
From Unlabeled Micrographs to 3D Structure: Self-Supervised Reconstruction with Score Models

Julian L. Möbius
Friedrich Schiller University

Michael Habeck
Friedrich Schiller University

Abstract

Single-particle cryo-electron microscopy (cryo-EM) aims to recover the 3D structure of biological macromolecules from large, noisy 2D micrographs. Each micrograph contains multiple particle projections at unknown positions and orientations, corrupted by the contrast transfer function and high noise levels. Standard cryo-EM pipelines rely on particle picking to extract candidate particles from micrographs, followed by a sequence of processing stages for alignment, classification, and 3D reconstruction. We introduce a self-supervised framework that removes the need for explicit particle picking by learning the distribution of random micrograph crops via denoising score matching. The learned score function is first used for fully self-supervised denoising through a Noisier2Noise estimator, producing noise-reduced micrographs without clean supervision. The score model is then fine-tuned on crops from these denoised micrographs to better capture underlying structural variability. At inference time, a user specifies a region of interest in a denoised micrograph, and a 3D volume is reconstructed using a score-distillation-based procedure. This approach replaces manual or heuristic particle picking in noisy or contrast-enhanced micrographs with a learned generative representation. We validate the method on two challenging simulated datasets and show that it can handle both conformational and compositional heterogeneity.

1 Introduction

In single-particle cryo-EM, many copies of a biological macromolecule or molecular assembly are frozen in vitreous ice and imaged with an electron microscope [3, 19]. Each micrograph is a large, extremely noisy 2D image containing on the order of $10^2 - 10^3$ particle instances at unknown locations and orientations. Image formation is further degraded by the microscope point spread function (PSF), or equivalently by the contrast transfer function (CTF) in Fourier space. Each particle instance contributes a noisy projection of an underlying 3D density. Let $\mathcal{X}_{\text{raw}}^m$ denote the m -th micrograph in a dataset of M micrographs. Each particle contributing to $\mathcal{X}_{\text{raw}}^m$ is associated with a latent volume $\mathbf{V}_i$, which may represent a distinct molecular conformation or assembly state, an unknown in-plane position, an unknown 3D rotation $\mathcal{R}_i$, and noise realization $\mathbf{N}_i$. For a local image region containing a particle, we assume the forward model

$$[\text{PSF}_m * \mathbf{f}(\mathbf{V}_i, \mathcal{R}_i)] + \mathbf{N}_i, \quad \mathbf{f}(\mathbf{V}_i, \mathcal{R}_i) = \Pi \circ \mathcal{R}_i \circ \mathbf{V}_i \quad (1)$$

where Π denotes orthogonal projection onto the imaging plane, so that $\mathbf{f}(\mathbf{V}_i, \mathcal{R}_i)$ is the clean 2D projection induced by rotating $\mathbf{V}_i$ by $\mathcal{R}_i$. The projected image is then convolved with the micrograph-specific PSF and corrupted by additive noise, typically approximated as

white Gaussian noise. In heterogeneous datasets, the latent volumes $\mathbf{V}_i$ may vary across particles, reflecting conformational or compositional heterogeneity. The goal is to recover the underlying 3D molecular density map or maps from the observed micrographs.

Existing reconstruction software does not operate directly on raw micrographs. Instead, it first constructs a particle stack by detecting candidate particles, approximately centering and boxing them, and often filtering them through class averaging. Widely used software packages such as RELION [23] and cryoSPARC [22] support this preprocessing pipeline as well as subsequent homogeneous *ab initio* reconstruction. In the homogeneous setting, all particle images are assumed to originate from a single underlying 3D structure, i.e., $\mathbf{V}_i \equiv \mathbf{V}$, making reconstruction substantially more tractable. Heterogeneous *ab initio* reconstruction is considerably more challenging, as particle pose and structural state must be inferred jointly while recovering a distribution of conformations. Existing methods differ primarily in whether particle poses are assumed to be known. CryoDRGN [32, 33], 3DFlex [21], and OPUS-DSD [15] operate in a refinement regime, where structural variability is estimated from particle images together with poses obtained from a prior consensus reconstruction. In contrast, CryoDRGN2 [34], cryoFIRE [13], and CryoDRGN-AI [14] tackle fully *ab initio* heterogeneous reconstruction by jointly estimating pose, conformation, and structure. However, all of these methods still rely on picked particles as input. Appendix A reviews additional related work on denoising and reconstruction in cryo-EM.

We propose an alternative framework for both homogeneous and heterogeneous *ab initio* reconstruction that avoids particle-picking and pose-estimation heuristics and instead performs the relevant data extraction in a fully self-supervised manner. Specifically, we learn the distribution p_{raw} of local image regions from random micrograph crops $\mathbf{x}_{\text{raw}} \sim p_{\text{raw}}$ using a score-based model in the spirit of diffusion models [25]. This learned score representation enables self-supervised denoising and PSF correction, yielding a distribution of clean micrographs $\mathcal{X}_{\text{clean}}$, allowing it to capture the underlying structure of noise-free, PSF-corrected observations more accurately. At inference time, the final score model supports selecting a particle projection of interest directly in the clean micrographs and lifting the corresponding image content into 3D via score-distillation [20]. As a result, particle picking is no longer required to build the training dataset and instead becomes a sparse, user-driven query mechanism for individual reconstructions. Figure 1 provides an overview.

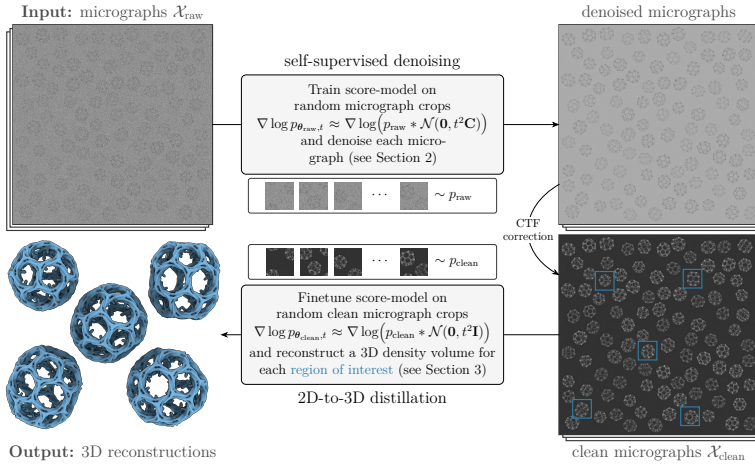


Figure 1: Overview of our 2D-to-3D reconstruction pipeline directly from raw micrographs without classical particle picking. All shown images and 3D volumes are real results from the simulated micrograph dataset containing a composition of four different architectures of clathrin cages.

2 Self-Supervised Micrograph Denoising

In the denoising and CTF correction step, our goal is to remove the noise component $\mathbf{N}_m \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{C})$ from each micrograph $\mathcal{X}_{\text{raw}}^m := \text{PSF}_m * \mathcal{X}_{\text{clean}}^m + \mathbf{N}_m$ to obtain $\text{PSF}_m * \mathcal{X}_{\text{clean}}^m$. For this we train a *denoiser* $\mathbf{D}_{\theta_{\text{raw}}}$ [8] on the noisy micrograph crops to estimate the score of the mollified target distribution

$$\nabla \log p_{\text{raw},t} := \nabla \log (p_{\text{raw}} * \mathcal{N}(\mathbf{0}, t^2 \mathbf{C})) \approx 1/t^2 \mathbf{C}^{-1} (\mathbf{D}_{\theta_{\text{raw}}}(\mathbf{x}, t) - \mathbf{x}) =: \nabla \log p_{\theta_{\text{raw}},t}. \quad (2)$$

The target p_{raw} is induced by uniformly sampling each crop $\mathbf{x}_{\text{raw}} \in \mathbb{R}^{c \times c}$ from $\mathcal{X}_{\text{raw}}$ and augmenting it with multiple-of-90° rotations and flips (arbitrary in-plane rotations on noisy data would induce critical interpolation errors). The crop size c is chosen to be around 120 - 200% of the expected particle size. We follow Karras et al. [8] and perform **denoising score matching** [25, 28] to learn a representation of p_{raw} :

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{x}_{\text{raw}} \sim p_{\text{raw}}, t, \boldsymbol{\epsilon}} \left[\lambda(t) \|\mathbf{D}_{\boldsymbol{\theta}}(\mathbf{x}_{\text{raw}} + t\boldsymbol{\epsilon}, t) - \mathbf{x}_{\text{raw}}\|_2^2 \right], \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}). \quad (3)$$

Here, $\mathbf{C}$ is learned from background patches of the data and characterizes the noise distribution of the micrographs. The first training uses p_{raw} and data-like colored perturbations (Appendix B explains further details). **Noisier2Noise** then restores complete micrographs without clean targets [17]. At the estimated acquisition-noise scale σ , Noisier2Noise forms $\mathbf{z} = \mathcal{X}_{\text{raw}}^m + \alpha\sigma\boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$ and estimates the noise free micrograph as

$$\text{PSF}_m * \mathcal{X}_{\text{clean}}^m \approx ((1 + \alpha^2)D_{\boldsymbol{\theta}_{\text{raw}}}(\mathbf{z}, \alpha\sigma) - \mathbf{z})/\alpha^2. \quad (4)$$

In our experiments, $\alpha = 1/4$. Because we choose $D_{\boldsymbol{\theta}_{\text{raw}}}$ to be a fully convolutional UNet (see Appendix B for more details on the UNet architecture), Eq. (4) can be applied to full micrographs even though the *denoiser* was trained on small crops. To approximately remove the effect of the PSF, we apply a Wiener filter (with $\lambda = 0.005$ in our experiments):

$$\mathcal{X}_{\text{clean}}^m \approx \mathcal{F}^{-1} \left\{ \frac{\text{CTF}_m}{\text{CTF}_m^2 + \lambda} \mathcal{F}\{\text{PSF}_m * \mathcal{X}_{\text{clean}}^m\} \right\}, \quad \text{with } \text{CTF}_m := \mathcal{F}\{\text{PSF}_m\}. \quad (5)$$

Denoising and PSF correction results in the set of all M estimated clean micrographs $\mathcal{X}_{\text{clean}}$ and its corresponding clean crop distribution p_{clean} , obtained by uniformly sampling $\mathbf{x}_{\text{clean}} \in \mathbb{R}^{c \times c}$ and applying arbitrary in-plane rotations and flips. Denoising examples can be seen in Fig. 1, 3 and 5.

3 Self-Supervised 2D-to-3D Reconstruction

We further train the UNet-based *denoiser* using denoising score matching replacing p_{raw} with p_{clean} in Eq. (3). After training on p_{clean} , the user can select a region in one of the denoised micrographs $\mathcal{X}_{\text{clean}}$ that shows a particle projection of interest. This single cropped particle is called the anchor $\mathbf{x}_a$ and will be unfolded into a 3D volume $\mathbf{V}_a$. For each anchor, an algorithm based on **score-distillation** [20] optimizes a volume whose random projections follow p_{clean} while its observed view agrees with $\mathbf{x}_a$. Different anchors can therefore query different conformations from the shared micrograph representation. We fix the clean anchor’s orientation to the identity and minimize

$$\mathcal{J}(\mathbf{V}_a) = w_s \mathbb{E}_{\mathbf{R}} [t^2 \text{D}_{\text{KL}}(\mathcal{N}(\mathbf{f}(\mathbf{V}_a, \mathbf{R}), t^2 \mathbf{I}) \| p_{\boldsymbol{\theta}_{\text{clean}}, t})] + \frac{w_a}{2} \|\mathbf{f}(\mathbf{V}_a, \mathbf{I}_3) - \mathbf{x}_a\|_2^2. \quad (6)$$

Here $p_{\boldsymbol{\theta}_{\text{clean}}, t}$ denotes the estimate of the clean micrograph density induced by the *denoiser* $\mathbf{D}_{\boldsymbol{\theta}_{\text{clean}}}$ (see Eq. 2) that was trained on p_{clean} . The first term minimizes the expected Kullback-Leibler divergence between the mollified distribution of current volume projections and the distribution of clean micrograph crops. Since the reverse KL-divergence is mode seeking, the current volume projections do not need to match the full distribution but only a mode, which perfectly fits into the heterogeneous setup where a single conformation only contributes to a subset of all particle projections. Writing $\mathbf{J}(\mathbf{V}_a, \mathbf{R}) = \partial \mathbf{f}(\mathbf{V}_a, \mathbf{R}) / \partial \mathbf{V}_a$, the voxel update follows

$$\begin{aligned} \nabla_{\mathbf{V}_a} \mathcal{J}(\mathbf{V}_a) = & w_s \mathbb{E}_{\mathbf{R}, \boldsymbol{\epsilon}} [\mathbf{J}(\mathbf{V}_a, \mathbf{R})^\top (\mathbf{f}(\mathbf{V}_a, \mathbf{R}) - D_{\boldsymbol{\theta}_{\text{clean}}}(\mathbf{f}(\mathbf{V}_a, \mathbf{R}) + t\boldsymbol{\epsilon}, t))] \\ & + w_a \mathbf{J}(\mathbf{V}_a, \mathbf{I}_3)^\top (\mathbf{f}(\mathbf{V}_a, \mathbf{I}_3) - \mathbf{x}_a), \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \end{aligned} \quad (7)$$

The first line is the score-distillation-like gradient; the second is the clean-anchor gradient (Appendix C details a full derivation of $\nabla_{\mathbf{V}_a} \mathcal{J}(\mathbf{V}_a)$). During reconstruction, stochastic gradient descent only updates $\mathbf{V}_a$, thus the parameters of the score model $\boldsymbol{\theta}_{\text{clean}}$ remain fixed and the expectation is approximated by a Monte Carlo estimate (see Algorithm 1 in Appendix C).

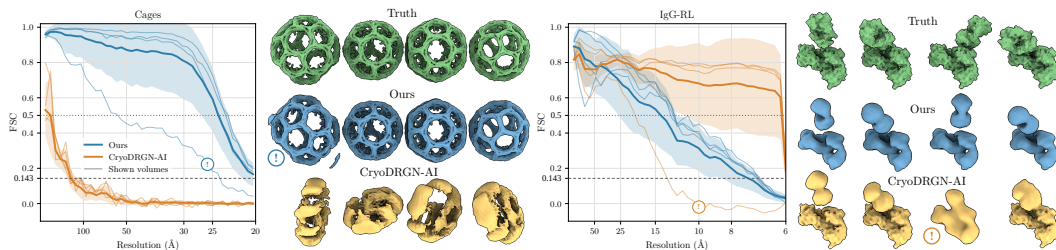


Figure 2: Mean FSC curves and representative reconstructions from the cages and IgG-RL dataset. Shaded regions show one standard deviation, and thin curves correspond to the displayed volumes. Circled exclamation marks indicate representative failure cases.

4 Results

Direct 3D reconstruction from raw micrographs departs fundamentally from the standard cryo-EM pipeline, and there is currently no established benchmark for this setting. For heterogeneous reconstruction, the most informative metric is per-image FSC from CryoBench [7], but it is only defined for synthetic data with known ground-truth conformations. We therefore evaluate our method on simulated datasets and present proof-of-concept results.

We simulated two heterogeneous datasets using ASPIRE [31]. The first consists of 200 micrographs containing random mixtures of four clathrin cage architectures (EMD-0114, EMD-0115, EMD-0116, EMD-0118), which differ in symmetry, size, and pentagon-hexagon composition. Such mixed cage populations commonly arise in structural biology [18]; see also EMPIAR-10295. The micrographs are 1024×1024 at $10 \text{ \AA}/\text{pixel}$ and contain 20,000 particles corrupted by additive white noise. The second dataset contains 250 simulated micrographs of size 1024×1024 at $3 \text{ \AA}/\text{pixel}$, comprising 100,000 projections uniformly sampled from 100 conformations of CryoBench’s IgG-RL benchmark [7] and embedded in colored noise. Appendix D provides the PSF/CTF parameters, signal-to-noise ratios, and additional results on a real dataset based on $3\times$ downsampled micrographs from the RELION-5.0 tutorial for the deposited β -galactosidase images in EMPIAR-10204 (Appendix D.1).

We compare our method to CryoDRGN-AI [7, 14], the state-of-the-art approach for heterogeneous *ab initio* reconstruction. A fully fair comparison is challenging because existing cryo-EM reconstruction methods require picked particles, whereas our method operates directly on raw micrographs. For the simulated datasets, ASPIRE provides perfectly centered, tightly boxed ground-truth particle images and corresponding CTF parameters, which we use as input to CryoDRGN-AI. In contrast, our method uses only the raw micrographs and CTF information. For each dataset, we randomly select 100 particle images and reconstruct the corresponding volumes. In our method, these images serve as anchor views. We then compare each reconstruction to the ground-truth conformation that generated the selected particle by aligning the two volumes and computing the Fourier shell correlation (FSC). We report the mean FSC over the same 100 samples for both methods, together with the first non-cherry-picked reconstructions from the particle stack in Fig. 2.

5 Discussion and Conclusion

We propose a different route to cryo-EM reconstruction: first learn a score model of noisy, unlabeled micrographs, then use this representation to recover structure directly from the data. Our experiments provide initial evidence that this strategy is viable for challenging reconstruction problems. On the clathrin-cage dataset, our method recovers detailed 3D structure directly from raw micrographs, whereas CryoDRGN-AI fails despite operating on the easier input of perfectly picked particles. On IgG-RL, however, CryoDRGN-AI performs better, likely because it captures the static core well, while neither method fully resolves the flexible component. We discuss limitations and possible improvements in Appendix E. Overall, these results suggest score-based modeling of raw micrographs as a promising direction for *ab initio* heterogeneous cryo-EM reconstruction.

Acknowledgments and Disclosure of Funding

This work was supported by the Carl Zeiss Stiftung within the project “Interactive Inference”.

References

- [1] Tamir Bendory, Nicolas Boumal, William Leeb, Eitan Levin, and Amit Singer. Toward single particle reconstruction without particle picking: Breaking the detection limit. *SIAM Journal on Imaging Sciences*, 16(2):886–910, 2023.
- [2] Tristan Bepler, Kotaro Kelley, Alex J Noble, and Bonnie Berger. Topaz-Denoise: general deep denoising models for cryoEM and cryoET. *Nature communications*, 11(1):5208, 2020.
- [3] Joachim Frank. *Three-Dimensional Electron Microscopy of Macromolecular Assemblies: Visualization of Biological Molecules in Their Native State*. Oxford University Press, 2006.
- [4] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.
- [5] Harshit Gupta, Thong H Phan, Jaejun Yoo, and Michael Unser. Multi-CryoGAN: Reconstruction of continuous conformations in Cryo-EM using generative adversarial networks. In *European Conference on Computer Vision*, pages 429–444. Springer, 2020.
- [6] Harshit Gupta, Michael T McCann, Laurene Donati, and Michael Unser. CryoGAN: A new reconstruction paradigm for single-particle cryo-EM via deep adversarial learning. *IEEE Transactions on Computational Imaging*, 7:759–774, 2021.
- [7] 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. *Advances in Neural Information Processing Systems*, 37:89468–89512, 2024.
- [8] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.
- [9] Kwanyoung Kim and Jong Chul Ye. Noise2score: tweedie’s approach to self-supervised image denoising without clean images. *Advances in Neural Information Processing Systems*, 34:864–874, 2021.
- [10] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [11] Alexander Krull, Tim-Oliver Buchholz, and Florian Jug. Noise2void-learning denoising from single noisy images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2129–2137, 2019.
- [12] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2noise: Learning image restoration without clean data. *arXiv preprint arXiv:1803.04189*, 2018.
- [13] Axel Levy, Gordon Wetzstein, Julien NP Martel, Frederic Poitevin, and Ellen Zhong. Amortized inference for heterogeneous reconstruction in cryo-em. *Advances in neural information processing systems*, 35:13038–13049, 2022.
- [14] Axel Levy, Rishwanth Raghu, J Ryan Feathers, Michal Grzadkowski, Frédéric Poitevin, Jake D Johnston, Francesca Vallese, Oliver Biggs Clarke, Gordon Wetzstein, and Ellen D Zhong. CryoDRGN-AI: neural ab initio reconstruction of challenging cryo-EM and cryo-ET datasets. *Nature Methods*, 22(7):1486–1494, 2025.

- [15] Zhenwei Luo, Fengyun Ni, Qinghua Wang, and Jianpeng Ma. OPUS-DSD: deep structural disentanglement for cryo-EM single-particle analysis. *Nature Methods*, 20(11):1729–1738, 2023.
- [16] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.
- [17] Nick Moran, Dan Schmidt, Yu Zhong, and Patrick Coady. Noisier2noise: Learning to denoise from unpaired noisy data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12064–12072, 2020.
- [18] Kyle L Morris, Joseph R Jones, Mary Halebian, Shenping Wu, Michael Baker, Jean-Paul Armache, Amaury Avila Ibarra, Richard B Sessions, Alexander D Cameron, Yifan Cheng, et al. Cryo-em of multiple cage architectures reveals a universal mode of clathrin self-assembly. *Nature structural & molecular biology*, 26(10):890–898, 2019.
- [19] Eva Nogales and Sjors H. W. Scheres. Cryo-EM: A unique tool for the visualization of macromolecular complexity. *Molecular Cell*, 58(4):677–689, 2015.
- [20] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.
- [21] Ali Punjani and David J Fleet. 3DFlex: determining structure and motion of flexible proteins from cryo-EM. *Nature Methods*, 20(6):860–870, 2023.
- [22] 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.
- [23] Sjors HW Scheres. Relion: implementation of a bayesian approach to cryo-em structure determination. *Journal of structural biology*, 180(3):519–530, 2012.
- [24] Shakarim Soltanayev and Se Young Chun. Training deep learning based denoisers without ground truth data. *Advances in neural information processing systems*, 31, 2018.
- [25] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.
- [26] Structura Biotechnology Inc. Job: Micrograph Denoiser (BETA). <https://guide.cryosparc.com/processing-data/all-job-types-in-cryosparc/exposure-curation/job-micrograph-denoiser-beta>, 2026. cryoSPARC Guide. Accessed: 2026-03-03.
- [27] Dmitry Tegunov and Patrick Cramer. Real-time cryo-electron microscopy data preprocessing with warp. *Nature methods*, 16(11):1146–1152, 2019.
- [28] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.
- [29] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, Steven Liu, William Berman, Yiyi Xu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. URL <https://github.com/huggingface/diffusers>.
- [30] Thorsten Wagner and Stefan Raunser. The evolution of SPHIRE-crYOLO particle picking and its application in automated cryo-EM processing workflows. *Communications biology*, 3(1):61, 2020.
- [31] Garrett Wright, Joakim Andén, Vineet Bansal, Junchao Xia, Chris Langfield, Josh Carmichael, Kris Sowattanangkul, Robbie Brook, Yunpeng Shi, Ayelet Heimowitz, Gabi Pragier, Itay Sason, Amit Moscovich, Yoel Shkolnisky, and Amit Singer. Computationalcryoem/aspire-python: v0.14.2, December 2025.

- [32] 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*, 2020.
- [33] Ellen D Zhong, Tristan Bepler, Bonnie Berger, and Joseph H Davis. CryoDRGN: reconstruction of heterogeneous cryo-EM structures using neural networks. *Nature methods*, 18(2):176–185, 2021.
- [34] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4066–4075, 2021.

A Related work

A.1 Heterogeneous *ab initio* reconstruction in Cryo-EM

Approaches to cryo-EM heterogeneity commonly use latent-variable generative models to represent a family of structures rather than a single density. In VAE-based formulations [10], an encoder maps each particle image to a latent conformational coordinate and a decoder maps that coordinate to a 3D density. These methods differ importantly in the information they require. CryoDRGN [32, 33], 3DFlex [21], and OPUS-DSD [15] operate in a refinement setting: in addition to boxed particle images and CTF parameters, they use poses obtained from an earlier homogeneous consensus reconstruction. They therefore model structural variability after major pose estimation has already been performed.

CryoDRGN2 [34], cryoFIRE [13] and CryoDRGN-AI [14] address the more complex fully *ab initio* setting, in which pose and conformation are inferred jointly from the particle stack. CryoDRGN2 extends the first version of CryoDRGN to the *ab initio* setting, cryoFIRE uses an amortized framework, and CryoDRGN-AI is the most recent version of the fleet for processing challenging heterogeneous data, combining ideas from the former methods.

Our pose estimation free approach is akin to other distribution-matching approaches such as CryoGAN [6] and Multi-CryoGAN [5] which also avoid assigning a pose to every observation by matching generated and observed projection distributions with a GAN [4].

However, all of the methods take extracted particle images as their basic input. Our method instead learns the projection statistics from random micrograph crops and uses only a small number of selected anchors to request particular reconstructions.

Bendory et al. [1] likewise seek to reconstruct 3D volumes without particle picking. Their method estimates low-order autocorrelations directly from entire micrographs and recovers the underlying volume by solving a nonlinear inverse problem. However, their formulation neglects CTF effects and assumes a single homogeneous structure. The method therefore remains a proof of concept based on simulated data and, in its current form, is not directly applicable to the CTF-affected, heterogeneous data considered in this work.

A.2 Deep learning based micrograph denoising

Self-supervised denoising methods learn without clean targets under different assumptions. Noise2Noise [12] requires two conditionally independent noisy observations of the same underlying signal, an assumption that can be difficult to satisfy in practice. Noise2Void [11] predicts masked pixels from their neighborhoods. SURE [24] and Noise2Score [9] estimate a score directly from noisy data, and Noisier2Noise [17] trains on observations to which additional synthetic noise has been added. The latter is particularly compatible with our overall setup based on denoising score matching.

Several cryo-EM pipelines use Noise2Noise-style denoisers to improve the visual contrast of micrographs and facilitate inspection or particle picking [2, 22, 26, 27, 30]. Cryo-EM micrographs are produced by motion-correcting and averaging a sequence of movie frames. Paired Noise2Noise inputs can therefore be constructed by splitting these frames into disjoint

groups, for example odd- and even-numbered frames, and averaging each group into a separate half-micrograph. The resulting denoised micrographs provide improved visual contrast and may facilitate particle picking, but they do not recover sufficient additional information for downstream 2D or 3D reconstruction [26]. We use the currently in cryo-EM unexplored Noisier2Noise estimator where we do not need to separate our micrographs into movie frames.

A.3 Score-based diffusion models

Score-based diffusion models represent a data distribution p_{data} through the gradient of its logarithmic Gaussian-mollified density [25]. Let $\mathbf{x}_0 \sim p_{\text{data}}$ and perturb it by $\mathbf{x}_t = \mathbf{x}_0 + t\epsilon$, where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $t > 0$. The perturbed samples follow the Gaussian-mollified density

$$p_t(\mathbf{x}) = \int p_{\text{data}}(\mathbf{x}_0) \mathcal{N}(\mathbf{x}; \mathbf{x}_0, t^2 \mathbf{I}) d\mathbf{x}_0 = (p_{\text{data}} * \mathcal{N}(\mathbf{0}, t^2 \mathbf{I}))(\mathbf{x}). \quad (8)$$

As $t \rightarrow 0$, p_t approaches the data distribution and increasing t progressively removes fine-scale structure. The corresponding score $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ is a vector field pointing in the direction in which the mollified log-density increases most rapidly. It therefore describes the local geometry of p_t without requiring its unknown normalization constant.

Denoising score matching learns this vector field from samples alone [8, 25, 28]. A *denoiser* $\mathbf{D}_{\theta}(\mathbf{x}, t)$ is trained by

$$\min_{\theta} \mathbb{E}_{t, \mathbf{x}_0, \epsilon} \left[\lambda(t) \|\mathbf{D}_{\theta}(\mathbf{x}_0 + t\epsilon, t) - \mathbf{x}_0\|_2^2 \right], \quad \mathbf{x}_0 \sim p_{\text{data}}, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \quad (9)$$

For a fixed t , the population optimum is the conditional mean $\mathbf{D}^*(\mathbf{x}, t) = \mathbb{E}[\mathbf{x}_0 \mid \mathbf{x}_t = \mathbf{x}]$. Tweedie’s identity then relates this denoising estimate directly to the score:

$$\nabla_{\mathbf{x}} \log p_t(\mathbf{x}) = \frac{\mathbf{D}^*(\mathbf{x}, t) - \mathbf{x}}{t^2} \approx \frac{\mathbf{D}_{\theta}(\mathbf{x}, t) - \mathbf{x}}{t^2}. \quad (10)$$

Thus, the *denoiser* is not only a direction for removing noise, after scaling by t^{-2} , it estimates the score of the mollified data distribution.

We follow the EDM parameterization of Karras et al. [8], which preconditions an underlying neural network $\mathbf{F}_{\theta}$ as

$$\begin{aligned} \mathbf{D}_{\theta}(\mathbf{x}, t) &= c_{\text{skip}}(t)\mathbf{x} + c_{\text{out}}(t)\mathbf{F}_{\theta}(c_{\text{in}}(t)\mathbf{x}, c_{\text{noise}}(t)), \\ c_{\text{skip}}(t) &= \frac{\sigma_{\text{data}}^2}{t^2 + \sigma_{\text{data}}^2}, & c_{\text{out}}(t) &= \frac{t\sigma_{\text{data}}}{\sqrt{t^2 + \sigma_{\text{data}}^2}}, \\ c_{\text{in}}(t) &= \frac{1}{\sqrt{t^2 + \sigma_{\text{data}}^2}}, & c_{\text{noise}}(t) &= \frac{1}{4} \log t. \end{aligned} \quad (11)$$

These analytic coefficients keep the inputs, targets, and skip connection at comparable scales across noise levels, leaving the UNet to learn the remaining denoising function. In our experiments, we normalized the datasets to $\sigma_{\text{data}} = 1$.

After training, new samples can be obtained by drawing $\mathbf{x}(t_{\text{max}}) \sim \mathcal{N}(\mathbf{0}, t_{\text{max}}^2 \mathbf{I})$ and integrating the probability-flow ODE from t_{max} to t_{min} [25]:

$$\frac{d\mathbf{x}}{dt} = -t \mathbf{s}_t(\mathbf{x}) \approx \frac{\mathbf{x} - \mathbf{D}_{\theta}(\mathbf{x}, t)}{t}. \quad (12)$$

Our objective is not image generation, so we do not integrate Eq. (12). Instead, we use the learned score as a neural representation of local cryo-EM image statistics. For the raw-crop model, isotropic noise is replaced by the estimated covariance $\mathbf{C}$, and Eq. (10) becomes $t^{-2}\mathbf{C}^{-1}(\mathbf{D}_{\theta}(\mathbf{x}, t) - \mathbf{x})$, as used in Eq. (2). Fine-tuning on clean crops returns to white perturbations. The resulting *denoiser* supplies both the self-supervised denoising estimate and the score-distillation gradient used for 3D reconstruction.

A.4 Score Distillation Sampling

Score Distillation Sampling (SDS) was introduced to generate differentiable 3D assets using a pretrained text-to-image diffusion model [20] trained on internet scale of natural RGB images. In its original setting, a NeRF [16] is rendered from random viewpoints and updated so that the rendered images become likely under a text-conditioned image distribution. The update can be interpreted as minimizing a reverse KL divergence between noisy rendered images and the diffusion model’s mollified density. Thus SDS is able to generate 3D assets without ever being trained on 3D data. We adapt this principle to an unconditional cryo-EM score model: the optimized representation is a voxel density, its rendered images are orthogonal projections, and the clean-anchor term fixes one observed view. Thus, the score constrains projections across random orientations while the anchor selects the particular structure to reconstruct.

B Training of the *denoiser*

We train one score model per dataset in two consecutive stages. Both stages use the single-channel, fully convolutional UNet2DModel implementation from the `diffusers` library [29]. The network is attention-free and contains two residual layers per block with group normalization. The five down and up blocks have channel sizes 64, 128, 256, 256 and 384. Its convolutional form permits training on $c \times c$ crops and subsequent application to complete micrographs. Before each stage, the corresponding micrograph collection is standardized using its global pixel mean and standard deviation; these statistics are stored with the checkpoint and reused for denoising and reconstruction.

In the first stage, every training crop is sampled independently and uniformly from the raw micrographs $\mathcal{X}_{\text{raw}}$. We apply a random multiple-of-90° rotation and random flips. Restricting rotations to exact right angles avoids interpolation artifacts. The perturbation in Eq. (3) is colored using the background-noise covariance $\mathbf{C}$ estimated which we set to the ground truth covariance of the noise. In practice however $\mathbf{C}$ has to be estimated from a few background crops. This makes the synthetic corruption resemble the acquisition noise required by the Noisier2Noise estimator.

The second stage starts from the final noisy-model checkpoint rather than a random initialization. We transfer both the UNet weights and Adam optimizer state, recompute the normalization statistics on $\mathcal{X}_{\text{clean}}$, and continue denoising score matching with white Gaussian perturbations. Because the micrographs have already been denoised, clean crops can be augmented with arbitrary in-plane rotations and flips. We first extract the smallest larger crop that contains the desired $c \times c$ field after rotation and then center-crop. This keeps interpolation and padding artifacts outside the returned sample.

For both stages we use Adam with its default momentum parameters and mixed-precision training. Following Karras et al. [8], the UNet is EDM-preconditioned with $\sigma_{\text{data}} = 1$. As we are not interested in training a full diffusion model for data generation we only need to train on small t . We draw $z \sim \mathcal{N}(0, 1)$ and set

$$t = \text{clip}_{[t_{\min}, t_{\max}]}(1.05 t_{\max} \text{sigmoid}(z) - 0.025 t_{\max}), \quad t_{\min} = 0.002. \quad (13)$$

We used $t_{\max} = 2$ in the first training stage and $t_{\max} = 4$ in the second training stage. The denoising objective, network architecture, and optimizer are otherwise unchanged between stages. Only the crop distribution, augmentation, $t_{\max}$, and perturbation covariance change. Dataset-specific settings are summarized in Table 1.

C Score-Distillation based 2D-to-3D reconstruction

Derivation of $\nabla_{\mathbf{V}_a} \mathcal{J}(\mathbf{V}_a)$. Let

$$\mathbf{J}(\mathbf{V}_a, \mathbf{R}) = \frac{\partial \mathbf{f}(\mathbf{V}_a, \mathbf{R})}{\partial \mathbf{V}_a}$$

denote the Jacobian of the projection operator with respect to the volume. The reconstruction objective is

$$\begin{aligned} \mathcal{J}(\mathbf{V}_a) &= w_s t^2 \mathbb{E}_{\mathbf{R} \sim \mathcal{U}(\text{SO}(3))} [D_{\text{KL}}(\mathcal{N}(\mathbf{f}(\mathbf{V}_a, \mathbf{R}), t^2 \mathbf{I}) \parallel p_{\text{clean}, t})] \\ &\quad + \frac{w_a}{2} \|\mathbf{f}(\mathbf{V}_a, \mathbf{I}_3) - \mathbf{x}_a\|_2^2. \end{aligned} \quad (14)$$

We derive the score-distillation and anchor gradients separately.

By Tweedie's identity,

$$\mathbf{D}_{\theta_{\text{clean}}}(\mathbf{x}, t) = \mathbf{x} + t^2 \nabla_{\mathbf{x}} \log p_{\text{clean}, t}(\mathbf{x}). \quad (15)$$

Therefore,

$$\mathbb{E}_{\substack{\mathbf{R} \sim \mathcal{U}(\text{SO}(3)) \\ \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})}} [\mathbf{J}(\mathbf{V}_a, \mathbf{R})^\top (\mathbf{f}(\mathbf{V}_a, \mathbf{R}) - \mathbf{D}_{\theta_{\text{clean}}}(\mathbf{f}(\mathbf{V}_a, \mathbf{R}) + t\boldsymbol{\epsilon}, t))] \quad (16)$$

$$= -\mathbb{E}_{\mathbf{R}, \boldsymbol{\epsilon}} [\mathbf{J}(\mathbf{V}_a, \mathbf{R})^\top (t^2 \nabla_{\mathbf{x}} \log p_{\text{clean}, t}(\mathbf{f}(\mathbf{V}_a, \mathbf{R}) + t\boldsymbol{\epsilon}) + t\boldsymbol{\epsilon})] \quad (17)$$

$$= -\nabla_{\mathbf{V}_a} \mathbb{E}_{\mathbf{R}, \boldsymbol{\epsilon}} [t^2 \log p_{\text{clean}, t}(\mathbf{f}(\mathbf{V}_a, \mathbf{R}) + t\boldsymbol{\epsilon}) + t\boldsymbol{\epsilon}^\top \mathbf{f}(\mathbf{V}_a, \mathbf{R})] \quad (18)$$

$$= -\nabla_{\mathbf{V}_a} \left(t^2 \mathbb{E}_{\mathbf{R}, \boldsymbol{\epsilon}} [\log p_{\text{clean}, t}(\mathbf{f}(\mathbf{V}_a, \mathbf{R}) + t\boldsymbol{\epsilon})] + t \underbrace{\mathbb{E}_{\mathbf{R}, \boldsymbol{\epsilon}} [\boldsymbol{\epsilon}^\top \mathbf{f}(\mathbf{V}_a, \mathbf{R})]}_{=0} \right) \quad (19)$$

$$= -t^2 \nabla_{\mathbf{V}_a} \mathbb{E}_{\mathbf{R}, \boldsymbol{\epsilon}} [\log p_{\text{clean}, t}(\mathbf{f}(\mathbf{V}_a, \mathbf{R}) + t\boldsymbol{\epsilon})]. \quad (20)$$

The second term in Eq. (19) vanishes because $\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}$ and $\boldsymbol{\epsilon}$ is independent of $\mathbf{R}$.

Writing the expectation as an integral gives

$$(20) = -t^2 \nabla_{\mathbf{V}_a} \int_{\text{SO}(3)} \int_{\mathbb{R}^d} \mathcal{N}(\boldsymbol{\epsilon}; \mathbf{0}, \mathbf{I}) \log p_{\text{clean}, t}(\mathbf{f}(\mathbf{V}_a, \mathbf{R}) + t\boldsymbol{\epsilon}) \, d\boldsymbol{\epsilon} \, d\mathbf{R} \quad (21)$$

$$= -t^2 \nabla_{\mathbf{V}_a} \int_{\text{SO}(3)} \int_{\mathbb{R}^d} \mathcal{N}(\mathbf{x}; \mathbf{f}(\mathbf{V}_a, \mathbf{R}), t^2 \mathbf{I}) \log p_{\text{clean}, t}(\mathbf{x}) \, d\mathbf{x} \, d\mathbf{R}. \quad (22)$$

Here, d is the dimension of the image space, and we used the substitution

$$\mathbf{x} = \mathbf{f}(\mathbf{V}_a, \mathbf{R}) + t\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} = \frac{\mathbf{x} - \mathbf{f}(\mathbf{V}_a, \mathbf{R})}{t}. \quad (23)$$

Under this change of variables,

$$\mathcal{N}(\boldsymbol{\epsilon}; \mathbf{0}, \mathbf{I}) = t^d \mathcal{N}(\mathbf{x}; \mathbf{f}(\mathbf{V}_a, \mathbf{R}), t^2 \mathbf{I}), \quad d\boldsymbol{\epsilon} = t^{-d} d\mathbf{x}. \quad (24)$$

The negative entropy of the Gaussian is

$$\begin{aligned} &\int_{\mathbb{R}^d} \mathcal{N}(\mathbf{x}; \mathbf{f}(\mathbf{V}_a, \mathbf{R}), t^2 \mathbf{I}) \log \mathcal{N}(\mathbf{x}; \mathbf{f}(\mathbf{V}_a, \mathbf{R}), t^2 \mathbf{I}) \, d\mathbf{x} \\ &= -\frac{d}{2} (\log(2\pi t^2) + 1). \end{aligned} \quad (25)$$

This quantity is independent of both $\mathbf{V}_a$ and $\mathbf{R}$. We may therefore add it inside the gradient in Eq. (22), obtaining

$$(22) = t^2 \nabla_{\mathbf{V}_a} \int_{\text{SO}(3)} \left[\int_{\mathbb{R}^d} \mathcal{N}(\mathbf{x}; \mathbf{f}(\mathbf{V}_a, \mathbf{R}), t^2 \mathbf{I}) \log \frac{\mathcal{N}(\mathbf{x}; \mathbf{f}(\mathbf{V}_a, \mathbf{R}), t^2 \mathbf{I})}{p_{\text{clean}, t}(\mathbf{x})} \, d\mathbf{x} \right] d\mathbf{R} \quad (26)$$

$$= t^2 \nabla_{\mathbf{V}_a} \int_{\text{SO}(3)} D_{\text{KL}}(\mathcal{N}(\mathbf{f}(\mathbf{V}_a, \mathbf{R}), t^2 \mathbf{I}) \parallel p_{\text{clean}, t}) \, d\mathbf{R} \quad (27)$$

$$= t^2 \nabla_{\mathbf{V}_a} \mathbb{E}_{\mathbf{R} \sim \mathcal{U}(\text{SO}(3))} [D_{\text{KL}}(\mathcal{N}(\mathbf{f}(\mathbf{V}_a, \mathbf{R}), t^2 \mathbf{I}) \parallel p_{\text{clean}, t})]. \quad (28)$$

Consequently, the gradient of the score-distillation term is

$$\nabla_{\mathbf{V}_a} \mathcal{J}_{\text{SDS}}(\mathbf{V}_a) = w_s \mathbb{E}_{\mathbf{R}, \boldsymbol{\epsilon}} [\mathbf{J}(\mathbf{V}_a, \mathbf{R})^\top (\mathbf{f}(\mathbf{V}_a, \mathbf{R}) - \mathbf{D}_{\theta_{\text{clean}}}(\mathbf{f}(\mathbf{V}_a, \mathbf{R}) + t\boldsymbol{\epsilon}, t))]. \quad (29)$$

The anchor term is a standard least-squares objective:

$$\mathcal{J}_{\text{anchor}}(\mathbf{V}_a) = \frac{w_a}{2} \|\mathbf{f}(\mathbf{V}_a, \mathbf{I}_3) - \mathbf{x}_a\|_2^2. \quad (30)$$

Its gradient is

$$\nabla_{\mathbf{V}_a} \mathcal{J}_{\text{anchor}}(\mathbf{V}_a) = w_a \mathbf{J}(\mathbf{V}_a, \mathbf{I}_3)^\top (\mathbf{f}(\mathbf{V}_a, \mathbf{I}_3) - \mathbf{x}_a). \quad (31)$$

Combining the score-distillation and anchor gradients gives

$$\begin{aligned} \nabla_{\mathbf{V}_a} \mathcal{J}(\mathbf{V}_a) = & w_s \mathbb{E}_{\mathbf{R}, \epsilon} [\mathbf{J}(\mathbf{V}_a, \mathbf{R})^\top (\mathbf{f}(\mathbf{V}_a, \mathbf{R}) - \mathbf{D}_{\theta_{\text{clean}}}(\mathbf{f}(\mathbf{V}_a, \mathbf{R}) + t\epsilon, t))] \\ & + w_a \mathbf{J}(\mathbf{V}_a, \mathbf{I}_3)^\top (\mathbf{f}(\mathbf{V}_a, \mathbf{I}_3) - \mathbf{x}_a), \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \end{aligned} \quad (32)$$

which is precisely the reconstruction gradient in Eq. (7). A pseudo code of stochastic gradient descent corresponding to Eq. (7) can be seen in Algorithm 1. Here the first projection is kept at the identity orientation and compared with the clean anchor. The remaining projections use independently sampled rotations and score residuals. The BACKPROJECT operation denotes application of the transpose Jacobian of $\mathbf{f}$ and averages the resulting voxel updates over the projection batch.

Algorithm 1 3D Reconstruction

Require: denoiser $\mathbf{D}_{\theta_{\text{clean}}}$, clean anchor $\mathbf{x}_a \in \mathbb{R}^{c \times c}$, steps S , projection batch size $K > 1$, mollification degree t , learning rate $\eta > 0$, weights $w_s, w_a \geq 0$

- 1: $\mathbf{V}_a[x, y, z] \leftarrow 0$ for all $x, y, z \in \{1, \dots, c\}$ ▷ initialize volume
- 2: **for** $s = 1, \dots, S$ **do**
- 3: $\mathbf{R}_1 \leftarrow \mathbf{I}_3$
- 4: $\mathbf{R}_k \sim \mathcal{U}(\text{SO}(3))$ for $k = 2, \dots, K$ ▷ sample views
- 5: $\mathbf{x}_k \leftarrow \mathbf{f}(\mathbf{V}_a, \mathbf{R}_k)$ for $k = 1, \dots, K$
- 6: $\epsilon_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ for $k = 2, \dots, K$
- 7: $\mathbf{r}_k \leftarrow w_s [\mathbf{x}_k - \mathbf{D}_{\theta_{\text{clean}}}(\mathbf{x}_k + t\epsilon_k, t)]$ for $k = 2, \dots, K$
- 8: $\mathbf{r}_1 \leftarrow w_a (\mathbf{x}_1 - \mathbf{x}_a)$ ▷ anchor residual
- 9: $\Delta \mathbf{V}_a \leftarrow \text{BACKPROJECT}(\{\mathbf{r}_k, \mathbf{R}_k\}_{k=1}^K)$
- 10: $\mathbf{V}_a \leftarrow \mathbf{V}_a - \eta \Delta \mathbf{V}_a$
- 11: **end for**
- 12: **return** $\mathbf{V}_a$

D Experiments

D.1 Real homogeneous reconstruction

Data and preprocessing. We evaluate our method on a real homogeneous cryo-EM dataset from the RELION 5.0 tutorial. The tutorial subset contains 25 raw micrographs of β -galactosidase obtained from EMPIAR-10204. We Fourier-cropped each micrograph by a factor of three, resulting in a pixel size of 2.655 Å. For this experiment, we modeled the micrograph noise as additive white noise. After self-supervised denoising, we corrected the CTF of each micrograph using its deposited per-micrograph CTF estimate. The resulting clean micrographs have size 1232×1264 pixels.

Visual inspection of the clean micrographs revealed a particularly challenging dataset. Large regions of many micrographs contain impurities together with blurred, distorted, or fragmented particles; only a smaller fraction of the observations resemble well-defined β -galactosidase projections. We nevertheless applied no micrograph exclusion, particle filtering, or quality-based crop selection. Instead, the clean-crop score model was trained on random crops drawn uniformly from all processed micrographs. The representative field in Figure 3 is micrograph 20170629_00044.

Annealed reconstruction orientations. Direct application of the score-distillation reconstruction in Algorithm 1 was unstable on these data. We attribute this behavior to the clean-crop score model representing not only intact molecular projections, but also the

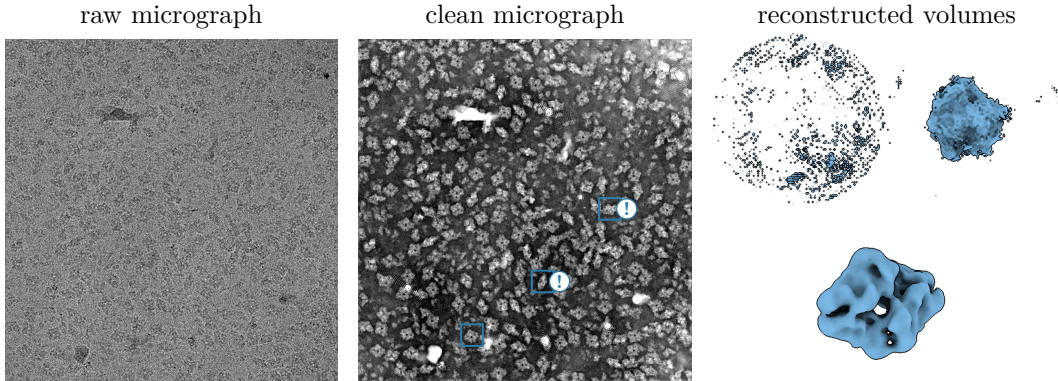


Figure 3: Representative processing and reconstruction results for the EMPIAR-10204 experiment. From left to right, the figure shows a raw micrograph, the corresponding clean micrograph after denoising and CTF correction, and three anchor-conditioned reconstructions. Both micrographs use their respective 1st to 99th percentile display ranges. Blue boxes mark the three 80×80 anchor crops. Circled exclamation marks identify the two failure cases.

abundant impurities, fragmented particles, and background structures. During reconstruction, the learned score can therefore favor projections resembling these unwanted observations rather than projections of an intact molecule.

To stabilize reconstruction, we annealed the sampled viewing directions around the orientation of the anchor image. At the beginning of optimization, the sampled rotations remain close to the anchor orientation. The maximum angular deviation then increases linearly from 0° to 90° during the first half of the reconstruction. At each iteration, rotation axes are sampled at random and rotation angles are drawn between zero and the current maximum angle. At the midpoint of optimization, we disable the angular restriction and sample unrestricted random orientations for the entire second half. Thus, the volume first develops under anchor-proximal views, is exposed to progressively broader orientations, and only then receives score-distillation updates from the full orientation distribution. With this modification, the algorithm produced recognizable reconstructions for compact, non-elongated β -galactosidase anchor projections. Figure 3 shows a representative raw micrograph, the corresponding denoised and CTF-corrected micrograph, and one successful reconstruction together with two representative failures.

Comparison with RELION 5.0. We compare our successful reconstruction with the homogeneous *ab initio* model produced by RELION 5.0 [23]. This RELION *ab initio* volume is also distributed in the RELION 5.0 tutorial project folder. The RELION workflow operates on a particle stack that was curated by class averaging and manual selection, leaving 4598 well-resolved, approximately centered, and boxed particles at $3.54 \text{ \AA}$ per pixel. In addition, the D_2 symmetry of β -galactosidase was supplied to RELION a priori. Both choices provide a substantial advantage over our reconstruction, which uses no assumed molecular symmetry and is conditioned only on a selected anchor from the processed micrographs.

For evaluation, we aligned both reconstructed volumes to the refined deposited map EMD-6840 and then resampled them on the voxel grid of EMD-6840. We computed an unmasked FSC against this reference. Figure 4 shows the FSC curves together with the aligned and resampled density maps. As expected from the more favorable input stack and the known symmetry, RELION reaches a higher spatial resolution. Despite using all micrographs without quality-based exclusion and learning from random crops that are dominated by difficult observations, our method still recovers a recognizable *ab initio* density of β -galactosidase. This experiment therefore demonstrates feasibility on real homogeneous data while also highlighting the sensitivity of score-distillation reconstruction to contamination in the learned clean-projection distribution.

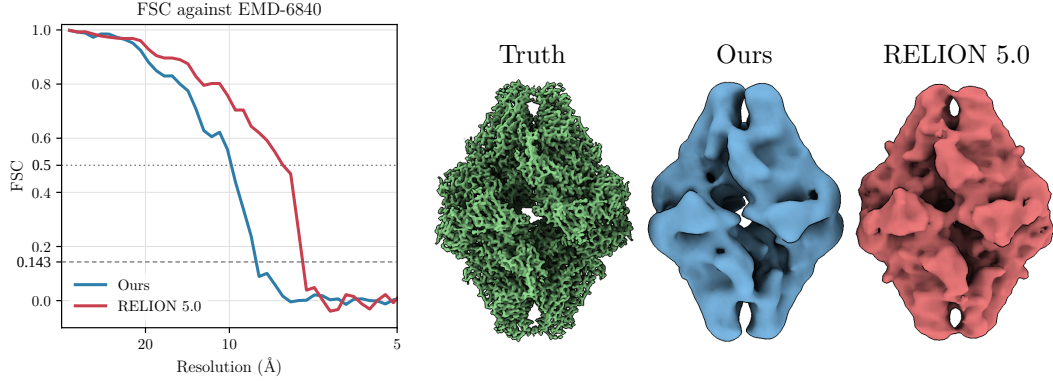


Figure 4: Homogeneous reconstruction of β -galactosidase. The FSC plot on the left compares our reconstruction and the RELION 5.0 *ab initio* model against the deposited EMD-6840 map without masking. RELION uses a curated stack of approximately centered particles and the known D_2 symmetry, whereas our method operates on the raw micrographs and a single selected anchor crop without assuming symmetry.

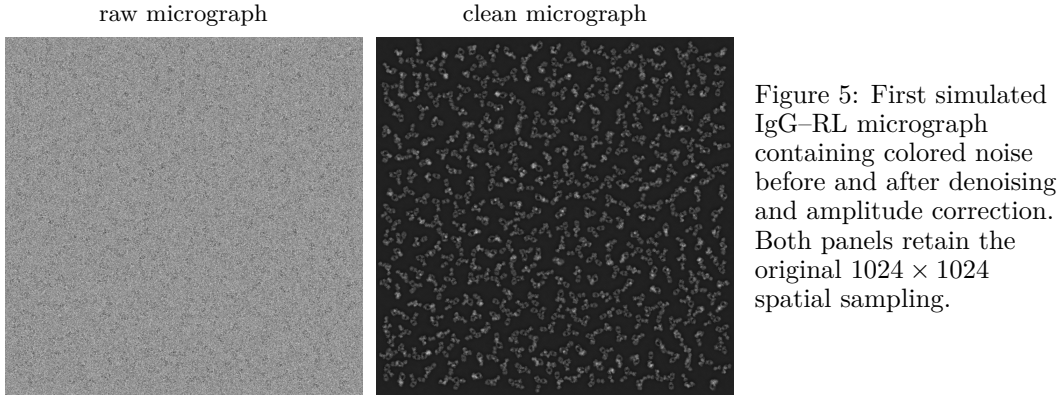


Figure 5: First simulated IgG-RL micrograph containing colored noise before and after denoising and amplitude correction. Both panels retain the original 1024×1024 spatial sampling.

D.2 Synthetic micrograph generation

Both synthetic datasets were generated with ASPIRE’s `MicrographSimulation` module. For every particle, ASPIRE sampled a volume index uniformly and an independent viewing rotation uniformly from $SO(3)$. The resulting projections were CTF-filtered, placed at random micrograph locations subject to the dataset-specific boundary and separation constraints below, and finally corrupted by additive noise. We retained the particle locations, source-volume indices, and CTF parameters as ground truth for evaluation.

CTF parameters. All particles in one micrograph share the same radial CTF. We used an accelerating voltage of 300 kV, spherical aberration $C_s = 2.26$ mm, amplitude-contrast parameter $\alpha = 0.1$, and envelope parameter $B = 0$. The radial model sets the two principal defocus values equal and therefore introduces no astigmatism. Defocus was varied deterministically across micrographs: values were linearly spaced from 1.0 – 1.5 μm for the cages and from 0.6 – 1.5 μm for IgG-RL.

Clathrin cages. The dataset contains four 85^3 clathrin-cage volumes and 200 micrographs of size 1024^2 at 10 Å per pixel (we downsampled the original volumes from EMD-0114, EMD-0115, EMD-0116 and EMD-0118). Each micrograph contains 100 particles, for 20 000 projections in total. Particle centers were restricted to lie at least 45 pixels from the micrograph boundary with a requested minimum center-to-center distance of 80 pixels. We added independent, zero-mean white Gaussian noise.

IgG-RL. The dataset contains 100 different conformations of size 128^3 and 250 micrographs of size 1024^2 at $3\text{ \AA}$ per pixel. Each micrograph contains 400 particles, giving 100 000 projections. Particle centers used a 30-pixel boundary and a requested minimum center-to-center distance of 40 pixels. We generated a stationary, zero-mean colored Gaussian field whose radial power spectral density is

$$S_N(\omega_x, \omega_y) = A \exp \left[-\frac{\omega_x^2 + \omega_y^2}{2(0.35\pi)^2} \right], \quad (33)$$

where π is the Nyquist frequency and A is chosen such that the mean of the discrete power spectrum is 9×10^{-6} . Thus, the expected real-space noise variance is 9×10^{-6} and its standard deviation is 3×10^{-3} . Samples were obtained by multiplying the Fourier coefficients of white Gaussian noise by $\sqrt{S_N}$. Consequently, the noise remains Gaussian but is spatially correlated, with more power at low frequencies and a smoother appearance than white noise (See Fig. 5).

Signal-to-noise ratio. We estimated the full-micrograph SNR from ten paired simulations per dataset. Each pair shares the same particles, poses, positions, amplitudes, and CTF; one image is noise-free and the other uses the dataset noise model. We retained the original particle density and linearly spaced the ten defocus values over the full ranges above. This gives an SNR of 0.197 (-7.06 dB) for the cages and 0.402 (-3.95 dB) for IgG-RL. These are full-field values: particle-free pixels contribute noise but no signal, so the estimates also reflect the respective particle densities.

Relation to CryoBench SNR labels. These values are not directly comparable to the nominal CryoBench SNRs [7]. CryoBench’s simulation software `cryosim` defines SNR from global pixel variance over the complete boxed image, including huge parts of noise-only background, and CryoBench reports this value before subsequent Fourier cropping. Its IgG-RL label of 0.01 (-20 dB) refers to 256^2 images before cropping to 128^2 . Halving the Fourier width retains approximately one quarter of the white-noise power and therefore raises the effective SNR by $10 \log_{10}(4) = 6.02\text{ dB}$, to about -13.98 dB . Expressing the same particle signal in an 88^2 rather than 128^2 box (which still shows large signal free parts) adds another $10 \log_{10}(128^2/88^2) = 3.25\text{ dB}$, giving a box-matched value of approximately -10.73 dB . The underlying particle has not changed; only the processing stage and amount of noise-only support have. Our SNRs are instead measured on dense 1024^2 micrographs containing 100 or 400 particles and hence a considerably smaller fraction of signal-free area, while all particle signals contribute to the numerator. Moreover, IgG-RL uses colored noise, for which total noise power does not describe its spectral overlap with the particle signal. Scalar SNRs therefore document each simulation but do not, without matched processing and support, provide a direct ranking of reconstruction difficulty.

D.3 CryoDRGN-AI baseline

We ran CryoDRGN-AI (v4.2.0) using the command `cryodrgn abinit` with `uninvert-data`, and `window-radius-gt-real 0.99`. We increased training from the default 30 epochs to 100 epochs, as in CryoBench, and retained all remaining optimization and hyperparameter defaults.

Table 1: Overview of training, denoising, and reconstruction settings for each dataset.

Setting	Cages	IgG-RL	EMPIAR-10204
<i>Raw-crop training</i>			
Crop size c	128	96	96
Learning rate	1×10^{-4}	1×10^{-4}	1×10^{-4}
$t_{\min}$	0.002	0.002	0.002
$t_{\max}$	2	2	2
Training steps	$\sim 78\text{k}$	$\sim 136\text{k}$	$\sim 136\text{k}$
Batch size	64	128	128
Runtime	$\sim 9.8\text{ h}$	$\sim 11.3\text{ h}$	$\sim 11.3\text{ h}$
<i>Denoising</i>			
Noise scale σ	0.92	0.85	0.9
Noisier2Noise factor α	1/4	1/4	1/4
Runtime all micrographs	few seconds	few seconds	few seconds
<i>Clean-crop fine-tuning</i>			
Learning rate	1×10^{-4}	1×10^{-4}	1×10^{-4}
$t_{\min}$	0.002	0.002	0.002
$t_{\max}$	4	4	4
Training steps	$\sim 155\text{k}$	$\sim 234\text{k}$	$\sim 234\text{k}$
Batch size	64	128	128
Runtime	$\sim 11.6\text{ h}$	$\sim 19.5\text{ h}$	~ 19.5
<i>Reconstruction</i>			
Clean-anchor size c	96	80	80
Projection batch size K	2	4	4
Mollification degree t	1	1.5	1
Score weight w_s	$K - 1$	$K - 1$	$K - 1$
Anchor weight w_a	1	$1 - (s/S)^{10}$	$1 - (s/S)$
Learning rate η	2.5	2.5	1
Reconstruction steps S	10k	10k	10k
Runtime per anchor	$\sim 2.30\text{ min}$	$\sim 2.10\text{ min}$	~ 2.10
<i>Compute</i>			
Hardware	A100-SXM4-40GB	A100-SXM4-40GB	A100-SXM4-40GB
Numerical precision	16-mixed	16-mixed	16-mixed

E Limitations

Runtime. State-of-the-art methods such as CryoDRGN-AI are significantly faster. Training its VAE-based model, including pose search, took approximately 1 h for the 20k cage particles and 5 h for the 100k IgG-RL particles. In comparison, the two training stages of our approach took approximately $9.8\text{ h} + 11.6\text{ h}$ for the cage micrographs, and $11.3\text{ h} + 19.5\text{ h}$ for the IgG-RL micrographs. However, these runtimes are not directly comparable: CryoDRGN-AI receives already picked, tightly boxed particles as input, whereas our method operates directly on micrographs and implicitly performs the relevant data extraction. Further runtime details are provided in Table 1.

Real data. Our score-based approach is still at an early stage of development, and a thorough evaluation on real experimental data remains future work. In particular, it remains to be determined whether the Noisier2Noise estimator retains sufficient structural information under strong noise with a complex and difficult-to-model distribution. For EMPIAR-10204, our experiments indicate that a white-noise assumption is sufficient to reconstruct a low-to-medium-resolution volume. However, this result may not generalize to other datasets. Future work should therefore investigate how accurately the noise distribution can be estimated from background patches and how errors in this estimate affect the quality of the denoised micrographs and subsequent reconstructions.

Micrograph quality. We hypothesize that low-quality micrographs, or low-quality regions within otherwise usable micrographs, negatively affect both denoising and 3D reconstruction. Compared with our experiments on idealized simulated micrographs, reconstructing β -galactosidase from EMPIAR-10204 is considerably less stable. During score-distillation, the reconstruction may be driven toward artifacts and other unwanted structures present in the training data rather than toward genuine particles of interest. One possible solution would be to inspect the denoised micrographs and exclude low-quality micrographs or mask regions dominated by experimental artifacts, contamination, or severely distorted particles. Such global micrograph-level selection and masking would still require substantially less manual effort than particle-by-particle picking.

High resolution. In its current form, our approach is not suitable for high-resolution reconstruction, which is already particularly challenging in the general heterogeneous setting. Applying PSF/CTF correction after denoising introduces unavoidable errors because frequencies near the zeros of the CTF cannot be recovered reliably by the Wiener filter used in our current implementation. In future work, we therefore plan to investigate whether the CTF can instead be incorporated directly into the forward model used during score-distillation. This would allow the score model to be trained on denoised but non-CTF-corrected micrographs and could avoid the explicit inversion of poorly observed frequencies.