CryoAI: Amortized Inference of Poses for Ab Initio Reconstruction of 3D Molecular Volumes from Real Cryo-EM Images

Axel Levy[1,2,*], Frédéric Poitevin[1,*], Julien Martel[2,*], Youssef Nashed[3], Ariana Peck[1], Nina Miolane[4], Daniel Ratner[3], Mike Dunne[1], and Gordon Wetzstein[2]

[1]LCLS, SLAC National Accelerator Laboratory, Menlo Park, CA, USA
[2]Stanford University, Department of Electrical Engineering, Stanford, CA, USA
[3]ML Initiative, SLAC National Accelerator Laboratory, Menlo Park, CA, USA
[4]University of California Santa Barbara, Department of Electrical and Computer Engineering, Santa Barbara, CA, USA

Abstract. Cryo-electron microscopy (cryo-EM) has become a tool of fundamental importance in structural biology, helping us understand the basic building blocks of life. The algorithmic challenge of cryo-EM is to jointly estimate the unknown 3D poses and the 3D electron scattering potential of a biomolecule from millions of extremely noisy 2D images. Existing reconstruction algorithms, however, cannot easily keep pace with the rapidly growing size of cryo-EM datasets due to their high computational and memory cost. We introduce cryoAI, an *ab initio* reconstruction algorithm for homogeneous conformations that uses direct gradient-based optimization of particle poses and the electron scattering potential from single-particle cryo-EM data. CryoAI combines a learned encoder that predicts the poses of each particle image with a physics-based decoder to aggregate each particle image into an implicit representation of the scattering potential volume. This volume is stored in the Fourier domain for computational efficiency and leverages a modern coordinate network architecture for memory efficiency. Combined with a symmetric loss function, this framework achieves results of a quality on par with state-of-the-art cryo-EM solvers for both simulated and experimental data, one order of magnitude faster for large datasets and with significantly lower memory requirements than existing methods.

Keywords: Cryo-electron Microscopy, Neural Scene Representation

1 Introduction

Understanding the 3D structure of proteins and their associated complexes is crucial for drug discovery, studying viruses, and understanding the function of the fundamental building blocks of life. Towards this goal, cryo-electron microscopy (cryo-EM) of isolated particles has been developed as the go-to method for imaging and studying molecular assemblies at near-atomic resolution [21,31,39]. In a

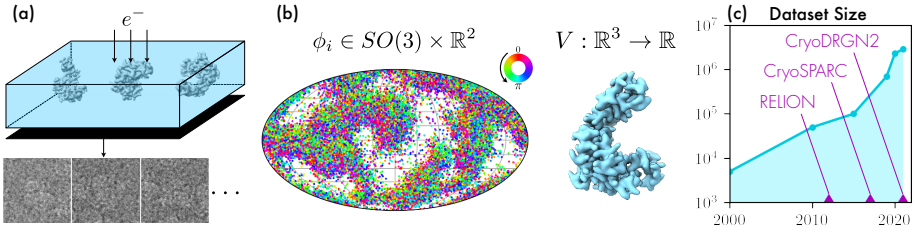


Fig. 1. (a) (Top) Illustration of a cryo-EM experiment. Molecules are frozen in a random orientation and their electron scattering potential (i.e., volume) V interacts with an electron beam imaged on a detector. (Bottom) Noisy projections (i.e., particles) of V selected from the full micrograph measured by the detector. (b) Output of a reconstruction algorithm: poses ϕ_i and volume V. Each pose is characterized by a rotation in $SO(3)$ (hue represents in-plane rotation) and a translation in $\mathbb{R}^2$ (not shown). An equipotential surface of V is shown on the right. (c) Evolution of the maximum number of images collected in one day [29] and established and emerging state-of-the-art reconstruction methods.

cryo-EM experiment, a purified solution of the molecule of interest is frozen in a thin layer of vitreous ice, exposed to an electron beam, and randomly oriented projections of the electron scattering potential (i.e., the volume) are imaged on a detector (Fig. 1 (a)). These raw *micrographs* are then processed by an algorithm that reconstructs the volume and estimates the unknown pose, including orientation and centering shift, of each particle extracted from the micrographs (Fig. 1 (b)).

Recent advances in sample preparation, instrumentation, and data collection capabilities have resulted in very large amounts of data being recorded for each cryo-EM experiment [4,29] (Fig. 1 (c)). Millions of noisy (images of) particles, each with an image size on the order of 100^2–400^2 pixels, need to be processed by the reconstruction algorithm to jointly estimate the pose of each particle and the unknown volume. Most existing algorithms that have been successful with experimental cryo-EM data address this problem using a probabilistic approach that iteratively alternates between updating the volume and the estimated poses [44,37,59,61]. The latter "orientation matching" step, however, is computationally expensive, requiring an exhaustive search in a 5-dimensional space ($\phi_i \in SO(3) \times \mathbb{R}^2$) for each particle. In spite of using smart pose search strategies and optimization schedules, the orientation matching step is the primary bottleneck of existing cryo-EM reconstruction algorithms, requiring hours to estimate a single volume and scaling poorly with increasing dataset sizes.

We introduce cryoAI, a technique that uses direct gradient-based optimization to jointly estimate the poses and the electron scattering potential of a non-deformable molecule (*homogeneous* reconstruction). Our method operates in an unsupervised manner over a set of images with an encoder–decoder pipeline. The encoder learns a discriminative model that associates each particle image with a pose and the decoder is a generative physics-based pipeline that uses the

predicted pose and a description of the volume to predict an image. The volume is maintained by an implicit, i.e., neural network–parameterized, representation in the decoder, and the image formation model is simulated in Fourier space, thereby avoiding the approximation of integrals via the Fourier-slice theorem (see Sec. 3.1). By learning a mapping from images to poses, cryoAI avoids the computationally expensive step of orientation matching that limits existing cryo-EM reconstruction methods. Our approach thus amortizes over the size of the dataset and provides a scalable approach to working with modern, large-scale cryo-EM datasets. We demonstrate that cryoAI performs homogeneous reconstructions of a comparable resolution but with nearly one order of magnitude faster runtime than state-of-the-art methods using datasets containing millions of particles.

Specifically, our contributions include

- a framework that learns to map images to particle poses while reconstructing an electron scattering potential for homogeneous single-particle cryo-EM;
- demonstration of reconstruction times and memory consumption that amortize over the size of the dataset, with nearly an order of magnitude improvement over existing algorithms on large datasets;
- formulations of a symmetric loss function and an implicit Fourier-domain volume representation that enable the high-quality reconstructions we show.

Source code is available at `https://github.com/compSPI/cryoAI`.

2 Related Work

Estimating the 3D structure of an object from its 2D projections with known orientations is a classical problem in tomography and has been solved using backprojection-based methods [18,43] or compressive sensing–style solvers [8,12]. In cryo-EM, the reconstruction problem is complicated by several facts: (1) the poses of the unknown object are also unknown for all projections; (2) the signal-to-noise ratio (SNR) is extremely low (around -20 dB for experimental datasets [6,5]); (3) the molecules in a sample can deform and be frozen in various (unknown) conformations. Unlike homogeneous reconstruction methods, heterogeneous methods take into account the deformations of the molecule and reconstruct a discrete set or a low-dimensional manifold of conformations. Although they give more structural information, most recent heterogeneous methods [59,36,62,9] assume the poses to be known. For each particle i, a pose ϕ_i is defined by a rotation $R_i \in SO(3)$ and a translation $\mathbf{t}_i \in \mathbb{R}^2$. In this work, we do not assume the poses to be known and aim to estimate the electron scattering function V of a unique underlying molecule in a homogeneous setting. We classify previous work on pose estimation into two inference categories [11]: non-amortized and amortized.

Non-amortized Inference refers to a class of methods where the posterior distribution of the poses $p(\phi_i|Y_i, V)$ is computed independently for each image Y_i. Common-line approaches [51,47,55,16,35,57], projection-matching strategies [33,3] and Bayesian formulations [24,10,45,37] belong to this category. The

software package RELION [44] widely popularized the Bayesian approach by performing Maximum-A-Posteriori (MAP) optimization through Expectation–Maximization (EM). Posterior distributions over the poses (and the optional *conformational* states) are computed for each image in the expectation step and all frequency components of the volume are updated in the maximization step, which makes the approach computationally costly. The competing software cryoSPARC [37] proposed to perform MAP optimization jointly using stochastic gradient descent (SGD) to optimize the volume V and branch-and-bound algorithms [22] to estimate the poses ϕ_i. While a gradient-based optimization scheme for V circumvents the costly updates in the maximization step of RELION, a pose must be estimated for each image by aligning each 2D projection Y_i with the estimated 3D volume V. Although branch-and-bound algorithms can accelerate the pose search, this step remains computationally expensive and is one of the bottlenecks of the method in terms of runtime. Ullrich *et al.* [50] proposed a variational and differentiable formulation of the optimization problem in the Fourier domain. Although they demonstrated that their method can estimate the volume when poses are known, they also showed that jointly optimizing the pose posterior distributions by SGD fails due to the high non-convexity of the problem. Instead of parameterizing the volume with a 3D voxel array, Zhong *et al.* proposed in cryoDRGN [60,59,61] to use a coordinate-based representation (details in Sec. 3.4) to directly approximate the electron scattering function in Fourier space. Their neural representation takes 3D Fourier coordinates and a latent vector encoding the conformational state as input, therefore accounting for continuous deformations of the molecule. The latest published version of cryoDRGN [59] reports excellent results on the reconstruction of conformation heterogeneities but assumes the poses to be determined by a consensus reconstruction. Poses are jointly estimated with V in cryoDRGN-BNB [60] and cryoDRGN2 [61], but in spite of a frequency-marching strategy, the use of a branch-and-bound algorithm and a later introduced multi-resolution approach the global 5D pose search remains the most computationally expensive step in their pipeline.

Amortized Inference techniques, on the other hand, learn a parameterized function $q_\xi(Y_i)$ that approximates the posterior distribution of the poses $p(\phi_i|Y_i, V)$ [14]. At the expense of optimizing the parameter ξ, these techniques avoid the orientation matching step which is the main computational bottleneck in non-amortized methods. Lian *et al* [23] demonstrated the possibility of using a convolutional neural network to approximate the mapping between cryo-EM images and orientations, but their method cannot perform end-to-end volume reconstruction. In cryoVAEGAN [27], Miolane *et al.* showed that the in-plane rotation could be disentangled from the contrast transfer function (CTF) parameters in the latent space of an encoder. Rosenbaum *et al.* [41] were the first to demonstrate volume reconstruction from unknown poses in a framework of amortized inference. In their work, distributions of poses and conformational states are predicted by the encoder of a Variational Autoencoder (VAE) [20]. In their model-based decoder, the predicted conformation is used to deform a base

backbone frame of Gaussian blobs and the predicted pose is used to make a projection of these blobs. The reconstructed image is compared to the measurement in order to optimize the parameters of both the encoder and the decoder. While this method is able to account for conformational heterogeneity in a dataset, it requires *a priori* information about the backbone frame. CryoPoseNet [30] proposed a non-variational autoencoder framework that can perform homogeneous reconstruction with a random initialization of the volume, avoiding the need for prior information about the molecule. Although it demonstrated the possibility of using a non-variational encoder to predict the orientations R_i, cryoPoseNet assumes the translations $\mathbf{t}_i$ to be given and the volume is stored in real space in the decoder (while the image formation model is in Fourier space, see Sec. 3.1), thereby requiring a 3D Fourier transform at each forward pass and making the overall decoding step slow. The volume reconstructed by cryoPoseNet often gets stuck in local minima, which is a problem we also address in this paper (see Sec. 3.5). Finally, the two last methods only proved they could be used with simulated datasets and, to the best of our knowledge, no amortized inference technique for volume estimation from unknown poses have been proven to work with experimental datasets in cryo-EM.

Previous methods differ in the way poses are inferred in the generative model. Yet, the only variable of interest is the description of the conformational state (for heterogeneous methods) and associated molecular volumes, while poses can be considered "nuisance" variables. As a result, recent works have explored methods that avoid the inference of poses altogether, such as GAN-based approaches [1]. CryoGAN [17], for example, used a cryo-EM simulator and a discriminator neural network to optimize a 3D volume. Although preliminary results are shown on experimental datasets, the reconstruction cannot be further refined with other methods due to the absence of predicted poses.

Our approach performs an amortized inference of poses and therefore circumvents the need for expensive searches over $SO(3) \times \mathbb{R}^2$, as in non-amortized techniques. In the implementation, no parameter needs to be statically associated with each image. Consequently, the memory footprint and the runtime of our algorithm does not scale with the number of images in the dataset. We introduce a loss function called "symmetric loss" that prevents the model from getting stuck in local minima with spurious planar symmetries. Finally, in contrast to previous amortized inference techniques, our method can perform volume reconstruction on experimental datasets.

3 Methods

3.1 Image Formation Model and Fourier-slice Theorem

In a cryo-EM sample, the charges carried by each molecule and their surrounding environment create an electrostatic potential that scatters probing electrons, which we refer to as the electron scattering "volume," and consider as a mapping

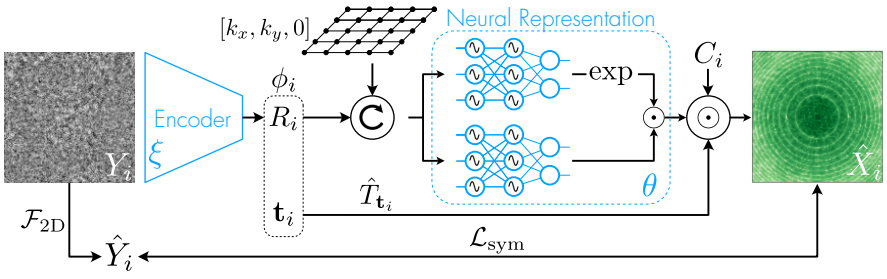


Fig. 2. Overview of our pipeline. The encoder, parameterized by ξ learns to map images Y_i to their associated pose $\phi_i = (R_i, \mathbf{t}_i)$. The matrix R_i rotates a slice of 3D coordinates in Fourier space. The coordinates are fed into a neural representation of $\hat{V}$, parameterized by θ. The output is multiplied by the CTF C_i and the translation operator $\hat{T}_{\mathbf{t}_i}$ to build $\hat{X}_i$, a noise-free estimation of $\mathcal{F}_{2D}[Y_i] = \hat{Y}_i$. $\hat{X}_i$ and $\hat{Y}_i$ are compared *via* the symmetric loss $\mathcal{L}_{sym}$. Differentiable parameters are represented in blue.

$$V : \mathbb{R}^3 \to \mathbb{R}. \tag{1}$$

In the sample, each molecule i is in an unknown orientation $R_i \in SO(3) \subset \mathbb{R}^{3\times3}$. The probing electron beam interacts with the electrostatic potential and its projections

$$Q_i : (x, y) \mapsto \int_z V\left(R_i \cdot [x, y, z]^T\right) dz \tag{2}$$

are considered mappings from $\mathbb{R}^2$ to $\mathbb{R}$. The beam then interacts with the lens system characterized by the Point Spread Function (PSF) P_i and individual particles are cropped from the full micrograph. The obtained images may not be perfectly centered on the molecule and small translations are modeled by $\mathbf{t}_i \in \mathbb{R}^2$. Finally, taking into account signal arising from the vitreous ice into which the molecules are embedded as well as the non-idealities of the lens and the detector, each image Y_i is generally modeled as

$$Y_i = T_{\mathbf{t}_i} * P_i * Q_i + \eta_i \tag{3}$$

where $*$ is the convolution operator, $T_{\mathbf{t}}$ the $\mathbf{t}$-translation kernel and η_i white Gaussian noise on $\mathbb{R}^2$ [53,44].

With a formulation in real space, both the integral over z in Eq. (2) and the convolution in Eq. (3) make the simulation of the image formation model computationally expensive. A way to avoid these operations is to use the Fourier-slice Theorem [7], which states that for any volume V and any orientation R_i,

$$\mathcal{F}_{2D}\left[Q_i\right] = \mathcal{S}_i\left[\mathcal{F}_{3D}\left[V\right]\right], \tag{4}$$

where $\mathcal{F}_{2D}$ and $\mathcal{F}_{3D}$ are the 2D and 3D Fourier transform operators and $\mathcal{S}_i$ the "slice" operator defined such that for any $\hat{V} : \mathbb{R}^3 \to \mathbb{C}$,

$$\mathcal{S}_i[\hat{V}] : (k_x, k_y) \mapsto \hat{V}\left(R_i \cdot [k_x, k_y, 0]^T\right). \tag{5}$$

That is, $\mathcal{S}_i[\hat{V}]$ corresponds to a 2D slice of $\hat{V}$ with orientation R_i and passing through the origin. In a nutshell, if $\hat{Y}_i = \mathcal{F}_{\text{2D}}[Y_i]$ and $\hat{V} = \mathcal{F}_{\text{3D}}[V]$, the image formation model in Fourier space can be expressed as

$$\hat{Y}_i = \hat{T}_{\mathbf{t}_i} \odot C_i \odot \mathcal{S}_i[\hat{V}] + \hat{\eta}_i, \tag{6}$$

where $\odot$ is the element-wise multiplication, $C_i = \mathcal{F}_{\text{2D}}[P_i]$ is the Contrast Transfer Function (CTF), $\hat{T}_{\mathbf{t}}$ the $\mathbf{t}$-translation operator in Fourier space (phase shift) and $\hat{\eta}_i$ complex white Gaussian noise on $\mathbb{R}^2$. Based on this generative model, cryoAI solves the inverse problem of inferring $\hat{V}$, R_i and $\mathbf{t}_i$ from $\hat{Y}_i$ assuming C_i is known.

3.2 Overview of CryoAI

CryoAI is built with an autoencoder architecture (see Fig. 2). The encoder takes an image Y_i as input and outputs a predicted orientation R_i along with a predicted translation $\mathbf{t}_i$ (Sec. 3.3). R_i is used to rotate a 2-dimensional grid of L^2 3D-coordinates $[k_x, k_y, 0] \in \mathbb{R}^3$ which are then fed into the neural network $\hat{V}_\theta$. This neural network is an implicit representation of the current estimate of the volume $\hat{V}$ (in Fourier space), and this query operation corresponds to the "slicing" defined by Eq. (5) (Sec. 3.4). Based on the estimated translation $\mathbf{t}_i$ and given CTF parameters C_i, the rest of the image formation model described in Eq. (6) is simulated to obtain $\hat{X}_i$, a noise-free estimation of $\hat{Y}_i$. These images are compared using a loss described in Sec. 3.5 and gradients are backpropagated throughout the differentiable model in order to optimize both the encoder and the neural representation.

3.3 Pose Estimation

CryoAI uses a Convolutional Neural Network (CNN) to predict the parameters R_i and $\mathbf{t}_i$ from a given image, thereby avoiding expensive orientation matching computations performed by other methods [44,37,61]. The architecture of this encoder has three layers.

1. *Low-pass filtering*: $Y_i \in \mathbb{R}^{L \times L}$ is fed into a bank of Gaussian low-pass filters.
2. *Feature extraction*: the filtered images are stacked channel-wise and fed into a CNN whose architecture is inspired by the first layers of VGG16 [46], which is known to perform well on image classification tasks.
3. *Pose estimation*: this feature vector finally becomes the input of two separate fully-connected neural networks. The first one outputs a vector of dimension 6 of $S^2 \times S^2$ [63] (two vectors on the unitary sphere in $\mathbb{R}^3$) and converted into a matrix $R_i \in \mathbb{R}^{3\times3}$ using the PyTorch3D library [38]. The second one outputs a vector of dimension 2, directly interpreted as a translation vector $\mathbf{t}_i \in \mathbb{R}^2$.

We call ξ the set of differentiable parameters in the encoder described above. We point the reader to Supp. B for more details about the architecture of the encoder.

3.4 Neural Representation in Fourier Space (FourierNet)

Instead of using a voxel-based representation, we maintain the current estimate of the volume using a neural representation. This representation is parameterized by θ and can be see seen as a mapping $\hat{V}_\theta : \mathbb{R}^3 \to \mathbb{C}$.

In imaging and volume rendering, neural representations have been used to approximate signals defined in real space [32,2,13,25,49]. Neural Radiance Field (NeRF) [26] is a successful technique to maintain a volumetric representation of a real scene. A view-independent NeRF model, for example, maps real 3D-coordinates $[x, y, z]$ to a color vector and a density scalar using positional encoding [52] and a set of fully-connected layers with ReLU activation functions. Sinusoidal Representation Networks (SIRENs) [48] can also successfully approximate 3D signed distance functions with a shallow fully-connected neural network using sinusoidal activation functions. However, these representations are tailored to approximate signals defined in real space. Here, we want to directly represent the Fourier transform of the electrostatic potential of a molecule. Since this potential is a smooth function of the spatial coordinates, the amplitude of its Fourier coefficients $\hat{V}(\mathbf{k})$ is expected to decrease with $|\mathbf{k}|$, following a power law (see Supp. C for more details). In practice, this implies that $|\hat{V}|$ can vary over several orders of magnitude and SIRENs, for example, are known to poorly approximate these types of functions [48]. The first method to use neural representations for volume reconstruction in cryo-EM, cryoDRGN [60,61], proposed to use a Multi-Layer Perceptron (MLP) with positional encoding in Hartley space (where the FST still applies).

With our work, we introduce a new kind of neural representation (FourierNet), tailored to represent signals defined in the Fourier domain, inspired by the success of SIRENs for signals defined in real space. Our idea is to allow a SIREN to represent a signal with a high dynamic range by raising its output in an exponential function. Said differently, the SIREN only represents a signal that scales logarithmically with the approximated function. Since Fourier coefficients are defined on the complex plane, we use a second network in our implicit representation to account for the phase variations. This architecture is summarized in Fig. 2 and details on memory requirements are given in Supp. C. Input coordinates $[k_x, k_y, k_z]$ are fed into two separate SIRENs outputting 2-dimensional vectors. For one of them, the exponential function is applied element-wise and the two obtained vectors are finally element-wise mutliplied to produce a vector in $\mathbb{R}^2$, mapped to $\mathbb{C}$ with the Cartesian coordinate system. Since $\hat{V}_\theta$ must represent the Fourier transform of real signals, we know that it should verify $\hat{V}_\theta(-\mathbf{k}) = \hat{V}_\theta(\mathbf{k})^*$. We enforce this property by defining

$$\hat{V}_\theta(\mathbf{k}) = \hat{V}_\theta(-\mathbf{k})^* \quad \text{if } k_x < 0. \tag{7}$$

Benefits of this neural representation are shown on 2-dimensional signals in the Supp. C.

The neural representation is queried for a set of L^2 3D-coordinates $[k_x, k_y, k_z]$, thereby producing a discretized slice $\mathcal{S}_i[\hat{V}_\theta] \in \mathbb{C}^{L\times L}$. The rest of the image

formation model (6) is simulated by element-wise multiplying $\mathcal{S}_i[\hat{V}_\theta]$ by the CTF C_i and a translation matrix,

$$\hat{X}_i = \hat{T}_{\mathbf{t}_i} \odot C_i \odot \mathcal{S}_i[\hat{V}_\theta], \tag{8}$$

where $\hat{T}_{\mathbf{t}_i}$ is defined by

$$\hat{T}_{\mathbf{t}_i}(\mathbf{k}) = \exp\left(-2j\pi \mathbf{k} \cdot \mathbf{t}_i\right). \tag{9}$$

The parameters of the CTF are provided by external CTF estimation softwares such as CTFFIND [40]. The whole encoder–decoder pipeline can be seen as a function that we call $\Gamma_{\xi,\theta}$, such that $\hat{X}_i = \Gamma_{\xi,\theta}(Y_i)$.

3.5 Symmetric Loss

In the image formation model of Eq. (3), the additive noise η_i is assumed to be Gaussian and uncorrelated (white Gaussian noise) [53,44], which means that its Fourier transform $\hat{\eta}_i$ follows the same kind of distribution. Therefore, maximum likelihood estimation on a batch $\mathcal{B}$ amounts to the minimization of the L2-loss.

Nonetheless, we empirically observed that using this loss often led the model to get stuck in local minima where the estimated volume showed spurious planar symmetries (see Sec. 4.3). We hypothesize that this behaviour is linked to the fundamental ambiguity contained in the image formation model in which, given unknown poses, one cannot distinguish two "mirrored" versions of the same volume [42]. We discuss this hypothesis in more detail in Supp. D. To solve this problem, we designed a loss that we call "symmetric loss" defined as

$$\mathcal{L}_{\text{sym}} = \sum_{i \in \mathcal{B}} \min\left\{ \|\hat{Y}_i - \Gamma_{\xi,\theta}(Y_i)\|^2, \|\mathcal{R}_\pi[\hat{Y}_i] - \Gamma_{\xi,\theta}\left(\mathcal{R}_\pi\left[Y_i\right]\right)\|^2 \right\} \tag{10}$$

where $\mathcal{R}_\pi$ applies an in-plane rotation of π on $L \times L$ images. Using the symmetric loss, the model can be supervised on a set of images Y_i in which the predicted in-plane rotation (embedded in the predicted matrix R_i) can always fall in $[-\pi/2, \pi/2]$ instead of $[-\pi, \pi]$. As shown in Sec. 4.3 and explained in Supp. D, this prevents cryoAI from getting stuck in spuriously symmetrical states.

4 Results

We qualitatively and quantitatively evaluate cryoAI for *ab initio* reconstruction of both simulated and experimental datasets. We first compare cryoAI to the state-of-the-art method cryoSPARC [37] in terms of runtime on a simulated dataset of the *80S* ribosome with low levels of noise. We then compare our method with baseline methods in terms of resolution and pose accuracy on simulated datasets with and without noise (*spike*, *spliceosome*). Next, we show that cryoAI can perform *ab initio* reconstruction on an experimental cryo-EM dataset (*80S*), which is the first time for a method estimating poses in an amortized fashion. Finally, we highlight the importance of a tailored neural representation in the decoder and the role of the symmetric loss in an ablation study.

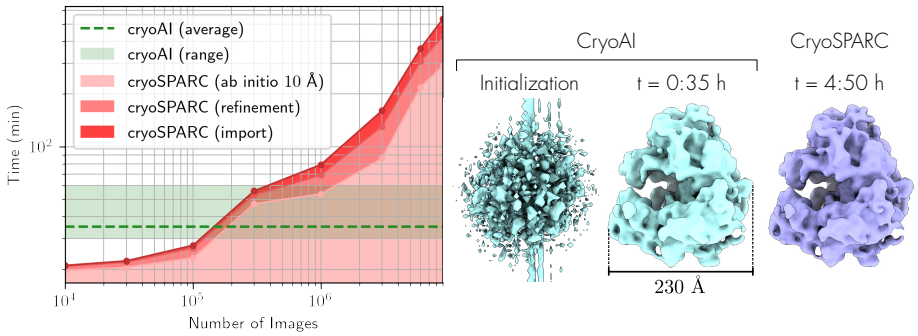


Fig. 3. (Left) Time to reach 10 Å of resolution with cryoAI (range and average over 5 runs per datapoint) and cryoSPARC vs. number of images in the simulated 80S dataset. (Right) Estimated volume at initialization and after 35 min of running cryoAI vs. cryoSPARC after convergence, with 9M images.

4.1 Reconstruction on Simulated Datasets

Experimental Setup. We synthesize three datasets from deposited Protein Data Bank (PDB) structures of the Plasmodium falciparum 80S ribosome (PDB: 3J79 and 3J7A) [56], the SARS-CoV-2 spike protein (PDB: 6VYB) [54] and the pre-catalytic spliceosome (PDB: 5NRL) [34]. First, a 3D grid map, the *ground-truth volume*, is generated in ChimeraX [15] from each atomic model using the steps described in Supp. A. Then a dataset is generated from the ground-truth volume using the image formation model described in Sec. 3.1. Images are sampled at $L = 128$. Rotations R_i are randomly generated following a uniform distribution over $SO(3)$ and random translations $\mathbf{t}_i$ are generated following a zero-mean Gaussian distribution ($\sigma = 20$ Å). The defocus parameters of the CTFs are generated with a log-normal distribution. We build noise-free (ideal) and noisy versions of each dataset (SNR = 0dB for 80S, SNR = -10dB for the others, see Supp. E for details). We compare cryoAI with three baselines: the state-of-the-art software cryoSPARC v3.2.0 [37] with default settings, the neural network–based method cryoDRGN2 [61] and the autoencoder-based method cryoPoseNet [30] (with the image formation model in real space in the decoder, see Supp. A). We quantify the accuracy of the reconstructed volume by computing the Fourier Shell Correlations (FSC) between the reconstruction and the ground truth and reporting the resolution at the 0.5 cutoff. All experiments are run on a single Tesla V100 GPU with 8 CPUs.

Convergence Time. We compare cryoAI with cryoSPARC in terms of runtime for datasets of increasing size in Fig. 3. We use the simulated 80S dataset and define the running time as the time needed to reach a resolution of 10 Å (2.65 pixels), which is a sufficiently accurate resolution to perform refinement with cryoSPARC (see workflow in Supp. A). With default parameters, cryoSPARC's ab initio reconstruction must process all images in the dataset. We show the time required by cryoSPARC for importing data and for the refine-

Table 1. Accuracy of pose and volume estimation for simulated data. Resolution (Res.) is reported using the FSC = 0.5 criterion, in pixels (↓). Rotation (Rot.) error is the median square Frobenius norm between predicted and ground truth matrices R_i (↓). Translation (Trans.) error is the mean square L2-norm, in pixels (↓).

Dataset		cryoPoseNet	cryoSPARC	cryoDRGN2	cryoAI
Spliceosome (ideal)	Res.	2.78	**2.13**	—	**2.13**
	Rot.	0.004	**0.0002**	—	0.0004
	Trans.	—	0.006	—	**0.001**
Spliceosome (noisy)	Res.	3.15	**2.61**	—	**2.61**
	Rot.	0.01	**0.002**	—	0.007
	Trans.	—	**0.007**	—	0.01
Spike (ideal)	Res.	16.0	2.33	—	**2.29**
	Rot.	5	0.0003	**0.0001**	0.0003
	Trans.	—	0.007	—	**0.001**
Spike (noisy)	Res.	16.0	3.56	**2.03**	2.91
	Rot.	6	0.02	**0.01**	**0.01**
	Trans.	—	0.008	—	**0.003**

ment step. CryoAI processes images batch-wise and does not statically associates variables to each image, making the convergence time (for reaching the specified resolution) independent from the size of the dataset. By contrast, the computation time of cryoSPARC increases with the number of images and can reach 5 hours with a dataset of 9M particles. We additionally show in Supp F the time required to estimate all the poses of the dataset with cryoAI's encoder.

Accuracy. We compare cryoAI with baseline methods on the *spike* and *spliceosome* datasets in Table 1. We compare the reconstructed variables (volume and poses) with their ground truth values (from simulation). Results of cryoDRGN2 are reported from available data in [59]. Images were centered for cryoPoseNet since the method does not predict $\mathbf{t}_i$. A "tight" adaptive mask was used with cryoSPARC. The performance of cryoAI is comparable with the baselines. The splicesome and the noise-free spike protein are reconstructed with state-of-the-art accuracy. In the noisy spike dataset, the accuracy of cryoAI and cryoSPARC decreases, which may be due to the pseudo-symmetries shown by the molecule (visual reconstruction in Supp. F). CryoPoseNet gets stuck for at least 24 hours in a state where the the resolution is very poor on both spike datasets.

4.2 Reconstruction on Experimental Datasets

Experimental Setup. We use the publicly available 80S experimental dataset EMPIAR-10028 [56,58,19] containing 105,247 images of length $L = 360$ (1.34 Å per pixel), downsampled to $L = 256$. The dataset is evenly split in two, each method runs independent reconstructions on each half and the FSC are measured between the two reconstructions. We compare cryoAI with cryoPoseNet and cryoSPARC. The dataset fed to cryoAI and cryoPoseNet is masked with

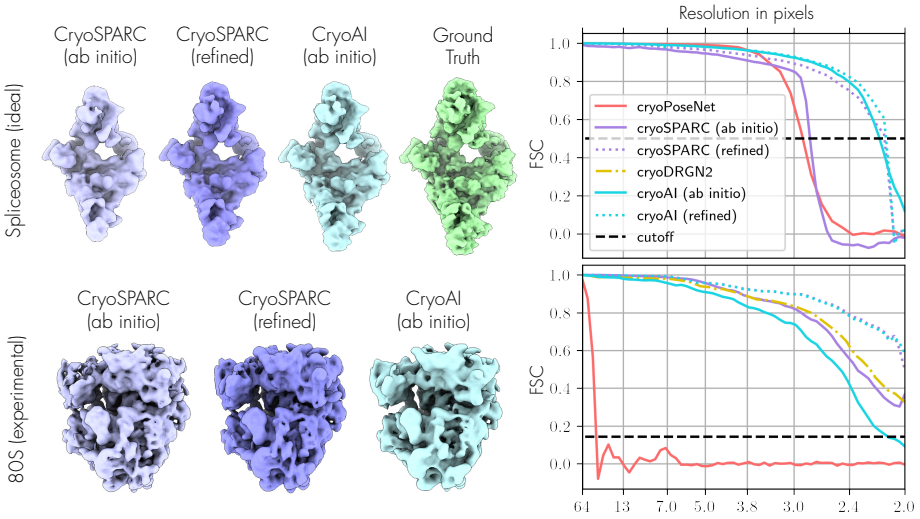


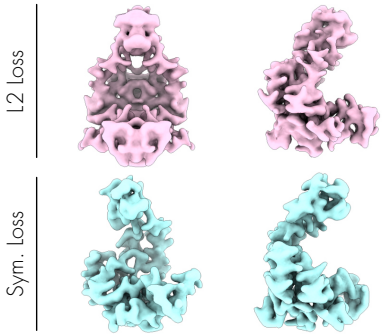
Fig. 4. (Top left) Volume reconstruction on a noise-free simulated dataset of the spliceosome ($L = 128$, pixel size = 4.25 Å). (Bottom left) Volume reconstruction for the experimental 80S dataset ($L = 128$, pixel size = 3.77 Å). (Right) Fourier Shell Correlations, reconstruction-to-ground-truth (top) or reconstruction-to-reconstruction (bottom). A resolution of 2.0 pixels corresponds to the Nyquist frequency. CryoAI can be refined using the software cryoSPARC.

a circular mask of radius 84 pixels, while cryoSPARC adaptively updates a "tight" mask. CryoAI and cryoPoseNet reconstruct a volume of size 128^3. For cryoSPARC, both the *ab initio* volume and the volume subsequently homogeneously refined from it were downsampled to the same size 128^3. We also demonstrate the possibility of refining cryoAI's output with the software cryoSPARC. Finally, we report the results published for cryoDRGN2 [61] that were obtained on a filtered version of the same dataset [58] downsampled to $L = 128$ prior reconstruction.

Results. We report quantitative and qualitative results in Fig. 4. CryoAI is the first amortized method to demonstrate proper volume reconstruction on an experimental dataset, although techniques predicting poses with an orientation-matching step (like cryoDRGN2) or followed by an EM-based refinement step (like cryoSPARC) can reach slightly higher resolutions. State-of-the-art results can be obtained with cryoSPARC's refinement, initialized from either cryoSPARC's or cryoAI's *ab initio*. Since simulated datasets were built using the same image formation model as the one cryoAI uses in its decoder, the gap in performance between the experimental and simulated datasets suggests that improvements could potentially be achieved with a more accurate physics model.

Ablation Study on Symmetrized Loss

Method/Loss		Time	Rot. Error
cryoPoseNet	L2	4:35 h	0.01
	Sym.	0:30 h	0.01
cryoAI	L2	1:35 h	0.005
	Sym.	**0:09 h**	**0.004**



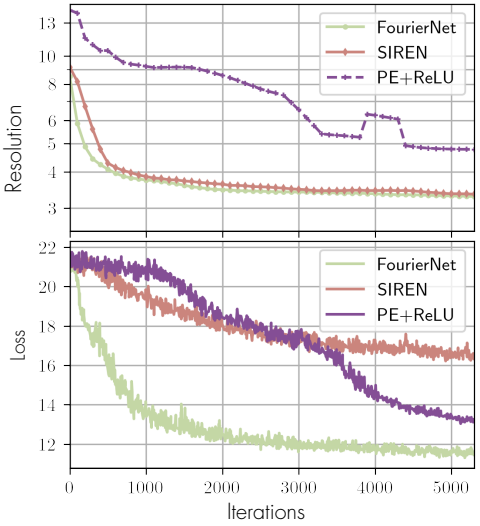


Fig. 5. (Top left) Ablation study on the symmetric loss with cryoAI and cryoPoseNet with simulated noise-free adenylate kinase ($L = 64$). We report the minimal convergence time out of 5 runs. CryoPoseNet is always slower and achieves worse results. The symmetric loss always accelerates convergence. (Bottom left) Volume reconstruction when using a L2 loss vs. the symmetric loss. The latter prevents the model from getting stuck in a symmetrical local minimum. (Right) Loss and resolution (in pixels, FSC = 0.143 cutoff) vs. number of iterations with a FourierNet, a SIREN [48] and an MLP with ReLU activation functions and positional encoding (32 images per batch).

4.3 Ablation Study

Importance of Symmetric Loss. The purpose of the symmetric loss is to prevent the model from getting stuck in local minima where the volume shows incorrect planar symmetries. Ullrich *et al.* showed in [50] that optimizing the poses using a gradient-based method often leads the model to fall in sub-optimal minima, due to the high non-convexity of the optimization problem. In [61], Zhong *et al.* implemented an autoencoder-based method (dubbed PoseVAE), and compared it to cryoDRGN2. The method is unable to properly reconstruct a synthetic hand, and a spurious planar symmetry appears in their reconstruction. We use a noisy dataset ($L = 128$) generated from a structure of Adenylate kinase (PDB 4AKE) [28]. We show in Fig. 5 that our method presents the same kind of artifact when using a L2 loss and validate that the symmetric loss prevents these artifacts. In Fig. 5, we compare our method to cryoPoseNet with and without the symmetric loss on a simulated ideal dataset of the same molecule ($L = 64$). Both methods use an autoencoder-based architecture and both converge significantly faster with the symmetric loss. With the same loss, cryoAI is always faster than cryoPoseNet since our method operates in Fourier space and avoids the approximation of integrals using the FST.

Comparison of Neural Representations. We replaced FourierNet with other neural representations in the decoder and compared the convergence rate of these models on the noisy Adenylate kinase dataset (L=128). In Fig. 5, we compare our architecture with a multi-layer perceptron (MLP) with sinusoidal activation functions (i.e., a SIREN [48]) and an MLP with ReLU activation function and positional encoding, as used by cryoDRGN2 [61]. We keep approximately 300k differentiable parameters in all representations. FourierNet significantly outperforms the two other architectures in terms of convergence speed.

5 Discussion

The amount of collected cryo-EM data is rapidly growing [29], which increases the need for efficient *ab initio* reconstruction methods. CryoAI proposes a framework of amortized inference to meet this need by having a complexity that does not grow with the size of the dataset. Since CryoAI jointly estimates volume and poses, it can be followed by reconstruction methods that address conformational heterogeneities, such as the ones available in cryoSPARC [37], RELION [44], or cryoDRGN [59]. The ever increasing size of cryo-EM datasets is necessary to provide sufficient sampling of conformational heterogeneities with increasing accuracy, in particular when imaging molecules that display complex dynamics. However, existing methods that tackle the more complex inference task of heterogeneous reconstruction also see their runtime suffer as datasets grow bigger, again showing the need for new developments that leverage amortized inference.

Future work on cryoAI includes adding features to the image formation model implemented in the decoder. CTFs, for example, are currently only characterized by three parameters (two defoci parameters and an astigmatism angle) but could be readily enhanced to account for higher-order effects (see e.g. [64]). A richer noise model, currently assumed to be Gaussian and white, could also improve the performance of the algorithm. In order to tackle the case of very noisy experimental datasets, adaptive masking techniques, such as those used by cryoSPARC, could be beneficial. In terms of hardware development, cryoAI would benefit from being able to run on more than a single GPU using data parallelism and/or model parallelism, thereby improving both runtime and efficiency. CryoAI, as described here, belongs to the class of homogeneous reconstruction methods; future developments should explore its performance in an heterogenous reconstruction setting, where conformational heterogeneity is baked in the generative model and the encoder is enhanced to predict descriptions of conformational states in low-dimensional latent space along with the poses.

Acknowledgment. We thank Wah Chiu for numerous discussions that helped shape this project. This work was supported by the U.S. Department of Energy, under DOE Contract No. DE-AC02-76SF00515. N.M. acknowledges support from the National Institutes of Health (NIH), grant No. 1R01GM144965-01. We acknowledge the use of the computational resources at the SLAC Shared Scientific Data Facility (SDF).

References

1. Akçakaya, M., Yaman, B., Chung, H., Ye, J.C.: Unsupervised deep learning methods for biological image reconstruction and enhancement: An overview from a signal processing perspective. IEEE Signal Processing Magazine **39**, 28–44 (2022)
2. Atzmon, M., Lipman, Y.: Sal: Sign agnostic learning of shapes from raw data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2565–2574 (2020)
3. Baker, T.S., Cheng, R.H.: A model-based approach for determining orientations of biological macromolecules imaged by cryoelectron microscopy. Journal of Structural Biology **116**, 120–130 (1996)
4. Baldwin, P.R., Tan, Y.Z., Eng, E.T., Rice, W.J., Noble, A.J., Negro, C.J., Cianfrocco, M.A., Potter, C.S., Carragher, B.: Big data in cryoEM: automated collection, processing and accessibility of em data. Current Opinion in Microbiology **43**, 1–8 (2018)
5. Bendory, T., Bartesaghi, A., Singer, A.: Single-particle cryo-electron microscopy: Mathematical theory, computational challenges, and opportunities. IEEE signal processing magazine **37**, 58–76 (2020)
6. Bepler, T., Kelley, K., Noble, A.J., Berger, B.: Topaz-denoise: general deep denoising models for cryoEM and cryoET. Nature communications **11**, 1–12 (2020)
7. Bracewell, R.N.: Strip integration in radio astronomy. Australian Journal of Physics **9**, 198–217 (1956)
8. Candes, E., Romberg, J., Tao, T.: Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Transactions on Information Theory **52**, 489–509 (2006)
9. Chen, M., Ludtke, S.J.: Deep learning-based mixed-dimensional Gaussian mixture model for characterizing variability in cryo-EM. Nature Methods **18**, 930–936 (2021)
10. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum Likelihood from Incomplete Data Via the EM Algorithm. Journal of the Royal Statistical Society: Series B (Methodological) **39**, 1–22 (1977)
11. Donnat, C., Levy, A., Poitevin, F., Miolane, N.: Deep Generative Modeling for Volume Reconstruction in Cryo-Electron Microscopy. arXiv: 2201.02867 (2022)
12. Donoho, D.: Compressed sensing. IEEE Transactions on Information Theory **52**, 1289–1306 (2006)
13. Genova, K., Cole, F., Vlasic, D., Sarna, A., Freeman, W.T., Funkhouser, T.: Learning shape templates with structured implicit functions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7154–7164 (2019)
14. Gershman, S., Goodman, N.: Amortized inference in probabilistic reasoning. In: Proceedings of the annual meeting of the cognitive science society. vol. 36 (2014)
15. Goddard, T.D., Huang, C.C., Meng, E.C., Pettersen, E.F., Couch, G.S., Morris, J.H., Ferrin, T.E.: Ucsf chimerax: Meeting modern challenges in visualization and analysis. Protein Science **27**, 14–25 (2018)
16. Greenberg, I., Shkolnisky, Y.: Common lines modeling for reference free ab-initio reconstruction in cryo-EM. Journal of Structural Biology **200**, 106–117 (2017)
17. Gupta, H., McCann, M.T., Donati, L., Unser, M.: CryoGAN: A New Reconstruction Paradigm for Single-Particle Cryo-EM Via Deep Adversarial Learning. IEEE Transactions on Computational Imaging **7**, 759–774 (2021)
18. Hertle, A.: On the Problem of Well-Posedness for the Radon Transform. In: Herman, G.T., Natterer, F. (eds.) Mathematical Aspects of Computerized Tomography. pp. 36–44. Springer (1981)

19. Iudin, A., Korir, P., Salavert-Torres, J., Kleywegt, G., Patwardhan., A.: Empiar: A public archive for raw electron microscopy image data. Nature Methods **13**, 387–388 (2016)
20. Kingma, D.P., Welling, M.: An introduction to variational autoencoders. arXiv preprint arXiv:1906.02691 (2019)
21. Kühlbrandt, W.: The resolution revolution. Science **343**, 1443–1444 (2014)
22. Lawler, E.L., Wood, D.E.: Branch-and-bound methods: A survey. Operations research **14**, 699–719 (1966)
23. Lian, R., Huang, B., Wang, L., Liu, Q., Lin, Y., Ling, H.: End-to-end orientation estimation from 2D cryo-EM images. Acta Crystallographica Section D: Structural Biology **78**, 174–186 (2022)
24. Mallick, S., Agarwal, S., Kriegman, D., Belongie, S., Carragher, B., Potter, C.: Structure and View Estimation for Tomographic Reconstruction: A Bayesian Approach. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). vol. 2, pp. 2253–2260 (2006)
25. Michalkiewicz, M., Pontes, J.K., Jack, D., Baktashmotlagh, M., Eriksson, A.: Implicit surface representations as layers in neural networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4743–4752 (2019)
26. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: European conference on computer vision. pp. 405–421. Springer (2020)
27. Miolane, N., Poitevin, F., Li, Y.T., Holmes, S.: Estimation of Orientation and Camera Parameters from Cryo-Electron Microscopy Images with Variational Autoencoders and Generative Adversarial Networks. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 4174–4183. IEEE (2020)
28. Müller, C., Schlauderer, G., Reinstein, J., Schulz, G.E.: Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding. Structure **4**, 147–156 (1996)
29. Namba, K., Makino, F.: Recent progress and future perspective of electron cryomicroscopy for structural life sciences. Microscopy **71**, i3–i14 (2022)
30. Nashed, Y.S.G., Poitevin, F., Gupta, H., Woollard, G., Kagan, M., Yoon, C.H., Ratner, D.: CryoPoseNet: End-to-End Simultaneous Learning of Single-Particle Orientation and 3D Map Reconstruction From Cryo-Electron Microscopy Data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 4066–4076 (2021)
31. Nogales, E.: The development of cryo-em into a mainstream structural biology technique. Nature Methods **13**, 24–27 (2016)
32. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: Deepsdf: Learning continuous signed distance functions for shape representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 165–174 (2019)
33. Penczek, P.A., Grassucci, R.A., Frank, J.: The ribosome at improved resolution: new techniques for merging and orientation refinement in 3D cryo-electron microscopy of biological particles. Ultramicroscopy **53**, 251–270 (1994)
34. Plaschka, C., Lin, P.C., Nagai, K.: Structure of a pre-catalytic spliceosome. Nature **546**, 617–621 (2017)
35. Pragier, G., Shkolnisky, Y.: A common lines approach for ab-initio modeling of cyclically-symmetric molecules. Inverse Problems **35**, 124005 (2019)
36. Punjani, A., Fleet, D.J.: 3d flexible refinement: structure and motion of flexible proteins from cryo-em. BioRxiv (2021)

37. Punjani, A., Rubinstein, J.L., Fleet, D.J., Brubaker, M.A.: cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. Nature Methods **14**, 290–296 (2017)
38. Ravi, N., Reizenstein, J., Novotny, D., Gordon, T., Lo, W.Y., Johnson, J., Gkioxari, G.: Accelerating 3D Deep Learning with PyTorch3D. arXiv: 2007.08501 (2020)
39. Renaud, J.P., Chari, A., Ciferri, C., ti Liu, W., Rémigy, H.W., Stark, H., Wiesmann, C.: Cryo-em in drug discovery: achievements, limitations and prospects. Nature Reviews Drug Discovery **17**, 471–492 (2018)
40. Rohou, A., Grigorieff, N.: Ctffind4: Fast and accurate defocus estimation from electron micrographs. Journal of structural biology **192**, 216–221 (2015)
41. Rosenbaum, D., Garnelo, M., Zielinski, M., Beattie, C., Clancy, E., Huber, A., Kohli, P., Senior, A.W., Jumper, J., Doersch, C., Eslami, S.M.A., Ronneberger, O., Adler, J.: Inferring a Continuous Distribution of Atom Coordinates from Cryo-EM Images using VAEs. arXiv:2106.14108 (Jun 2021)
42. Rosenthal, P.B., Henderson, R.: Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. Journal of molecular biology **333**, 721–745 (2003)
43. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D: Nonlinear Phenomena **60**, 259–268 (1992)
44. Scheres, S.H.: RELION: Implementation of a Bayesian approach to cryo-EM structure determination. Journal of Structural Biology **180**, 519–530 (2012)
45. Sigworth, F.J.: A maximum-likelihood approach to single-particle image refinement. Journal of Structural Biology **122**, 328–339 (1998)
46. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
47. Singer, A., Coifman, R.R., Sigworth, F.J., Chester, D.W., Shkolnisky, Y.: Detecting consistent common lines in cryo-EM by voting. Journal of Structural Biology **169**, 312–322 (2010)
48. Sitzmann, V., Martel, J., Bergman, A., Lindell, D., Wetzstein, G.: Implicit neural representations with periodic activation functions. Advances in Neural Information Processing Systems **33**, 7462–7473 (2020)
49. Sitzmann, V., Zollhöfer, M., Wetzstein, G.: Scene representation networks: Continuous 3d-structure-aware neural scene representations. Advances in Neural Information Processing Systems **32** (2019)
50. Ullrich, K., Berg, R.v.d., Brubaker, M., Fleet, D., Welling, M.: Differentiable probabilistic models of scientific imaging with the fourier slice theorem. arXiv preprint arXiv:1906.07582 (2019)
51. Vainshtein, B., Goncharov, A.: Determination of the spatial orientation of arbitrarily arranged identical particles of unknown structure from their projections. In: Soviet Physics Doklady. vol. 31, p. 278 (1986)
52. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
53. Vulović, M., Ravelli, R.B.G., van Vliet, L.J., Koster, A.J., Lazić, I., Lücken, U., Rullgård, H., Öktem, O., Rieger, B.: Image formation modeling in cryo-electron microscopy. Journal of Structural Biology **183**, 19–32 (2013)
54. Walls, A.C., Park, Y.J., Tortorici, M.A., Wall, A., McGuire, A.T., Veesler, D.: Structure, function, and antigenicity of the sars-cov-2 spike glycoprotein. Cell **181**, 281–292 (2020)

55. Wang, L., Singer, A., Wen, Z.: Orientation Determination of Cryo-EM Images Using Least Unsquared Deviations. SIAM Journal on Imaging Sciences **6**, 2450–2483 (2013)
56. Wong, W., Bai, X.c., Brown, A., Fernandez, I.S., Hanssen, E., Condron, M., Tan, Y.H., Baum, J., Scheres, S.H.: Cryo-em structure of the plasmodium falciparum 80s ribosome bound to the anti-protozoan drug emetine. Elife **3**, e03080 (2014)
57. Zehni, M., Donati, L., Soubies, E., Zhao, Z.J., Unser, M.: Joint Angular Refinement and Reconstruction for Single-Particle Cryo-EM. IEEE Transactions on Image Processing **29**, 6151–6163 (2020)
58. Zhong, E.: cryodrgn-empiar (2022), `https://github.com/zhonge/cryodrgn_empiar`
59. Zhong, E.D., Bepler, T., Berger, B., Davis, J.H.: CryoDRGN: reconstruction of heterogeneous cryo-EM structures using neural networks. Nature Methods **18**, 176–185 (2021)
60. Zhong, E.D., Bepler, T., Davis, J.H., Berger, B.: Reconstructing continuous distributions of 3D protein structure from cryo-EM images. arXiv:1909.05215 (2019)
61. Zhong, E.D., Lerer, A., Davis, J.H., Berger, B.: 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. pp. 4066–4075 (2021)
62. Zhong, E.D., Lerer, A., Davis, J.H., Berger, B.: Exploring generative atomic models in cryo-EM reconstruction. arXiv:2107.01331 (2021)
63. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the Continuity of Rotation Representations in Neural Networks. arXiv: 1812.07035 (2020)
64. Zivanov, J., Nakane, T., Forsberg, B.O., Kimanius, D., Hagen, W.J., Lindahl, E., Scheres, S.H.: New tools for automated high-resolution cryo-em structure determination in relion-3. elife **7**, e42166 (2018)