya, Chunxiao Bi, Alison Biester, et al. Updated resources for exploring experimentally-determined PDB structures and Computed Structure Models at the RCSB Protein Data Bank. *Nucleic Acids Research*, 53(D1):D564–D574, 2025. doi:10.1093/nar/gkae1091.

Hyungjin Chung, Jeongsol Kim, Michael T. McCann, Marc L. Klasky, and Jong Chul Ye. Diffusion Posterior Sampling for General Noisy Inverse Problems. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=OnD9zGAGT0k.

Hyungjin Chung, Jeongsol Kim, Geon Yeong Park, Hyelin Nam, and Jong Chul Ye. CFG++: Manifold-constrained Classifier Free Guidance for Diffusion Models. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=E77uvbOTtp.

Julius Erbach, Dominik Narnhofer, Andreas Dombos, Bernt Schiele, Jan Eric Lenssen, and Konrad Schindler. Solving Inverse Problems with FLAIR. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 151498–151535, 2025. doi:10.52202/085713-4560.

J. Ryan Feathers, Erica K. Richael, Kayla A. Simanek, J. Christopher Fromme, and Jon E. Paczkowski. Structure of the RhlR-PqsE complex from *Pseudomonas aeruginosa* reveals mechanistic insights into quorum-sensing gene regulation. *Structure*, 30(12):1626–1636.e4, 2022. doi:10.1016/j.str.2022.10.008.

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

Nabin Giri, Liguo Wang, and Jianlin Cheng. Cryo2StructData: A large labeled cryo-EM density map dataset for AI-based modeling of protein structures. *Scientific Data*, 11:458, 2024. doi:10.1038/s41597-024-03299-9.

Jeffrey Gu, Minkyu Jeon, Ambri Ma, Serena Yeung-Levy, and Ellen D. Zhong. CryoHype: Reconstructing a thousand cryo-EM structures with transformer-based hypernetworks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 35280–35290, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Gu_CryoHype_Reconstructing_a_thousand_cryo-EM_structures_with_transformer-based_hypernetworks_CVPR_2026_paper.html.

Alexander Hägele, Elie Bakouch, Atli Kosson, Loubna Ben Allal, Leandro Von Werra, and Martin Jaggi. Scaling Laws and Compute-Optimal Training Beyond Fixed Training Durations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 76232–76264, 2024. doi:10.52202/079017-2427.

Richard Henderson. The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules. *Quarterly Reviews of Biophysics*, 28 (2):171–193, 1995. doi:10.1017/S003358350000305X.

Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary Position Embedding for Vision Transformer. In *European Conference on Computer Vision (ECCV)*, pp. 289–305, 2024. doi:10.1007/978-3-031-72684-2_17.

Jonathan Ho and Tim Salimans. Classifier-Free Diffusion Guidance. In *NeurIPS Workshop on Deep Generative Models and Downstream Applications*, 2021. URL https://openreview.net/forum?id=qw8AKxfYbI.

Andrii Iudin, Paul K. Korir, Sriram Somasundharam, Simone Weyand, Cesare Cattavitello, Neli Fonseca, Osman Salih, Gerard J. Kleywegt, and Ardan Patwardhan. EMPIAR: the Electron Microscopy Public Image Archive. *Nucleic Acids Research*, 51(D1):D1503–D1511, 2023. doi:10.1093/nar/gkac1062.

Minkyu Jeon, Rishwanth Raghu, Miro Astore, Geoffrey Woollard, Ryan Feathers, Alkin Kaz, Sonya M. Hanson, Pilar Cossio, and Ellen D. Zhong. CryoBench: Diverse and challenging datasets for the heterogeneity problem in cryo-EM. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 37, pp. 89468–89512, 2024. doi:10.52202/079017-2840.

Jeongsol Kim, Bryan Sangwoo Kim, and Jong Chul Ye. FlowDPS: Flow-Driven Posterior Sampling for Inverse Problems. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12328–12337, 2025. doi:10.1109/ICCV51701.2025.01146.

Gerard J. Kleywegt, Paul D. Adams, Sarah J. Butcher, Catherine L. Lawson, Alexis Rohou, et al. Community recommendations on cryoEM data archiving and validation. *IUCrJ*, 11(2):140–151, 2024. doi:10.1107/S2052252524001246.

Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying Guidance in a Limited Interval Improves Sample and Distribution Quality in Diffusion Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 122458–122483, 2024. doi:10.52202/079017-3892.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Networks. In *International Conference on Machine Learning (ICML)*, volume 97, pp. 3744–3753, 2019. URL https://proceedings.mlr.press/v97/lee19d.html.

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 (ECCV)*, pp. 540–557, 2022a. doi:10.1007/978-3-031-19803-8_32.

Axel Levy, Gordon Wetzstein, Julien N. P. Martel, Frédéric Poitevin, and Ellen D. Zhong. Amortized Inference for Heterogeneous Reconstruction in Cryo-EM. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 13038–13049, 2022b. doi:10.52202/068431-0947.

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

Yilai Li, Jing Yuan, Yi Zhou, Zhenghua Wang, Suyi Chen, Fengyu Yang, Haibin Ling, Shahar Z. Kovalsky, Xiaoqing Zheng, and Quanquan Gu. A Generative Foundation Model for Cryo-EM Densities. bioRxiv preprint 2025.12.29.696802, 2025. URL https://doi.org/10.64898/2025.12.29.696802.

Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common Diffusion Noise Schedules and Sample Steps are Flawed. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5392–5399, 2024. doi:10.1109/WACV57701.2024.00532.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow Matching for Generative Modeling. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Yaron Lipman, Marton Havasi, Peter Holderrieth, Neta Shaul, Matt Le, Brian Karrer, Ricky T. Q. Chen, David Lopez-Paz, Heli Ben-Hamu, and Itai Gat. Flow Matching Guide and Code. arXiv preprint arXiv:2412.06264, 2024. URL https://doi.org/10.48550/arXiv.2412.06264.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

Michael Lustig, David Donoho, and John M. Pauly. Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magnetic Resonance in Medicine*, 58(6):1182–1195, 2007. doi:10.1002/mrm.21391.

Ségolène Martin, Anne Gagneux, Paul Hagemann, and Gabriele Steidl. PnP-Flow: Plug-and-Play Image Restoration with Flow Matching. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=5AtHrq3B5R.

Amit Moscovich, Amit Halevi, Joakim Andén, and Amit Singer. Cryo-EM reconstruction of continuous heterogeneity by Laplacian spectral volumes. *Inverse Problems*, 36(2):024003, 2020. doi:10.1088/1361-6420/ab4f55.

William Peebles and Saining Xie. Scalable Diffusion Models with Transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182, 2023. doi:10.1109/ICCV51070.2023.00387.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *AAAI Conference on Artificial Intelligence*, volume 32, pp. 3942–3951, 2018. doi:10.1609/aaai.v32i1.11671.

Eric F. Pettersen, Thomas D. Goddard, Conrad C. Huang, Elaine C. Meng, Gregory S. Couch, Tristan I. Croll, John H. Morris, and Thomas E. Ferrin. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Science*, 30(1):70–82, 2021. doi:10.1002/pro.3943.

Mehrsa Pourya, Bassam El Rawas, and Michael Unser. Flower: A Flow-Matching Solver for Inverse Problems. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=QGd34p02mI.

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

Ali Rahimi and Benjamin Recht. Random Features for Large-Scale Kernel Machines. In *Advances in Neural Information Processing Systems (NIPS)*, volume 20, 2007. URL https://papers.nips.cc/paper_files/paper/2007/hash/013a006f03dbc5392effeb8f18fda755-Abstract.html.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685, 2022. doi:10.1109/CVPR52688.2022.01042.

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. doi:10.1016/j.jmb.2003.07.013.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 36479–36494, 2022. doi:10.52202/068431-2643.

Sjors H. W. Scheres. RELION: Implementation of a Bayesian approach to cryo-EM structure determination. *Journal of Structural Biology*, 180(3):519–530, 2012. doi:10.1016/j.jsb.2012.09.006.

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 (NeurIPS)*, volume 37, pp. 55785–55809, 2024. doi:10.52202/079017-1773.

Yu-hsuan Shih, Garrett Wright, Joakim Andén, Johannes Blaschke, and Alex H. Barnett. cuFINUFFT: a load-balanced GPU library for general-purpose nonuniform FFTs. In *IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pp. 688–697, 2021. doi:10.1109/IPDPSW52791.2021.00105.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with Rotary Position Embedding. *Neurocomputing*, 568:127063, 2024. doi:10.1016/j.neucom.2023.127063.

Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 7537–7547, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/55053683268957697aa39fba6f231c68-Abstract.html.

Lanhui Wang, Yoel Shkolnisky, and Amit Singer. A Fourier-based Approach for Iterative 3D Reconstruction from Cryo-EM Images. arXiv preprint 1307.5824, 2013. URL https://doi.org/10.48550/arXiv.1307.5824.

Simon Wiedemann and Reinhard Heckel. A deep learning method for simultaneous denoising and missing wedge reconstruction in cryogenic electron tomography. *Nature Communications*, 15(1): 8255, 2024. doi:10.1038/s41467-024-51438-y.

Marcin Wojdyr. GEMMI: A library for structural biology. *Journal of Open Source Software*, 7(73): 4200, 2022. doi:10.21105/joss.04200.

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 *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8462–8471, 2025. doi:10.1109/ICCV51701.2025.00792.

Ellen D. Zhong, Tristan Bepler, Joseph H. Davis, and Bonnie Berger. Reconstructing continuous distributions of 3D protein structure from cryo-EM images. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://openreview.net/forum?id=SJxUjlBtwB.

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, 2021a. doi:10.1038/s41592-020-01049-4.

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 *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4046–4055, 2021b. doi:10.1109/ICCV48922.2021.00403.

Yi Zhou, Yilai Li, Jing Yuan, and Quanquan Gu. CryoFM: A Flow-based Foundation Model for Cryo-EM Densities. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=T4sMzjy7fO.

A ONE SOLVER STEP RETURNS THE CONDITIONAL MEAN

The statement below is about conditional flow matching in general, so it is written for generic random vectors $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$ and a conditioning variable c. Section 4.4 applies it with $\boldsymbol{x}_1$ being a random molecular volume and c the conditioning set S_c.

Proposition 1. *Let $(\boldsymbol{x}_1, c)$ have $\mathbb{E}\|\boldsymbol{x}_1\|_2^2 < \infty$, and let $\boldsymbol{x}_0$ be independent of $(\boldsymbol{x}_1, c)$ with $\mathbb{E}\|\boldsymbol{x}_0\|_2^2 < \infty$. For $t \in [0,1]$, let $\boldsymbol{x}_t = (1-t)\,\boldsymbol{x}_0 + t\,\boldsymbol{x}_1$, and define*

$$v^\star(\boldsymbol{x}, t, c) \;=\; \mathbb{E}\big[\,\boldsymbol{x}_1 - \boldsymbol{x}_0 \;\big|\; \boldsymbol{x}_t = \boldsymbol{x},\; c\,\big], \tag{7}$$

choosing, at every t, a version of this conditional expectation that is jointly measurable in $(\boldsymbol{x}, t, c)$. Then $v^\star$ minimizes the flow matching objective

$$\mathcal{L}(v) \;=\; \mathbb{E}_{t,\,\boldsymbol{x}_0,\,(\boldsymbol{x}_1,c)} \big\| v(\boldsymbol{x}_t, t, c) \;-\; (\boldsymbol{x}_1 - \boldsymbol{x}_0) \big\|_2^2 \tag{8}$$

over measurable velocity fields, for every distribution of t on $[0,1]$, and a single Euler step of the sampling ODE from $t = 0$ to $t = 1$ with $v^\star$ returns

$$\hat{\boldsymbol{x}} \;=\; \boldsymbol{x}_0 \;+\; v^\star(\boldsymbol{x}_0, 0, c) \;=\; \mathbb{E}\big[\,\boldsymbol{x}_1 \;\big|\; c\,\big], \tag{9}$$

the MMSE estimate of $\boldsymbol{x}_1$ given c.

Proof. Fix t. Among measurable functions of $(\boldsymbol{x}_t, c)$, the squared error $\mathbb{E}\|(\boldsymbol{x}_1 - \boldsymbol{x}_0) - f(\boldsymbol{x}_t, c)\|_2^2$ is minimized by the conditional mean $f = \mathbb{E}[\,\boldsymbol{x}_1 - \boldsymbol{x}_0 \mid \boldsymbol{x}_t, c\,]$, which is $v^\star(\cdot, t, \cdot)$. The objective averages this error over t, so $v^\star$ minimizes it whatever the distribution of t. Any other minimizer agrees with $v^\star$ almost everywhere under the law of $(t, \boldsymbol{x}_t, c)$. When t has a density, the objective therefore does not determine the field at $t = 0$.

At $t = 0$ the interpolant gives $\boldsymbol{x}_t = \boldsymbol{x}_0$. Since $\boldsymbol{x}_0$ is independent of $(\boldsymbol{x}_1, c)$, conditioning on it says nothing about $\boldsymbol{x}_1$ once c is given, so

$$v^\star(\boldsymbol{x}_0, 0, c) \;=\; \mathbb{E}\big[\,\boldsymbol{x}_1 - \boldsymbol{x}_0 \;\big|\; \boldsymbol{x}_0,\; c\,\big] \;=\; \mathbb{E}\big[\,\boldsymbol{x}_1 \;\big|\; c\,\big] \;-\; \boldsymbol{x}_0. \tag{10}$$

A single Euler step from $t = 0$ to $t = 1$ has step size 1, which gives Equation 9. ■

A trained network approximates $v^\star$, so the single-step reconstruction approximates $\mathbb{E}[\,\boldsymbol{x}_1 \mid c\,]$ to the accuracy the network achieves at $t = 0$.

B FORWARD MODEL AND SIMULATION DETAILS

This appendix describes the implementation details of the forward model stated in Section 3. Its constants are collected in Section 5, and the dataset they describe is built as in Appendix C.

B.1 THE PROJECTION OPERATOR

The operator $\mathcal{P}R_i$ of Equation 2 and its adjoint are evaluated in the Fourier domain. A rotation acts on the potential by $(R\,\boldsymbol{x})(u) = \boldsymbol{x}(Ru)$ for $u \in \mathbb{R}^3$, so R maps the coordinates of the microscope into the frame of the molecule. By the Fourier slice theorem, the discrete 2D transform of $\mathcal{P}R\,\boldsymbol{x}$ is the central plane of the 3D transform of $\boldsymbol{x}$ spanned by the first two columns of R. A projection therefore comes from sampling the volume's Fourier coefficients on that plane and performing an inverse 2D FFT. The sampling is a type-2 non-uniform fast Fourier transform and the adjoint is the corresponding type-1 transform, computed using cuFINUFFT (Barnett et al., 2019; Shih et al., 2021). Forward and adjoint are transposes up to floating-point precision, as the conjugate-gradient baseline of Appendix E requires.

Orientations are stored in the convention used by cryoDRGN and CryoBench (Zhong et al., 2021a; Jeon et al., 2024). On matched poses the operator reproduces CryoBench's real-space projector at correlation 0.9988; the shortfall is that projector's trilinear interpolation.

B.2 CONTRAST TRANSFER FUNCTION

The CTF follows the astigmatic parameterization of Table 2, computed with cryoDRGN's implementation (Zhong et al., 2021a). Only the three defocus quantities vary between images. They are drawn, as in CryoBench (Jeon et al., 2024), from the empirical distribution of EMPIAR-11247 (Feathers et al., 2022; Iudin et al., 2023); the rest are user-supplied constants. The remaining parameters are taken from the dataset. The envelope is ours: deposited records carry no B-factor term, so we add a Gaussian taper that removes at most 16% of the amplitude, at Nyquist, and neither creates nor moves a zero of the CTF.

Table 2: The simulated microscope. Only the first three quantities vary between images.

Symbol	Quantity	Value
$\mathrm{dfu}, \mathrm{dfv}$	defocus along the two principal axes	per image
dfang	astigmatism angle	per image
V	accelerating voltage	200 kV
C_s	spherical aberration	2.7 mm
A_c	amplitude contrast	0.1
ϕ	additional phase shift	0
B	Gaussian envelope (B-factor)	100Å^2

B.3 NOISE AND THE SNR CONVENTION

After applying the CTF, we add white Gaussian noise, $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(0, \sigma^2 I)$ with one value of σ for the entire dataset. A signal-to-noise ratio does not specify it, since implementations differ over the region measured, over whether it precedes or follows the CTF, and over variances against amplitudes. Ours is a variance ratio measured after the CTF over the disc inscribed in the image,

$$\sigma \;=\; \frac{s_{\mathrm{ref}}}{\sqrt{\mathrm{snr}}}, \qquad s_{\mathrm{ref}} \;=\; \mathrm{std}\big(\mathrm{CTF}(\mathcal{P}R\boldsymbol{x}) \,\big|\, \mathrm{disc}\big), \tag{11}$$

where s_{ref} is the standard deviation of the CTF-filtered clean projections over that disc, pooled over the training molecules and their poses. We use the disc because it is the exact support of every projection. Volumes are masked to the sphere inscribed in the box (Appendix C), and a sphere of radius $D/2$ projects to a disc of radius $D/2$ at every orientation, so the corners are zero. CryoBench and cryoDRGN measure over the whole image instead (Jeon et al., 2024; Zhong et al., 2021a). At this rung our label is 1.28 times theirs for the same physical noise, so our 0.1 corresponds to 0.078 measured over the whole image.

C BENCHMARK CONSTRUCTION

Sources and selection. Candidate entries come from the Cryo2StructData metadata (Giri et al., 2024), distributed through Harvard Dataverse under DOI `10.7910/DVN/JMN60H`. The metadata lists 7,386 entries. Filtering on molecular weight between 100 and 400 kDa, deposited resolution better than 4Å and a single-particle experimental method leaves 3,954. For each of the filtered molecules we read two fields from the PDB (Burley et al., 2025): the global symmetry of its first assembly, and the membership of every polymer entity in the 30% sequence-identity clusters the PDB maintains. Restricting to asymmetric (C_1) assemblies leaves 2,349, and the deduplication below leaves 1,134. A random sample of these under a fixed seed gives the benchmark's 1,000 molecules. Entries the PDB cannot resolve were dropped.

Deduplication and splits. The deduplication key of an entry is the sorted set of the 30%-identity cluster identifiers of all its polymer entities, so it identifies a complex by composition rather than by any single chain. We keep one representative per distinct key, the entry with the best deposited resolution. An entry for which the PDB reports no cluster membership is kept as a singleton. We then split the deduplicated set, stratifying across ten molecular-weight quantiles under a fixed seed. This gives 850 training, 75 validation and 75 test molecules with matched weight distributions (Table 3).

Table 3: Per-split statistics. Molecular weight and deposited resolution are given as the median with the interquartile range.

Split	Molecules	Molecular weight (kDa)	Deposited resolution (Å)
Train	850	199.4 [154.7–274.4]	3.20 [2.90–3.50]
Validation	75	194.8 [155.8–283.2]	3.10 [2.91–3.42]
Test	75	200.4 [156.0–268.8]	3.20 [2.80–3.60]

Volumes and poses. We fetch each selected entry as its first assembly and prepare it with `gemmi` (Wojdyr, 2022): we take the first model, strip waters, ions and ligands, resolve alternate locations to the primary conformer, and center the result on its centroid. We archive that prepared file, and everything downstream is regenerated from it. It is voxelized with ChimeraX's `molmap` utility (Pettersen et al., 2021) at 3Å map resolution on a 1.5Å grid, giving an intermediate 256^3 map. That map is Fourier-downsampled to the target rung by cropping the centered transform. With the box fixed at 384Å, the three rungs follow as 32^3, 64^3 and 128^3 at 12, 6 and 3Å per voxel. Downsampling this way band-limits the ground truth to the rung's Nyquist frequency by construction. The downsampled volume is masked to the sphere of radius $D/2$ inscribed in the box, so its support does not depend on orientation. Every projection is computed from that masked volume, before any normalization (Appendix D.2), at all 36,864 poses of a fixed grid uniform over $SO(3)$: 768 viewing directions by 48 in-plane rotations.

D Architecture and hyperparameters

Tables 4 and 5 give the architecture of the velocity network (Section 4.1) and of the projection encoder (Section 4.2), which together hold 91.2M parameters. Table 6 gives the optimization settings. The subsections below give the positional scheme and the initialization.

Table 4: Velocity network (Section 4.1).

Box size (rung)	64^3 at 6.0Å per voxel
Patch	4^3 voxels, giving 4096 tokens
Depth	6 blocks
Width d	768
Heads / head dimension d_h	8 / 96
Feed-forward	SwiGLU, hidden 2048
Self-attention position	three-axis axial RoPE, base 10^4
Cross-attention position	absolute coordinate on the query (Section 4.2)
Time conditioning	adaptive layer norm, zero-initialized
Dropout	0.2
Parameters	78.7M

Table 5: Projection encoder (Section 4.2).

Input	k projections at 64^2
Convolutional stem	2 layers, 32 and 64 channels, kernel 3, stride 1, GELU, GroupNorm
Patch embedding	linear on 16^2-pixel patches of the stem output, $M = 16$ per projection
Token width d_c	512
Patch slot embedding	learned, $M \times d_c$, zero-initialized
Geometry features	RFF, 64 bands, $\sigma_p = 3.0$, $\sigma_r = 1.0$
Fusion norm	LayerNorm, no affine
Set attention	1 block, 8 heads
Dropout	0.2
Parameters	12.5M

Table 6: Optimization settings. Every model reported in this paper, the ablation arms of Section 6.2 included, is trained with these settings.

Mixed-k ladder $\mathcal{K}$	$\{25, 50, 100, 300, 600, 1000\}$
Ladder weights π	$(\frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{4}, \frac{1}{4})$
Optimizer	AdamW
Learning rate	10^{-4}
Learning-rate schedule	warmup, constant, linear decay to zero (Hägele et al., 2024)
Warmup	5 epochs
Training budget	1200 epochs
Terminal decay	final 100 epochs
Micro-batch / accumulation / effective batch	$2/16/32$
Weight decay	10^{-4}
Gradient clipping	global norm 1.0
Precision	bfloat16
Flow-time distribution	uniform on $[0, 1]$
Checkpoint reported	final iterate; no EMA, no early stopping
Training seeds	3

D.1 Positional encoding

Rotary embeddings work for self-attention, where volume tokens share one grid and interact through their relative displacement. A projection token and a voxel block sit at no relative offset in a shared grid, so cross-attention matches them through an absolute coordinate that both carry.

The random Fourier features of Section 4.2 embed both positions and the frame F_i. For a vector v and a bandwidth σ,

$$\gamma_\sigma(v) = \big[\sin(2\pi\Omega v),\ \cos(2\pi\Omega v)\big], \qquad \mathbb{E}\,\gamma_\sigma(v)^\top \gamma_\sigma(v') \ \propto\ \exp\big(-2\pi^2\sigma^2\|v - v'\|^2\big), \tag{12}$$

where Ω is a Gaussian matrix drawn once that is never trained. This is the standard random-feature identity (Rahimi & Recht, 2007): the expected inner product of two embedded coordinates depends only on their separation, and decays with it. That is what makes the query-key term of Section 4.2 large when voxel block j lies near o_{ia}. Positions and orientations get separate bases, calibrated against different geometries. The positional bandwidth σ_p is set so that two positions one volume token apart correlate about one half. The rotation bandwidth σ_r is set against the angular spacing of the orientation grid, so that neighboring orientations are smoothly related while distant ones are told apart. Table 5 gives both.

D.2 Initialization and normalization

The zero-initialized layers are the readout that inverts the patchification, the output projection of every cross-attention, and the set attention block's output and feed-forward layers. The perceptrons ψ and χ of Equation 5 and the query embedding ϕ have two layers each, with the output layer zero-initialized as well. At initialization the predicted velocity is therefore zero, the frame modulation is the identity and neither positional term contributes. Every other parameter uses PyTorch's default initialization. The fusion normalization is a layer norm without affine parameters; the zero-initialized slot embedding ζ_a enters before it and the offset term after.

The flow target is a normalized volume: each is divided by a single constant measured on the training split, with no mean subtracted, so that scaling a volume scales its projections by the same factor. The network therefore learns its output in that scaled space. Conditioning projections are z-scored with constants from the same split.

E Baseline configurations

Both baselines reconstruct from exactly the same measurements that the model is scored on: the projections are exported once per evaluation cell and every method reads them, so images, defocus values and noise realization are shared. Every method receives the same true orientations.

Conjugate gradient. The oracle inverse (Section 6): the linear MMSE estimator under the forward model of Section 3, solved on our own projection operator and its transpose. Writing A for the stacked projection operator, C for the per-image CTF, m for the spherical mask and $\boldsymbol{y}$ for the measurements, it solves

$$m\big(A^\top C^2 A + \Lambda\big)\, m\,\hat{\boldsymbol{x}} \;=\; mA^\top C\,\boldsymbol{y}, \qquad \Lambda(\kappa) \;=\; \frac{D^3\sigma^2}{P_{\boldsymbol{x}}(\kappa)}, \tag{13}$$

by conjugate gradients for 200 iterations. Λ is a filter on the volume's Fourier transform, with value $\Lambda(\kappa)$ on the shell of radius κ. The regularizer carries no tunable constant: σ is the exact noise level (Appendix B.3) and $P_{\boldsymbol{x}}$ the radially averaged power spectrum of the training split. The projection operator is rank deficient at these counts, so the inversion is regularized rather than literal. FIRM solves the same normal equations with a non-uniform FFT, for the same setting of known orientations (Wang et al., 2013).

RELION. Run as an unmodified binary, version 5.1.0: the one-pass reconstruction tool RELION provides for images whose orientations are already known. Three treatments of the CTF are reported: full knowledge, phase flipping, and none. Every optional setting is left at its default, so the symmetry is C_1, the padding factor is 2, the resolution limit is Nyquist and the pixel size comes from the star file.

CTF treatments for small k. Phase flipping outperforms full CTF correction below several hundred projections, and the ordering inverts at larger k. This follows from one-pass reconstruction without a regularizing prior: the CTF correction divides by a CTF-weighted sum of contributions. When few views cover a shell that sum is small, so the correction amplifies noise. Phase flipping is nearly a no-op there for a related reason: at the defocus values simulated here the first zero of the CTF lies beyond the resolution any method reaches at small k, so the reconstruction sits entirely inside the first passband. On resolution it is therefore indistinguishable from no CTF correction at all below several hundred projections. The two are not equivalent: the high-band FSC separates them at every count (Table 8).

F Full results grid

All tables report the held-out test split with the complete row set: our five samplers, the plain rollout at $N \in \{1, 4, 10\}$ and the two guided variants, and the four classical settings. Table 1 is a curated subset of Table 7.

CryoFATHOM entries are the mean over three training seeds with their standard deviation, the classical baselines have no seed axis. Figure 4 plots the evaluation ladder of Table 7 against k, and Figure 5 resolves the same comparison across frequency at every count on the ladder. The training-ladder block's $k = 100$ row falls between the evaluation ladder's 64 and 128, where the ordering against the oracle inverse reverses (Section 6.1). Table 8 gives the high-band FSC of Section 6 for every method at every count, and Figure 6 compares the five samplers on resolution and on the high-band FSC.

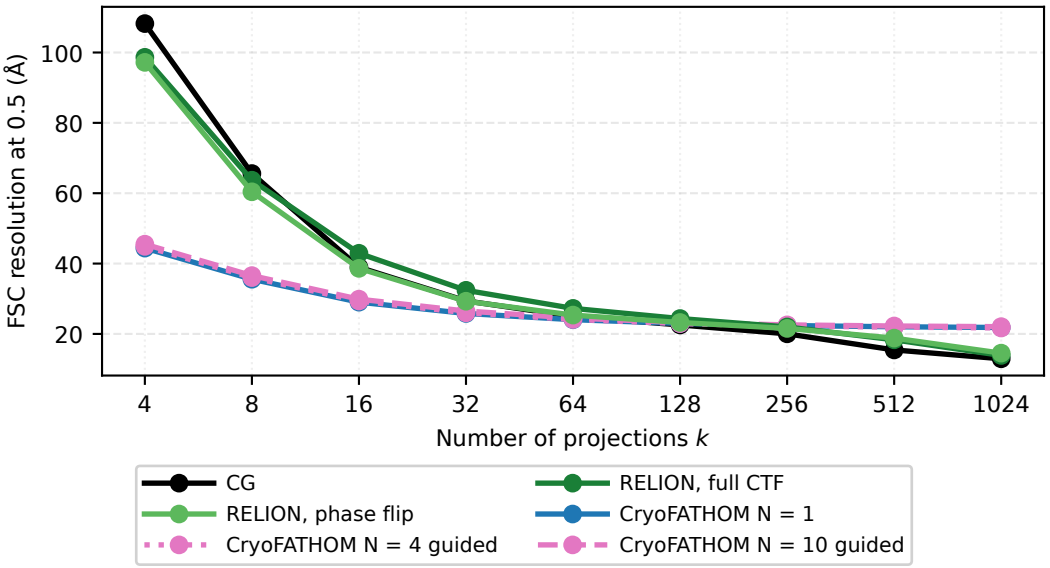


Figure 4: Spherical FSC resolution at 0.5 (lower is better) over the whole evaluation ladder. CryoFATHOM dominates for small values of k but its curve flattens around 128, so the classical methods beat it when the number of projections is large.

Table 7: Spherical FSC resolution at 0.5 (Å, lower is better) on the held-out test split. The upper block is the evaluation ladder, no value of which the model was trained at; the lower block is the six values of the training ladder. Below several hundred projections RELION's phase-flip and no-CTF columns are nearly indistinguishable on this metric. See also Table 8 where we show that the high-band FSC distinguishes between them.

	CryoFATHOM					Classical			
k	$N=1$	$N=4$	$N=10$	$N=4$ guided	$N=10$ guided	CG oracle inv.	RELION full CTF	RELION ph. flip	RELION no CTF
Evaluation ladder									
4	44.4±0.6	45.0±0.7	46.9±1.0	44.9±0.6	45.5±0.3	108.2	98.6	97.2	97.3
8	35.5±0.4	36.5±0.4	38.1±0.7	36.0±0.4	36.6±0.3	65.6	63.7	60.4	60.2
16	29.0±0.4	29.9±0.3	31.7±0.5	29.4±0.4	29.9±0.3	39.0	42.9	38.6	38.4
32	25.8±0.2	26.5±0.1	28.0±0.2	26.0±0.2	26.5±0.2	29.4	32.4	29.3	29.3
64	24.0±0.2	24.5±0.2	25.9±0.1	24.1±0.2	24.5±0.2	25.0	27.3	25.4	25.4
128	23.0±0.2	23.3±0.1	24.4±0.1	22.9±0.2	23.2±0.2	22.6	24.4	23.2	23.3
256	22.4±0.1	22.6±0.1	23.7±0.1	22.2±0.1	22.6±0.2	20.0	22.0	21.6	21.7
512	22.0±0.1	22.3±0.1	23.5±0.1	21.9±0.1	22.3±0.2	15.4	18.3	18.8	20.6
1024	21.8±0.0	22.2±0.1	23.3±0.1	21.7±0.1	22.1±0.2	12.9	13.9	14.5	20.1
Training ladder									
25	26.7±0.2	27.4±0.1	29.1±0.2	27.0±0.2	27.5±0.2	31.8	35.4	31.6	31.6
50	24.4±0.2	25.0±0.1	26.4±0.1	24.5±0.2	25.0±0.2	26.2	28.7	26.4	26.4
100	23.2±0.2	23.7±0.2	24.9±0.2	23.2±0.2	23.6±0.3	23.3	25.3	23.9	23.9
300	22.3±0.1	22.6±0.1	23.7±0.1	22.2±0.1	22.5±0.2	19.1	21.4	21.1	21.4
600	21.9±0.1	22.3±0.1	23.4±0.2	21.8±0.1	22.2±0.2	14.2	17.2	17.9	20.5
1000	21.8±0.1	22.2±0.1	23.3±0.2	21.7±0.1	22.1±0.2	13.0	14.0	14.6	20.1

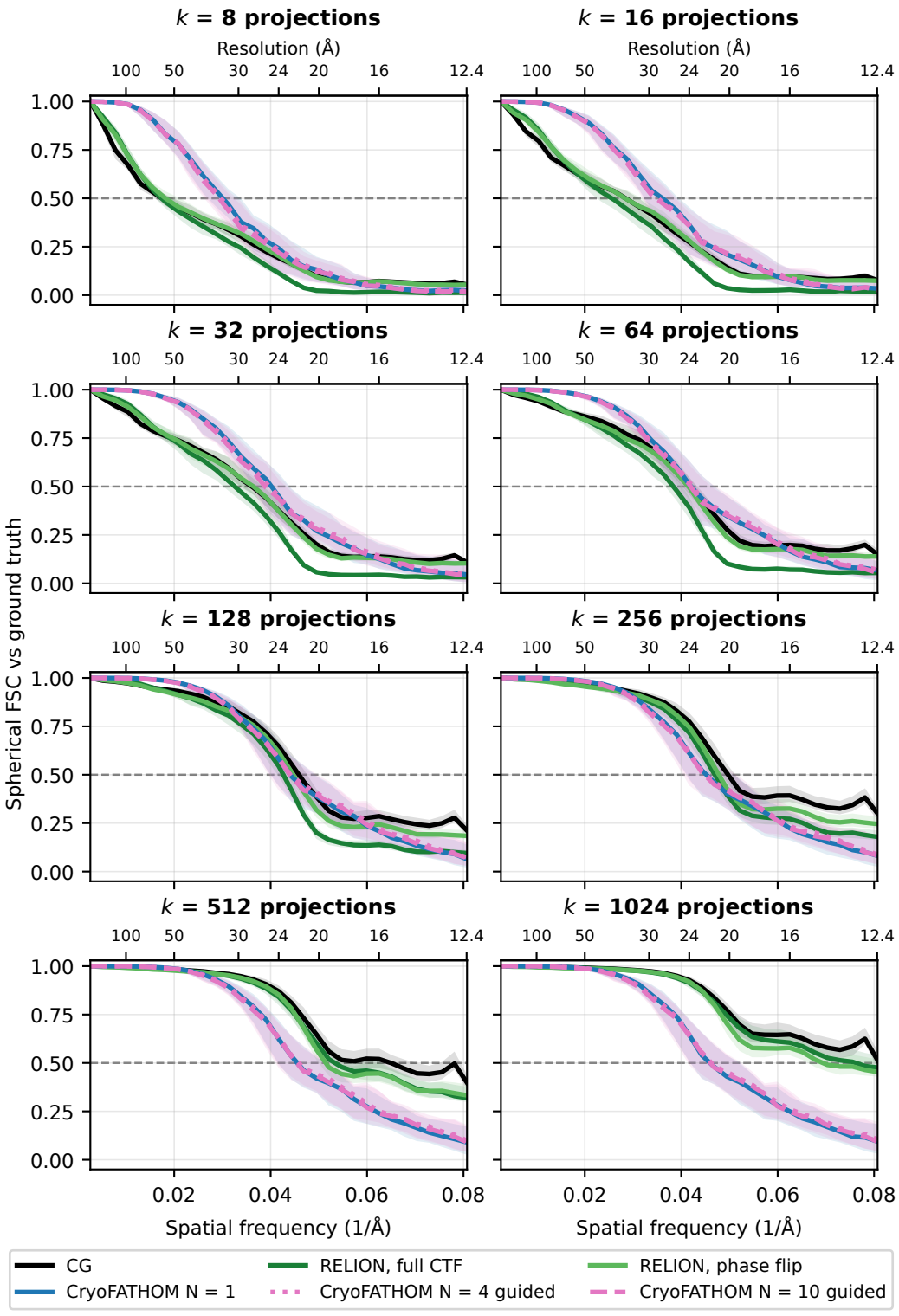


Figure 5: Shell-resolved correlation with ground truth across the evaluation ladder, median and interquartile range over the 75 test molecules. The crossover is visible here: our curves lead in the low and middle shells through $k = 64$, the two are close at $k = 128$, and the oracle inverse is clearly ahead from $k = 256$ upward.

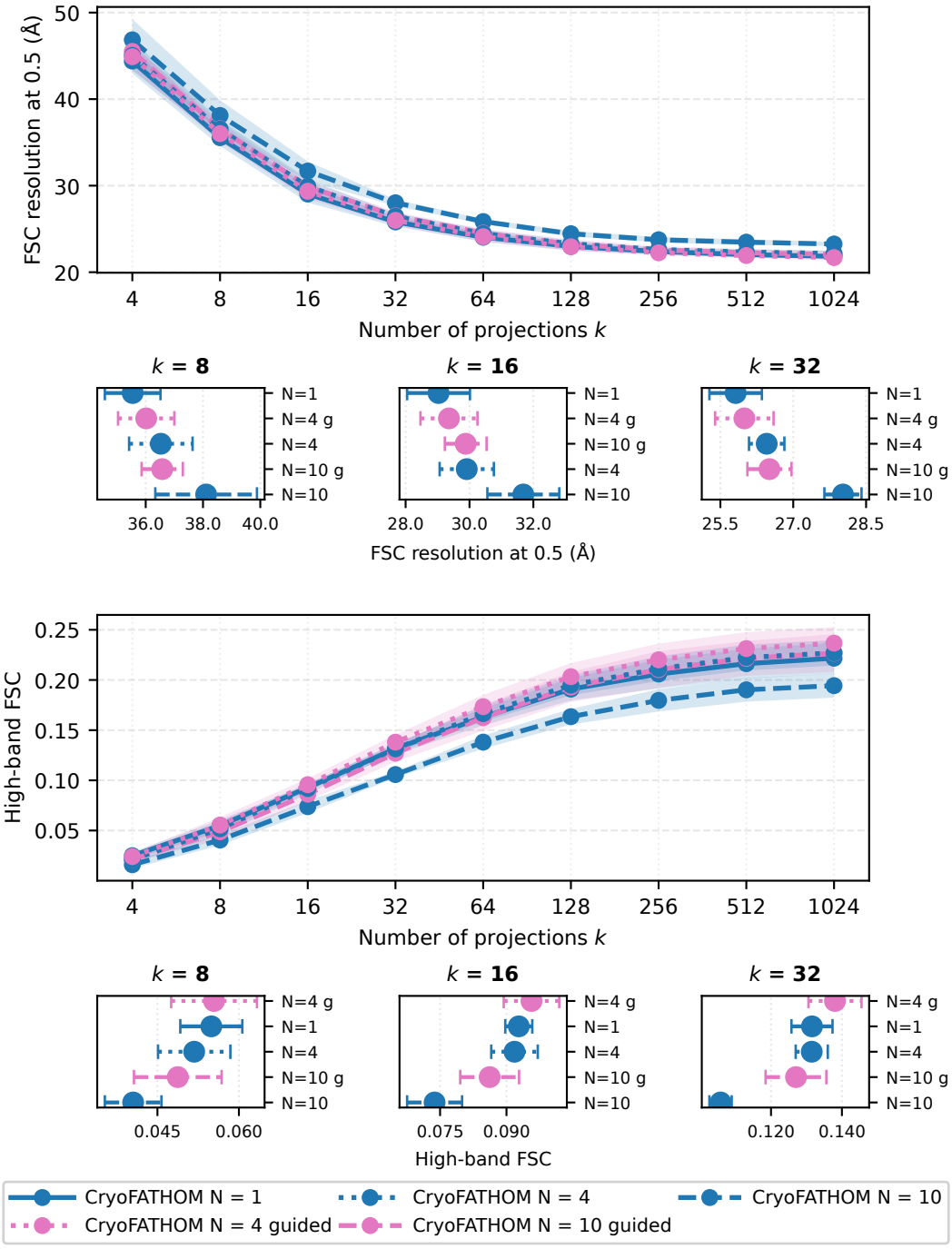


Figure 6: The performance of our five samplers by FSC resolution (top panel, lower is better) and by the high-band FSC (bottom panel, higher is better), bands and intervals over three training seeds; the inset panels resolve the settings at the counts where the curves crowd. Ten plain steps are the worst setting at every k on resolution, but guidance returns most of what they cost. On the high-band FSC, guided four-step sampling is the best setting from $k = 16$ upward.

Table 8: The high-band FSC on the held-out test split: the mean FSC over shells above 0.65 of Nyquist (Section 6), higher is better. Across-seed standard deviation is at most 0.008 and is not shown. RELION without CTF correction is anti-correlated with the truth past the first zero of the contrast transfer function (Appendix E).

	CryoFATHOM					Classical			
k	$N=1$	$N=4$	$N=10$	$N=4$ guided	$N=10$ guided	CG oracle inv.	RELION full CTF	RELION ph. flip	RELION no CTF
Mean FSC over shells above 0.65 *of Nyquist*									
4	0.025	0.021	0.016	0.024	0.021	0.045	0.010	0.043	−0.020
8	0.055	0.052	0.041	0.055	0.049	0.065	0.014	0.061	−0.030
16	0.093	0.092	0.074	0.096	0.086	0.092	0.022	0.085	−0.041
32	0.132	0.131	0.106	0.138	0.127	0.130	0.037	0.118	−0.057
64	0.163	0.167	0.138	0.173	0.162	0.186	0.064	0.160	−0.078
128	0.191	0.195	0.163	0.203	0.193	0.262	0.119	0.215	−0.104
256	0.206	0.211	0.180	0.220	0.210	0.361	0.231	0.290	−0.143
512	0.216	0.222	0.190	0.231	0.222	0.478	0.398	0.395	−0.200
1024	0.222	0.227	0.194	0.237	0.227	0.604	0.553	0.523	−0.273

G Guidance schedule and comparison to CFG++

This appendix defines the guidance recipe of Section 4.4 and its tuning protocol, compares it with CFG++ (Chung et al., 2025), and gives the amplitude calibration procedure.

G.1 The sampler and its tuning

Guidance is applied at inference time on a trained checkpoint, no model is trained differently because of it. One step of the guided solver, at step index i of N with $t = i/N$ is given by,

$$\boldsymbol{x}_{t+1/N} \;=\; \boldsymbol{x}_t \;+\; \frac{1}{N}\Big[\, v_\emptyset \;+\; W(i)\,\big(v_\theta(\boldsymbol{x}_t, t, S_c) - v_\emptyset\big)\Big], \qquad v_\emptyset \;=\; v_\theta(\boldsymbol{x}_t, t, \varnothing), \tag{14}$$

with $W(i) = w$ constant for our recipe, and a different W for CFG++ (Appendix G.2). The gate makes the second network evaluation conditional: when $t \geq t_{\mathrm{hi}}$ the step reduces to the plain conditional update and $v_\emptyset$ is never computed. Guidance is active on the first $\lceil t_{\mathrm{hi}} N \rceil$ steps and $\mathrm{NFE} = N + \lceil t_{\mathrm{hi}} N \rceil$, so a smaller gate is computationally cheaper.

In our experiments we first tried a learned null, training with the conditioning dropped on 10% of examples. That model reconstructed worse than one trained without conditioning dropout, and the guidance it enabled did not recover the difference. The null $\varnothing$ we use instead is the all-zero token set of Section 4.4: one projection's worth of tokens, M of them, all zero. It is supplied directly to the cross-attention rather than produced by the encoder, so neither the convolutional stem nor the set attention block runs on it. The cross-attention projections carry no bias, and the normalization maps zero to zero, so the output on all-zero tokens is identically zero and the block computes what it would with cross-attention removed. The number of such tokens supplied therefore cannot matter, hence a guided step costs well under twice a plain one at large k.

The constants w and t_{hi} are swept jointly for each solver-step count on the validation split, and the selected cell is applied unchanged when the model is evaluated on test. Table 9 reports the sweep. Arms are compared on resolution, subject to not degrading the high-band FSC: guidance is there to recover the fidelity that extra solver steps cost, so we do not take an arm that buys resolution by giving up high-frequency correlation. At $N = 1$ every strength degrades the reconstruction, monotonically in w, from -0.04% at $w = 1.25$ to -0.70% at $w = 4$, so the single step reconstruction is left unguided. At $N = 4$ the strength was chosen at $w = 3$ with the whole trajectory guided. For $N = 10$ the largest strength in our grid, $w = 8$, gives the best resolution, but it does so by trading away the high-band FSC: against $w = 6$ it is better on resolution by 0.16% and worse on high-band FSC by 0.78% where projections are scarce, on every seed. We therefore take $w = 6$, whose gate at $t_{\mathrm{hi}} = 0.5$ also halves the null evaluations.

Table 9: Guidance sweep on the validation split: strength w against gate t_{hi}, at each solver-step count. Entries are the percentage improvement in spherical FSC resolution at 0.5 over the plain rollout at the same N and the same $\boldsymbol{x}_0$, averaged over seven values of k, 75 molecules and three training seeds (positive is better). Selected cells in bold.

$N = 4$				
w	$t_{\mathrm{hi}} = 0.25$	0.50	0.75	1.00
1.5	+1.22	+1.18	+1.17	+1.18
2	+1.45	+1.53	+1.56	+1.59
2.5	+1.42	+1.65	+1.70	+1.71
3	+1.27	+1.64	+1.76	**+1.78**
3.5	+1.11	+1.55	+1.72	+1.80
4	+0.92	+1.46	+1.66	+1.75
5	+0.66	+1.28	+1.52	+1.64
$N = 10$				
w	$t_{\mathrm{hi}} = 0.20$	0.30	0.50	1.00
2	+3.90	+3.84	+3.79	+3.81
3	+4.65	+4.63	+4.57	+4.56
4	+4.94	+5.00	+5.04	+4.99
5	+5.09	+5.24	+5.32	+5.29
6	+5.17	+5.40	**+5.54**	+5.47
8	+5.22	+5.50	+5.69	+5.67

G.2 Comparison to CFG++

CFG++ (Chung et al., 2025) replaces the guided update by a denoise-and-renoise pair. We translate its DDIM update (their Algorithm 2) directly: the data endpoint is estimated from the guided velocity, the noise endpoint from the unconditional one, and the next state is their interpolation at the next time. With $\Delta v = v_\theta(\boldsymbol{x}_t, t, S_c) - v_\emptyset$, guided velocity $v_g = v_\emptyset + \lambda\,\Delta v$, and $s = t + \mathrm{d}t$,

$$\boldsymbol{x}_s \;=\; (1-s)\,\underbrace{\big(\boldsymbol{x}_t - t\,v_\emptyset\big)}_{\hat{\boldsymbol{x}}_0} \;+\; s\,\underbrace{\big(\boldsymbol{x}_t + (1-t)\,v_g\big)}_{\hat{\boldsymbol{x}}_1} \;=\; \boldsymbol{x}_t \;+\; \mathrm{d}t\left[\,v_\emptyset \;+\; \frac{\lambda\,s\,(1-t)}{\mathrm{d}t}\,\Delta v\right], \tag{15}$$

since the coefficient of $v_\emptyset$ collapses to $s - t = \mathrm{d}t$. This is Equation 14 under the weight

$$W_{\text{CFG++}}(i) \;=\; \lambda\,\frac{(i+1)(N-i)}{N}, \tag{16}$$

with $t = i/N$ and $\mathrm{d}t = 1/N$. In deterministic sampling, CFG++ is therefore plain CFG with this weight, and our implementation reproduces its path from Equation 16 to within 1.2×10^{-6}.

This weight has a bump shape: it equals λ at both ends, peaks at 1.5λ when $N = 4$ and 3λ when $N = 10$, and $W = 1$ is no guidance at all. Guidance can act only where the conditional and null fields differ, and in our model they differ most at the first step. Measured along the plain rollout at $N = 10$, averaged over seven values of k, the 75 validation molecules and three seeds, $\|\Delta v\|/\|v_\theta\|$ is 0.75 at $t = 0$, 0.05 at $t = 0.5$ and 0.06 at $t = 0.9$, while $\cos(v_\emptyset, v_\theta)$ rises from 0.66 to 0.999. The separation is largest at $t = 0$ in every trajectory we measured. CFG++ therefore puts most of its guidance where the two fields already agree.

Table 10 sweeps λ over $[0, 1]$, the range that CFG++ allows. Resolution falls away as λ decreases, by more than 30% at $\lambda = 0.25$. CFG++ bounds λ so that each data estimate stays between the unconditional and conditional ones, an argument that assumes a trained unconditional model. Our null is the same network with its conditioning pathway inactive, so we give that constraint up and guide with a larger weight. The result is specific to our setting and does not bear on CFG++ with a trained unconditional model, the case it was designed for.

G.3 Amplitude calibration

Guidance's excess is concentrated in the low shells. A single global factor would be dominated by it and would suppress the high-frequency shells, so the calibration works per shell instead. It

Table 10: CFG++ across $\lambda \in [0, 1]$, against the plain rollout at the same N and the same $\boldsymbol{x}_0$. The checkpoint, null, initial noise, calibration and the code path are shared, so the weight profile W is the only difference between arms. Entries are percentage improvement in spherical FSC resolution at 0.5 measured on the validation split (positive is better).

Arm	$N = 4$	$N = 10$
CFG++, $\lambda = 0.25$	-34.02	-31.99
CFG++, $\lambda = 0.5$	-6.91	-7.47
CFG++, $\lambda = 0.75$	-1.82	-1.60
CFG++, $\lambda = 1$	-0.17	$+0.20$
CFG, selected recipe	$\mathbf{+1.78}$	$\mathbf{+5.54}$

matches the guided volume's radial amplitude profile to that of its plain sibling. Let $\boldsymbol{x}_{\mathrm{g}}$ be the guided reconstruction and $\boldsymbol{x}_{\mathrm{p}}$ the plain rollout at the same N from the same $\boldsymbol{x}_0$ and the same conditioning set, let m be the spherical mask the evaluation applies, and let $P[\,\cdot\,](\kappa)$ be power averaged over the Fourier shell at radius κ, with $\kappa = 0$ the zero-frequency term. Then,

$$\boldsymbol{x}_{\mathrm{cal}} \;=\; \mathcal{F}^{-1}\Big[\, r(\kappa)\, \mathcal{F}\big[\, m \odot \boldsymbol{x}_{\mathrm{g}} \,\big] \Big], \qquad r(\kappa) \;=\; \sqrt{\frac{P\big[\, m \odot \boldsymbol{x}_{\mathrm{p}} \,\big](\kappa)}{P\big[\, m \odot \boldsymbol{x}_{\mathrm{g}} \,\big](\kappa)}}. \tag{17}$$

It reads no ground truth, fits no parameters and exposes no hyperparameter.

The Fourier shell correlation is a normalized per-shell inner product, so scaling shell κ by $r(\kappa)$ multiplies its numerator and denominator equally, and every reported resolution is identical with and without it. It exists so that displayed volumes are comparable across arms.

H Ablations

Each arm differs from the model of Section 4 on exactly one axis. The guidance recipe was tuned on the full model, so the arms are compared on the plain single-step sampler. The conclusions do not depend on which sampler is used.

Withholding the orientations removes both geometry terms of Equation 5: the frame modulation (α_i, β_i) and the ray offset o_{ia} are not computed. The slot embedding ζ_a remains, so a token still identifies which patch of its projection it came from. Coarser tokenization enlarges the patch embedding quadratically at identical patchify cost, so $M = 1$ carries more than twice the parameters of the full model. Between $M = 16$ and $M = 4$ the gap is 1.5Å at $k = 4$ and 0.6Å at $k = 16$ on all three seeds; from $k \approx 50$ upward the two are indistinguishable, with $M = 4$ marginally but never significantly ahead (Figure 9).

Table 11: Withholding the orientations. Spherical FSC resolution at 0.5 (Å, lower is better) on the held-out test split, mean $\pm$ standard deviation over three training seeds.

	With orientations (ours)			Without orientations		
k	N=1	N=4 g.	N=10 g.	N=1	N=4 g.	N=10 g.
4	44.4 ±0.6	44.9 ±0.6	45.5 ±0.3	66.3 ±0.6	66.2 ±0.8	66.5 ±0.7
8	35.5 ±0.4	36.0 ±0.4	36.6 ±0.3	64.8 ±0.6	65.2 ±0.4	65.7 ±0.2
16	29.0 ±0.4	29.4 ±0.4	29.9 ±0.3	64.6 ±0.5	64.8 ±0.3	65.3 ±0.5
32	25.8 ±0.2	26.0 ±0.2	26.5 ±0.2	64.2 ±0.5	64.5 ±0.7	65.1 ±0.5
64	24.0 ±0.2	24.1 ±0.2	24.5 ±0.2	63.3 ±0.8	63.7 ±0.8	64.6 ±0.3
128	23.0 ±0.2	22.9 ±0.2	23.2 ±0.2	63.4 ±0.5	63.8 ±0.5	64.3 ±0.3
256	22.4 ±0.1	22.2 ±0.1	22.6 ±0.2	63.3 ±0.8	63.6 ±0.7	64.2 ±0.4
512	22.0 ±0.1	21.9 ±0.1	22.3 ±0.2	63.0 ±0.6	63.2 ±0.4	64.3 ±0.3
1024	21.8 ±0.0	21.7 ±0.1	22.1 ±0.2	62.9 ±0.5	63.4 ±0.3	64.1 ±0.6

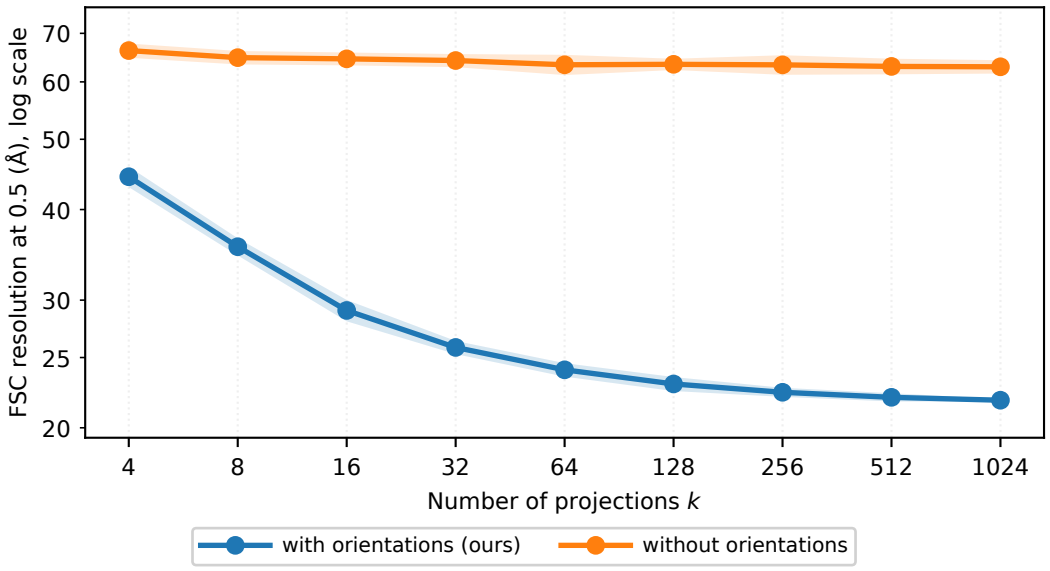


Figure 7: Withholding the orientations, bands over three seeds.

Table 12: Tokens per projection. Spherical FSC resolution at 0.5 (Å, lower is better) on the held-out test split, mean ± standard deviation over three training seeds. The single-token arm's spread across seeds is 20 to 31Å: it does not converge to a consistent solution.

	$M = 16$ (ours)			$M = 4$			$M = 1$		
k	N=1	N=4 g.	N=10 g.	N=1	N=4 g.	N=10 g.	N=1	N=4 g.	N=10 g.
4	44.4 ±0.6	44.9 ±0.6	45.5 ±0.3	45.9 ±0.3	46.3 ±0.4	46.6 ±0.4	93.6 ±19.6	114.1 ±30.0	115.1 ±30.0
8	35.5 ±0.4	36.0 ±0.4	36.6 ±0.3	36.4 ±0.5	36.9 ±0.5	37.3 ±0.5	83.8 ±24.5	89.6 ±28.7	91.2 ±28.8
16	29.0 ±0.4	29.4 ±0.4	29.9 ±0.3	29.7 ±0.5	29.9 ±0.5	30.4 ±0.5	80.0 ±27.8	83.2 ±30.5	84.4 ±30.6
32	25.8 ±0.2	26.0 ±0.2	26.5 ±0.2	25.9 ±0.5	26.1 ±0.5	26.6 ±0.6	78.2 ±30.0	82.0 ±31.4	82.0 ±31.4
64	24.0 ±0.2	24.1 ±0.2	24.5 ±0.2	24.0 ±0.4	24.1 ±0.4	24.3 ±0.4	75.8 ±30.2	78.8 ±31.6	79.6 ±31.6
128	23.0 ±0.2	22.9 ±0.2	23.2 ±0.2	22.8 ±0.4	22.9 ±0.4	23.2 ±0.4	76.0 ±31.3	78.8 ±31.7	79.2 ±31.9
256	22.4 ±0.1	22.2 ±0.1	22.6 ±0.2	22.2 ±0.4	22.1 ±0.4	22.5 ±0.4	75.7 ±31.8	78.0 ±32.0	79.1 ±32.2
512	22.0 ±0.1	21.9 ±0.1	22.3 ±0.2	21.9 ±0.4	21.8 ±0.4	22.1 ±0.4	74.4 ±31.0	77.3 ±31.5	77.7 ±31.5
1024	21.8 ±0.0	21.7 ±0.1	22.1 ±0.2	21.8 ±0.4	21.6 ±0.4	21.9 ±0.4	74.6 ±30.9	77.4 ±31.0	77.6 ±31.0

I Training compute accounting

All models reported here were trained on NVIDIA L40 accelerators with 48 GB of memory, one GPU per run, on a server with two AMD EPYC 9654 processors. At the training budget of Table 6 a run of the full model takes about 30 hours, and the three seeds behind every entry were trained concurrently on separate GPUs. The ablation arms take about 24 hours per run at $M = 4$ and $M = 1$, and about 30 hours without orientations. The twelve training runs behind our results therefore come to roughly 330 GPU-hours. Development and model selection across the 32^3 and 64^3 rungs consumed about ten times that.

J Additional qualitative reconstructions

Figures 11–13 show ten test molecules chosen for qualitative examination, at three measurement counts. Within a row every panel is built from the same measurements and the same noise realization. The RELION column uses phase flipping since it was best at these values of k (Appendix E).

The methods do not share an amplitude scale, so one density level would not mean the same thing in every column. A molecule's deposited molecular weight fixes how much space its density should

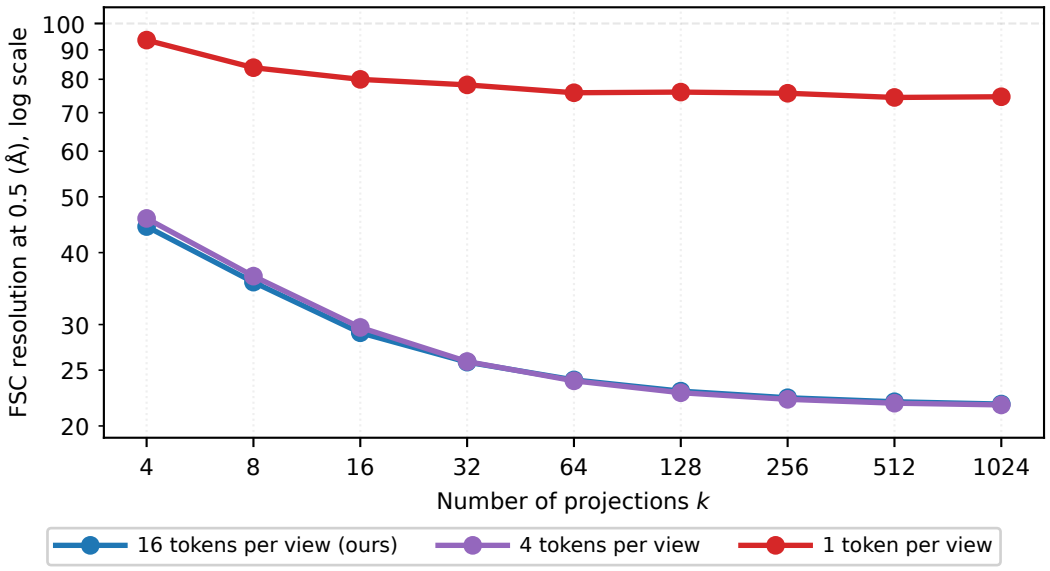


Figure 8: Tokens per projection at $N = 1$, on a logarithmic axis. One token per projection lands far from the other two, which are close enough that Figure 9 separates them.

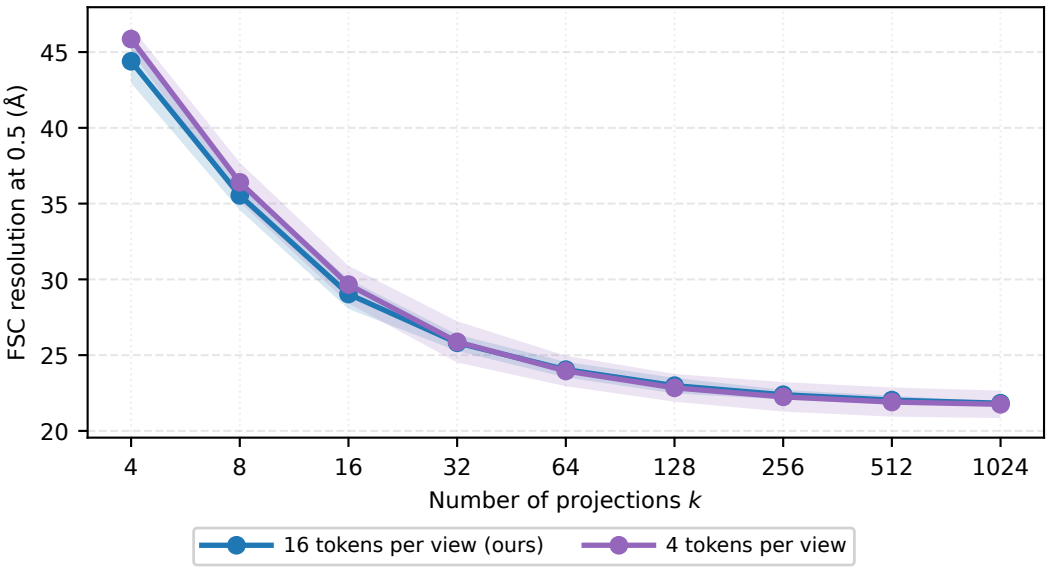


Figure 9: The same comparison restricted to $M = 16$ and $M = 4$, on a linear axis with bands over three seeds. The two separate only at the smallest counts and converge as measurements accumulate.

fill, and dividing by the voxel volume turns that into a count n. Deposited maps are checked on the same correspondence between contour level, enclosed volume and molecular weight (Kleywegt et al., 2024). Every panel of a row, the ground truth included, is drawn at the level that puts the n densest voxels of its own map inside the surface. Scaling a map up or down does not change which voxels those are, so no per-method correction is applied. Every column therefore shows the same amount of material. The panels of a row are rendered in ChimeraX (Pettersen et al., 2021), through one camera set on the ground truth. Small disconnected fragments are a known artifact of RELION reconstructions which we did not eliminate.

Every volume is band-limited by Fourier cropping (Appendix C), and sharp truncation leaves sinc side lobes around the density. In the smallest and thinnest molecules the enclosed-volume level falls

close to the lobes' own amplitude, so the surface encloses them and they render as speckle, as in 7UL4 and 6LHS. Reconstructions share that band limit, so the ringing sits on both sides of every comparison, not in one method.

Read as a series, the three sheets show the classical reconstructions narrowing the gap our prior opens at small k. At $k = 8$ the classical methods output highly-fragmented reconstructions while CryoFATHOM usually outputs a more sensible reconstruction with a single connected component that matches the ground truth. By $k = 16$ the classical methods recover the gross shape, and by $k = 32$ the conjugate-gradient column is coherent throughout and smoother than the guided samples. What separates the columns has become surface detail rather than form. The guided columns carry more surface detail than the single step at every count, and Appendix F reports how much of it is actually correlated with the truth.

Figure 10 shows the three molecules of Figure 1 in the plane of the measurements: one image of the conditioning set, the clean projection of the ground truth, and the projection of our reconstruction.

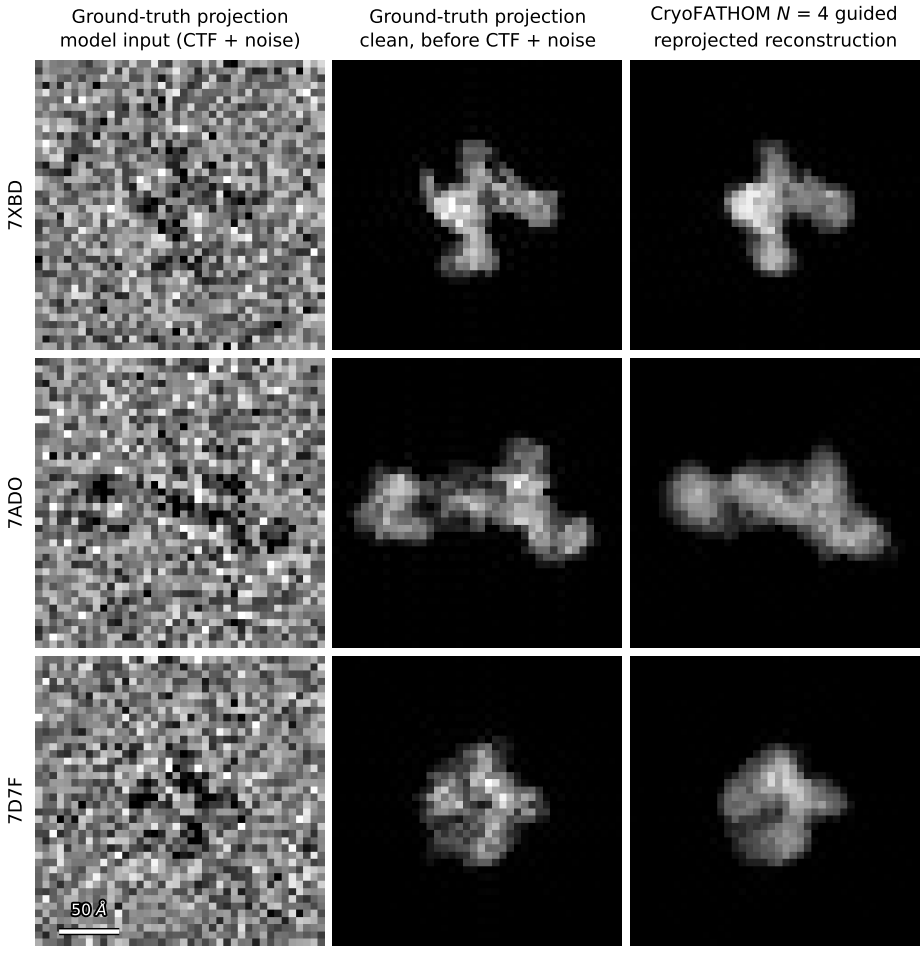


Figure 10: The three molecules of Figure 1 at $k = 16$. The left column shows the projection images with CTF and noise that are the inputs to the reconstruction methods. The middle column shows the same images without CTF and noise. The right column contains projections of reconstructed volumes using CryoFATHOM with $N = 4$ and guidance that Figure 1 renders as a surface. Panels are cropped to the central 240Å of the 384Å box; the scale bar is 50Å.

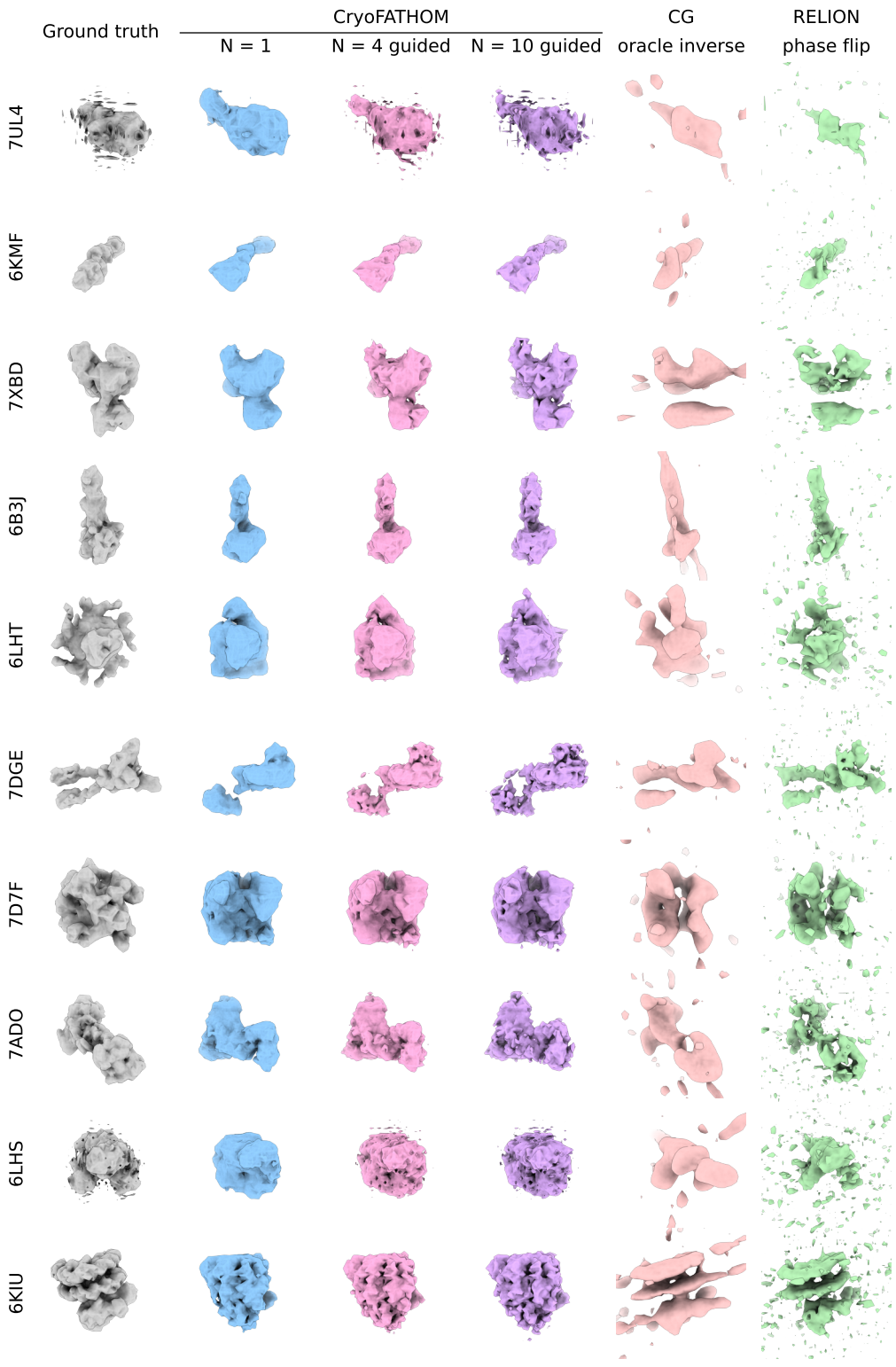


Figure 11: Reconstructed density surfaces from $k = 8$ projections. At this count the classical reconstructions are broken into fragments while CryoFATHOM successfully recovers the gross shape.

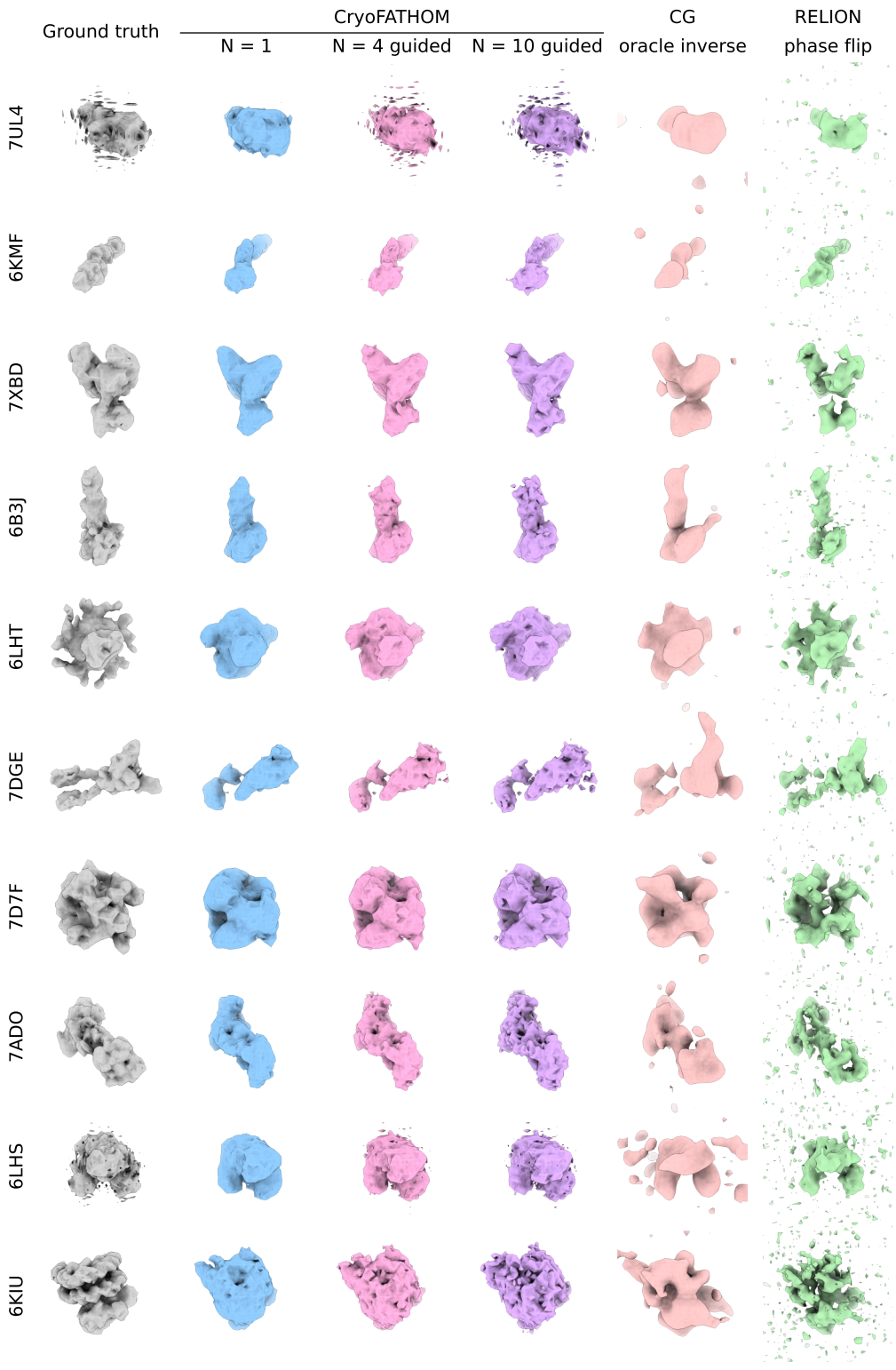


Figure 12: The same ten molecules at $k = 16$, under the same conventions. Rows 7XBD, 7ADO and 7D7F are the three shown in Figure 1.

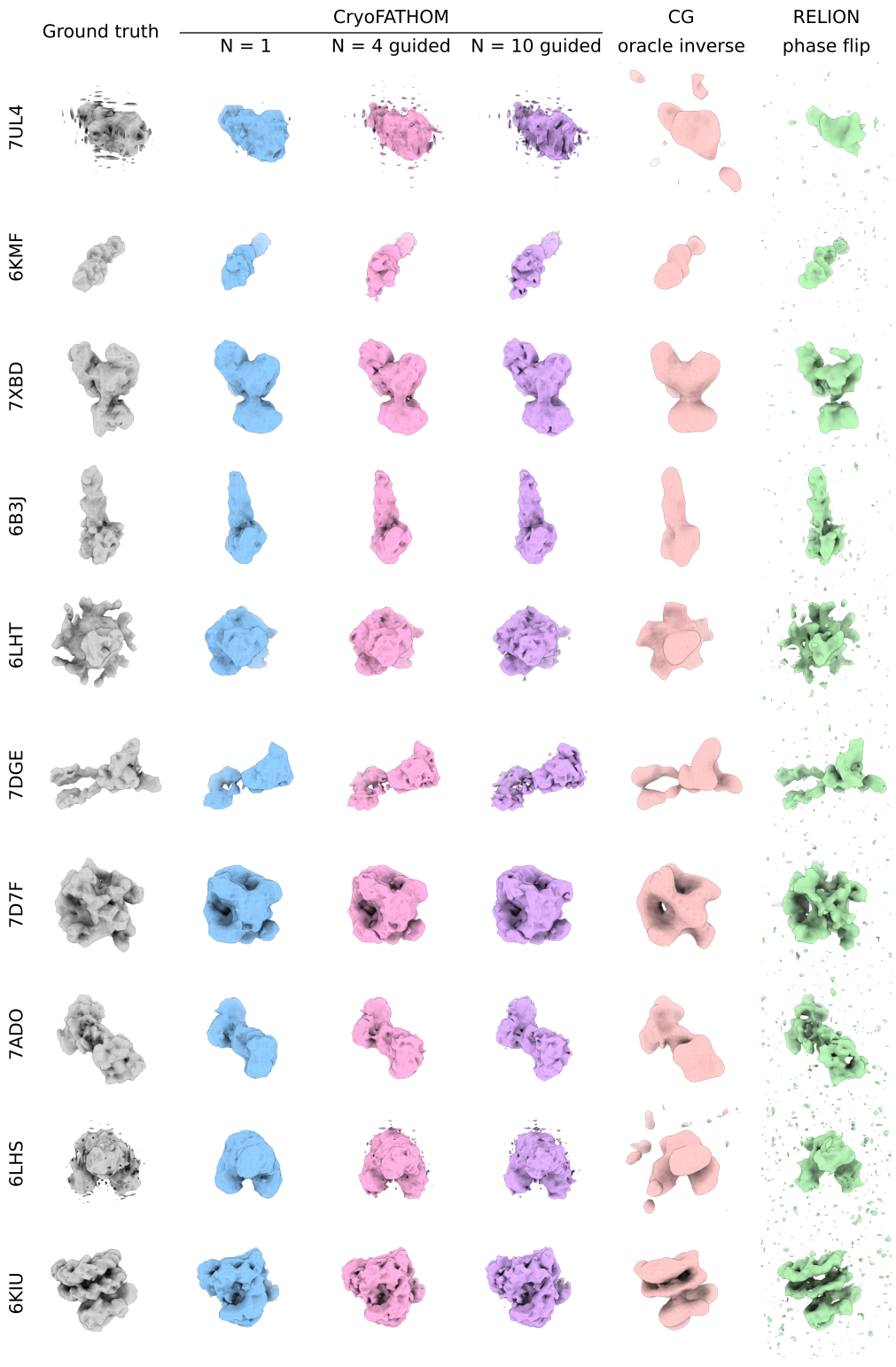


Figure 13: The same ten molecules at $k = 32$. The conjugate-gradient column has become coherent and remains smoother than the guided samples.