This ICCV workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

CryoPoseNet: End-to-End Simultaneous Learning of Single-particle Orientation and 3D Map Reconstruction from Cryo-electron Microscopy Data

Youssef S. G. Nashed
Machine Learning Initiative, SLAC National Accelerator Laboratory
Menlo Park, CA, USA
ynashed@slac.stanford.edu

Frederic Poitevin
Department of LCLS Data Analytics, SLAC National Accelerator Laboratory
Menlo Park, CA, USA
frederic.poitevin@stanford.edu

Harshit Gupta
Machine Learning Initiative, SLAC National Accelerator Laboratory
Menlo Park, CA, USA

Geoffrey Woollard
Department of Computer Science, University of British Columbia
Vancouver, Canada

Michael Kagan
Fundamental Physics Directorate, SLAC National Accelerator Laboratory
Menlo Park, CA, USA

Chun Hong Yoon
Department of LCLS Data Analytics, SLAC National Accelerator Laboratory
Menlo Park, CA, USA

Daniel Ratner
Machine Learning Initiative, SLAC National Accelerator Laboratory
Menlo Park, CA, USA
dratner@slac.stanford.edu

Abstract

Cryogenic electron microscopy (cryo-EM) provides images from different copies of the same biomolecule in arbitrary orientations. Here, we present an end-to-end unsupervised approach that learns individual particle orientations directly from cryo-EM data while reconstructing the 3D map of the biomolecule following random initialization. The approach relies on an auto-encoder architecture where the latent space is explicitly interpreted as orientations used by the decoder to form an image according to the physical projection model. We evaluate our method on simulated data and show that it is able to reconstruct 3D particle maps from noisy- and CTF-corrupted 2D projection images of unknown particle orientations.

1. Introduction

Determining the structure of biomolecules is an important step towards understanding their functional mechanism or designing new drugs. Structural determination techniques such as nuclear magnetic resonance spectroscopy, X-ray diffraction (XRD) crystallography or cryogenic electron microscopy (cryo-EM) have been very successful over the years [1]. Of those techniques, cryo-EM has been used increasingly in recent years to determine structures of many biomolecules at near-atomic resolution. This revolution has been possible because of better hardware and software techniques [18]. Cryo-EM is the fastest growing technique in terms of structures deposited in the Protein Data Bank, projected to be comparable to XRD within a few years. While the typical resolution reported for cryo-EM structures is still worse than those reported for XRD structures, recent developments have proven possible to distinguish individual atoms in cryo-EM maps [25]. Whereas XRD simultaneously measures millions of copies of a molecule, cryo-EM is a single-particle imaging (SPI) technique, with each cryo-EM image corresponding to a single molecular copy. While SPI reconstruction methods must learn the conformation and orientation of individual particles, this challenge is also an opportunity: rather than learning an average structure, SPI methods can resolve biologically relevant dynamics and avoid blurring from averaging over heterogeneous particles [27]. In this paper, we discuss a new approach to solving the single-particle orientation problem using a neural-network auto-encoder.

1.1. Cryo-EM 3D Reconstruction

In cryo-EM, a large number $(10^4 - 10^7)$ of identical copies of the same biomolecule are first frozen in a thin vitreous layer of ice. A beam of electron then goes through this sample and is acquired by a detector. It is typically assumed that only the phase of the incoming beam has been changed by the electrostatic potential of the sample, not its amplitude. Since the sample is typically prepared so the material that the electrons go through is very thin, it is also usually assumed that the linear projection approximation can be used to model the image formation process. The resulting image is called a micrograph. Individual particles are located and extracted from the micrograph as square patches of identical size $N \times N$ resulting in a particle stack which constitutes the starting dataset for tomographic reconstruction of the particle $N \times N \times N$ volume. Tomographic reconstruction is an ill-posed problem that faces multiple challenges. The individual pose (orientation and translational shift) of each particle in the stack is not known and needs to be estimated from images that have been corrupted in two different ways. First, the image is convolved with the point-spread function (PSF) of the microscope which acts as a filter reducing the information content of each image across its spectrum. Second, the main part of the signal in the image comes from the surrounding ice which is a major source of noise, in addition to other sources of noise such as shot noise resulting from dose fractionation and detector response. Indeed a balance between radiation damage and getting signal at all needs to be found during data collection. These difficulties make cryo-EM 3D reconstruction a very challenging inverse problem.

1.2. Related work

Many methods have been developed to tackle cryo-EM tomographic reconstruction [43, 11, 37, 44, 14, 35, 6, 31], including common-lines approaches [47, 21, 41, 51, 10, 29, 53], projection-matching strategies [28, 2], or Bayesian formulations [7, 40, 35, 31]. The Bayesian approach has been popularized by the widely used software RELION [34] which performs maximum-a-posteriori (MAP) optimization through Expectation-Maximization (EM) to reconstruct a map which maximizes the likelihood of the acquired data while still meeting some *a priori* condition about the map. In EM, during the Expectation-step, a conditional distribution over the poses and shifts is estimated for each projection using the current estimate of the map. In the next Maximization-step, these estimated distributions are used to update the map. These two steps are performed iteratively until meeting a convergence criterion. MAP optimization does not guarantee global convergence and therefore its reconstruction is dependent on the quality of the initial map. First implementations of the method suffered from their lack of robustness to map initialization, so a competing software cryoSPARC [31] proposed optimizing a MAP solution *ab initio* through stochastic gradient descent (SGD) - however, in doing so, individual poses are not estimated explicitly, thereby limiting the achievable resolution. The resulting low-resolution map can then be further refined with the EM algorithm. Because the EM reconstruction approach requires estimation of conditional distribution on poses for each measurement, the number of variables to determine grows directly with the number of measurements.

Neural networks (NNs) with their state-of-the-art performance on various different inverse problems [42, 26, 19, 5, 52, 20, 22] have recently been introduced in the cryo-EM processing pipeline, mainly in pre-processing steps such as denoising of the micrograph [3] and particle picking [50, 55, 45, 49, 4]. More recently a few methods using NNs have been proposed to solve the cryo-EM reconstruction problem [54, 32, 12, 13, 30, 23]. These methods require no prior training and are fully unsupervised. In [54] a modified variational autoencoder (VAE) [16] is used to reconstruct continuous conformations of dynamic biomolecules. In [12, 13], generative adversarial networks (GANs) [9] are modified to reconstruct structure of biomolecules with single and continuous conformations, respectively. In [23] a

composition of GANs and VAE is used to find the latent variables that explain the data, which then can be further used for reconstruction. In contrast to these methods, [30] argues that amortized inference as done by VAEs does not yield results that are precise enough and instead perform direct inference of the conformational coordinate using an auto-decoder approach. None of these methods use NNs to estimate the poses for each projection. Instead they either use an external traditional pose estimation routine [54] or bypass the pose assignment step altogether [12, 13].

1.3. Contributions

In this paper, we propose CryoPoseNet, a method which uses a modified auto-encoder [48, 39, 38] to simultaneously reconstruct the 3D map and estimate individual poses. Our method is fully unsupervised, and does not require any prior training nor an *ab-initio* solution. We use a combination of an encoder parameterized by an NN and a cryo-EM physics based decoder parameterized by a learnable 3D structure. The encoder outputs the estimated pose for a given input measurement, which is then fed to the decoder that outputs the simulated cryo-EM projection from the current structure. The encodings are constrained to correspond to the pose by the physics-based decoder. The weights of the encoder and the 3D structure in the decoder are simultaneously optimized in order to decrease the error between the given projection (input of the encoder) and the output of the decoder.

We leverage this auto-encoder approach to skip the EM-like step in traditional methods for estimating poses. As opposed to an EM approach, the number of encoder parameters is fixed and we do not require explicit estimation of pose distribution for each measurement. Therefore, the number of variables to be estimated in our method does not grow with the number of measurements.

2. Background and Current Methods

The reconstruction problem of cryo-EM requires estimation of the structure $\mathbf{x} \in \mathbb{R}^{N\times N\times N}$ from the measurements $\{\mathbf{y}_1, \dots, \mathbf{y}_M\}$. The acquisition of each measurement $\mathbf{y}_m \in \mathbb{R}^{N\times N}$ can be modeled as

$$\mathbf{y}_m = \underbrace{C_{\mathbf{d}_m} * S_{\mathbf{t}_m}\{P_{\boldsymbol{\theta}_m}\{\mathbf{x}\}\}}_{\mathbf{H}_{\boldsymbol{\varphi}_m}} + \mathbf{n}_m, \tag{1}$$

where $\mathbf{n}_m \in \mathbb{R}^{N\times N}$ is the additive noise. The imaging operator $\mathbf{H}_{\boldsymbol{\varphi}_m}$ depends on the imaging parameters $\boldsymbol{\varphi}_m = (\boldsymbol{\theta}_m, \boldsymbol{t}_m, \boldsymbol{d}_m) \in \mathbb{R}^8$. The imaging operator consists of the projection operator $P_{\boldsymbol{\theta}^m}$ which outputs the tomographic projection of the structure rotated by the Euler angles $\boldsymbol{\theta}_m = (\theta_{m,1}, \theta_{m,2}, \theta_{m,3})$; the shift operator $S_{\mathbf{t}_m}$, which shifts the projection by $\mathbf{t}_m = (t_{m,1}, t_{m,2})$ consisting of horizontal and vertical directions, respectively; and the convolution operator $C_{\mathbf{d}_m}$ which corrupts the image by the contrast transfer function (CTF) with parameters $\mathbf{d}_m = (d_{m,1}, d_{m,2}, \alpha_m)$ that consists of major defocus, minor defocus, and astigmatism angle, respectively.

As discussed earlier, the imaging parameters, $\boldsymbol{\theta}_m$ and $\boldsymbol{t}_m$, are unknown for each measurement. This coupled with the loss of information due to the CTF, and the high level of noise, makes the reconstruction of $\mathbf{x}$ a challenging problem.

Maximum Likelihood. A naive approach to solving the reconstruction problem would consist in searching for the unknown imaging parameters for each measurement and a global structure which maximizes the likelihood of the measurements. This quest can be described as

$$\mathbf{x}_{\text{rec}}, \tilde{\boldsymbol{\varphi}}_1, \dots, \tilde{\boldsymbol{\varphi}}_m = \arg \max_{\mathbf{x}, \tilde{\boldsymbol{\varphi}}_1, \dots, \tilde{\boldsymbol{\varphi}}_M} \sum_{m=1}^{M} \log p(\mathbf{y}_m | \mathbf{x}, \tilde{\boldsymbol{\varphi}}_m), \tag{2}$$

$$= \arg \max_{\mathbf{x}, \tilde{\boldsymbol{\varphi}}_1, \dots, \tilde{\boldsymbol{\varphi}}_M} \sum_{m=1}^{M} \|\mathbf{y}_m - \mathbf{H}_{\tilde{\boldsymbol{\varphi}}_m}\mathbf{x}\|^2, \tag{3}$$

where $p(\mathbf{y}_m|\mathbf{x}, \tilde{\boldsymbol{\varphi}}_m)$ denotes the likelihood of the measurement $\mathbf{y}_m$ given the imaging parameter $\tilde{\boldsymbol{\varphi}}_m$ and the structure $\mathbf{x}$. For an independent white Gaussian noise model, it takes the form (3).

However, since the formulation (3) is highly non-convex and filled with poor local minima [46], this naive approach would rarely give a reasonable solution. A brute-force approach to solving the reconstruction problem would follow an iterative procedure where instead of estimating all the parameters at the same time, the structure would be updated using the current estimates of the poses which would be updated in turn using the current estimate of the structure. At iteration k, this is given by

$$\tilde{\boldsymbol{\varphi}}_{m,k} = \arg \max_{\tilde{\boldsymbol{\varphi}}_m} \log p(\mathbf{y}_m | \mathbf{x}_k, \tilde{\boldsymbol{\varphi}}_m) \forall m \in [1, \dots, M], \tag{4}$$

$$\mathbf{x}_{k+1} = \arg \max_{\mathbf{x}} \log p(\mathbf{y}_m | \mathbf{x}, \tilde{\boldsymbol{\varphi}}_{k,m}) \tag{5}$$

This approach would not guarantee to find a global optimum either and the quality of the reconstruction would be highly dependent on the initialization. Moreover, the high level of noise makes the pose estimation error-prone.

Maximum Marginalized Likelihood. To remedy these issues, current methods use a marginalized likelihood formulation which instead of estimating a single pose for each measurement, effectively weighs the contribution of the poses from the whole search space. This is given by Eq.(6) where $p(\mathbf{y}_m|\mathbf{x})$ denotes the probability of the projection

given the structure.

$$\mathbf{x}_{\text{rec}} = \arg\max_{\mathbf{x}} \sum_{m=1}^{M} \log p(\mathbf{y}_m|\mathbf{x}), \tag{6}$$
$$= \arg\max_{\mathbf{x}} \sum_{m=1}^{M} \log \int p(\mathbf{y}_m|\mathbf{x}, \boldsymbol{\varphi}) p(\boldsymbol{\varphi}) \, \mathrm{d}\boldsymbol{\varphi},$$

Formulation (6) can be solved in two stages. First, a reasonable approximation to the global maximum can be found through SGD or some other method. Next, this *ab-initio* structure $\mathbf{x}_0$ is used to initialize an iterative EM procedure (also called iterative refinement) which will refine the solution further, at the cost of doing more operations for pose estimation.

The expectation step of the k-th iteration estimates a conditional distribution on the space of poses for each projection given the current estimate of the structure $\mathbf{x}_k$. This estimate is given by

$$p(\boldsymbol{\varphi}|\mathbf{x}_k, \mathbf{y}_m) = \frac{p(\mathbf{y}_m|\mathbf{x}_k, \boldsymbol{\varphi}) p(\boldsymbol{\varphi})}{\int_{\boldsymbol{\varphi}} p(\mathbf{y}_m|\mathbf{x}_k, \boldsymbol{\varphi}) p(\boldsymbol{\varphi}) \, \mathrm{d}\boldsymbol{\varphi}}. \tag{7}$$

The maximization step then uses these poses to update the structures by solving

$$\mathbf{x}_{k+1} = \arg\max_{\mathbf{x}} \sum_{m=1}^{M} \mathbb{E}_{p(\boldsymbol{\varphi}|\mathbf{x}_k, \mathbf{y}_m)} [\log p(\mathbf{y}_m|\mathbf{x}, \boldsymbol{\varphi})]. \tag{8}$$

For feasibility, the space of $\boldsymbol{\varphi}$ is discretized to compute the integrals in (7) and (8). This weighted form of pose estimation is less sensitive to the initial reference and yields better quality reconstructions. In most methods, prior knowledge over the structure is used to obtain a modified MAP formulation. However, the traditional approaches just described are computationally intensive procedures because estimating poses or conditional distributions over them against an ever changing reference structure scales poorly - the number of variables to estimate grows directly with the size of the data.

3. Proposed Method

To solve the scaling problem, we propose a neural network based representation of poses where we consider the unknown poses $\tilde{\boldsymbol{\varphi}}_m$ as the output of a tunable function E_γ for a given input measurement (as defined in (9)). In this work, we consider the shifts $\mathbf{t}_m$ and defocus parameters $\mathbf{d}_m$ known, but in principle they could also be added to E_γ outputs.

For a general error function R, we solve (10) which becomes (11) (similar to (3)) when a Gaussian noise model is assumed

$$\tilde{\boldsymbol{\varphi}}_m = E_\gamma(\mathbf{y}_m) \quad \forall m \in [1, \ldots, M]. \tag{9}$$

$$\mathbf{x}_{\text{rec}} = \arg\max_{\mathbf{x},\gamma} \sum_{m=1}^{M} R(y_m, \mathbf{H}_{E_\gamma(\mathbf{y}_m)}\mathbf{x}). \tag{10}$$

$$= \arg\max_{\mathbf{x},\gamma} \sum_{m=1}^{M} \|\mathbf{y}_m - \mathbf{H}_{E_\gamma(\mathbf{y}_m)}\mathbf{x}\|^2. \tag{11}$$

We minimize (10) using SGD as described in Algorithm 1. At each iteration, for a batch of measurements $\{\mathbf{y}_1, \ldots, \mathbf{y}_B\}$, we get the empirical estimate of the loss function in (10) by

$$L(\mathbf{x}, \gamma) = \sum_{b=1}^{B} R(y_b, \mathbf{H}_{E_\gamma(\mathbf{y}_b)}\mathbf{x}). \tag{12}$$

The structure and the encoder weights are optimized using the gradients $\nabla_{\mathbf{x}} L(\mathbf{x}, \gamma)$ and $\nabla_\gamma L(\mathbf{x}, \gamma)$, respectively. In summary, simultaneously with structure estimation we tune the weights of a neural network so that the imaging parameters maximize the likelihood of the measurements. Our scheme is shown in Figure 1.

Algorithm 1 Reconstruct cryo-EM data

Input: acquired dataset $\{\mathbf{y}_1, \ldots, \mathbf{y}_M\}$; number of reconstruction iterations, n_{rec}; size of the batches used for SGD, B; optimizer parameters; Error function R;
Initialization: $\mathbf{x}_{\text{rec}}$ with a uniform random distribution, and E_γ with random weights

1: **for** n_{rec} **do**
2: From acquired dataset, sample a batch$\{\mathbf{y}^1_{\text{batch}}, \ldots, \mathbf{y}^B_{\text{batch}}\}$
3: Obtain $\tilde{\boldsymbol{\varphi}}_b = E_\gamma(\mathbf{y}^b_{\text{batch}}) \forall b \in [1, \ldots, B]$
4: Compute $L(\mathbf{x}, \gamma) = \sum_{b=1}^{B} R(\mathbf{y}^b_{\text{batch}}, \mathbf{H}_{\tilde{\boldsymbol{\varphi}}_b}\mathbf{x}_{\text{rec}})$
5: Update $\mathbf{x}_{\text{rec}}$ using $\nabla_{\mathbf{x}} L(\mathbf{x}, \gamma)$
6: update γ using $\nabla_\gamma L_(\mathbf{x}, \gamma)$

Output: $\mathbf{x}_{\text{rec}}, \tilde{\boldsymbol{\varphi}}_1, \ldots, \tilde{\boldsymbol{\varphi}}_M$.

Encoder. The encoder NN is composed of a convolutional neural network (CNN) followed by a multilayer percepetron (MLP), also known as a fully-connected layer. The CNN extracts shift-invariant features from the measurements, which are then transformed by the MLP into orientation parameters as described below. We use a standard CNN encoder architecture of a *conv2D* $\rightarrow$ *conv2D* $\rightarrow$ *maxpool* block repeated three times. All convolution layers use 3×3 filters, and maxpooling downsamples by a factor of 2. The number of convolution filters per block is $32, 64, 128$, respectively, and the MLP is composed of two layers, each containing 512 units/neurons.

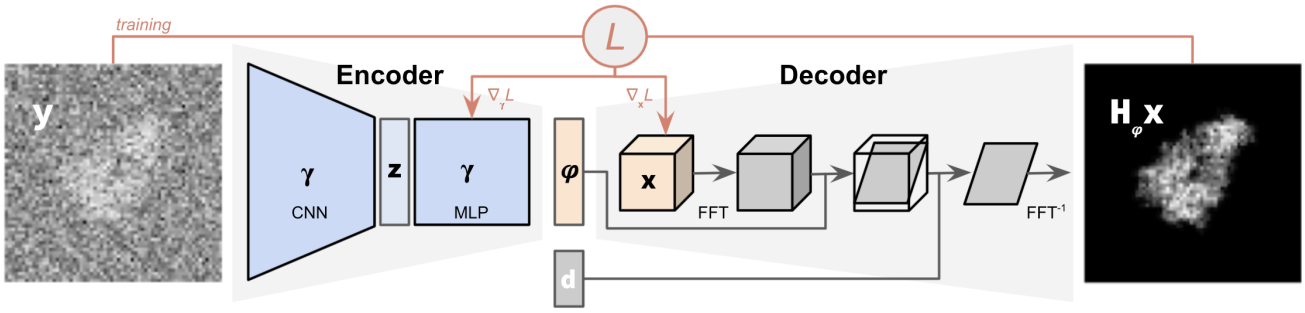

Figure 1. **CryoPoseNet architecture**.The input image $\mathbf{y}$ is encoded into a pose $\boldsymbol{\varphi}$ by the encoder. The decoder outputs the tomographic projection of the structure oriented at φ, and then corrupted by the CTF. The image formation is implemented using the Fourier-slice theorem, with the estimated 3D structure $\mathbf{x}$ Fourier transformed, sliced into a plane normal to $\boldsymbol{\varphi}$, and corrupted by a known CTF with parameter $\mathbf{d}$. The parameters of the encoder and decoder are respectively the weights $\boldsymbol{\gamma}$ of the CNN and MLP and the structure $\mathbf{x}$. They are optimized during reconstruction through minimization of the loss L that measures the dissimilarity between the input image $\mathbf{y}$ and the reconstructed one $\mathbf{H}_{\boldsymbol{\varphi}}\mathbf{x}$.

Differentiable cryo-EM physics model. The cryo-EM physics operator $\mathbf{H}_{\boldsymbol{\varphi}}$ uses the output of the encoder as the pose parameters $\boldsymbol{\varphi}$. In order to compute the gradients $\nabla_{\mathbf{x}} L(\mathbf{x}, \gamma)$ and $\nabla_{\gamma} L(\mathbf{x}, \gamma)$, the operator $\mathbf{H}_{\boldsymbol{\varphi}}$ needs to be differentiable with respect to the structure as well as the imaging parameters. We therefore implement a differentiable cryo-EM physics model which enables the learning of encoder weights using backpropagation. The image formation model is implemented in reciprocal space, in which the the 3D Fourier transform of $\mathbf{x}$ is computed once per batch. Predicted/Decoded projections are generated by the 2D inverse Fourier transform of a slice in the 3D Fourier volume extracted using the corresponding predicted pose. Since the encoder is optimized using backpropagation, the parameterization of the poses directly affects the encoder's performance and thereby, the quality of the reconstruction. We use the following parameterizations in our experiments:

- Euler Angles: For each measurement, the encoder yields a 3-dimensional output which is then fed to the imaging model as Euler angles (commonly used in cryo-EM software). The structure is then rotated using these angles.

- Quaternions: The three-dimensional MLP output is transformed into a 4-dimensional vector using a fixed transformation which has the properties of a unit quaternion $\mathbf{q} = (q_1, q_2, q_3, q_4)$. The imaging model then rotates the structure by $2\cos^{-1} q_1$, around the axis (q_2, q_3, q_4).

- $\mathcal{S}^2 \times \mathcal{S}^2$ (s2s2): The MLP outputs two 3-dimensional vectors. These are then orthonormalized to obtain $\mathbf{w}_1$ and $\mathbf{w}_2$. A third vector is then computed by the cross product of the first two, $\mathbf{w}_3 = \mathbf{w}_1 \times \mathbf{w}_2$. These three vectors define a local coordinate system that relates to a 3×3 rotation matrix. This matrix is then used to rotate the structure. Formally, this case is an algebraic parameterization of the Lie group of 3D rotations SO(3), using two orthonormal vectors $\mathbf{w}_1$ and $\mathbf{w}_2$.

NN pose estimation offers many advantages. First, instead of searching for individual independent poses $\{\boldsymbol{\varphi}_m\}$, we estimate a global function that maps the measurements to their respective poses. Since this mapping is global, it indirectly integrates the information from all the projections for each pose estimation. Second, since the information of poses has been condensed in the fixed size γ, the number of variables to estimate does not grow with the size of the data. Finally, the universal approximation property of the neural networks lets us learn complicated mappings between projections and their poses.

Simulation of datasets. All the experiments presented here used datasets generated from the atomic model 4AKE[1] of *E.coli* adenylate kinase [24], a small 47 kDa protein. TEM simulator [33] was used to generate a $128 \times 128 \times 128$ electrostatic potential map with pixel size 0.8 Å. Images associated with pose φ were obtained by rotating the centered map with φ and projecting it along the z-direction in real space after resampling on the original grid. To account for CTF corruption, the resulting image was Fourier transformed, padded to double length, multiplied by a pre-computed 2D CTF image with given defocus d. Gaussian white noise is added *a posteriori*. For each dataset, 10,000 images were generated by sampling SO(3) uniformly, out of which 9,000 images were used for reconstruction and the rest 1,000 were kept to assess the quality of pose estimation by the encoder. When relevant, the CTF range [0.4 µm, 1.2 µm] is sampled uniformly. The

[1] https://www.rcsb.org/structure/4AKE

noise is sampled from the normal distribution and scaled to match the desired signal-to-noise ratio (SNR). Table 1 lists the datasets generated.

Table 1. Datasets used in the numerical experiments. All datasets are comprised of a training set of 9,000 images, a held-out set of 1,000 images, with images being made of 128×128 pixels of size 0.8 Å. Abbreviations: (SNR) Signal-to-Noise Ratio. (CTF) Contrast Transfer Function. (MAE) Mean Absolute Error. (RMSE) Root Mean Square Error.

	Parameters		Performance		
	SNR (dB)	CTF (μm)	Res. (Å)	MAE (degrees)	RMSE ($\times 10^{-5}$)
A	-	-	1.79	0.62	6.2
B	10	-	2.13	0.69	7.3
C	5	-	2.28	0.84	8.7
D	0	-	2.43	0.86	10.2
E	-5	-	2.56	2.04	13.2
F	-10	-	2.78	3.20	17.5
G	0	$\mathcal{U}(0.4, 1.2)$	2.57	1.12	15.9

Reconstruction. For all the described datasets, reconstruction was carried with the following parameters: we use the Adam (Adaptive Moment Estimation) [15] optimizer with a learning rate of 5×10^{-4} , minibatch size $B = 32$, number of minibatch update steps $n_{\text{rec}} = 10,000$ which is equivalent to approximately 35.5 epochs or passes over the whole dataset. The framework is implemented using Tensorflow and runs on an NVIDIA Tesla V100 GPU in around 5 hours for a full reconstruction run.

4. Results

We evaluated the performance of our method on simulated datasets. We detail below how the methods performed in the absence of noise, in the presence of increasing noise, and in a more realistic setting with noise and CTF corruption.

SO(3) parameterization. In the absence of noise (dataset A), reconstruction converged in a few thousand steps, as can be seen from the loss curve on Fig.2-A. Following a similar convergence pattern, the MAE between estimated and ground truth poses φ decreased drastically in the first few hundred steps (see Fig.2-B). Three different approach to SO(3) parameterization have been tried, of which the Euler representation converges the slowest and s2s2 the fastest and ultimately most accurate. On close inspection (data not shown), the main reason for this discrepancy is attributed to parameterization singularities arising from both the Euler and quaternion formulations that could affect the numerical stability of gradients computed under such discontinuous representations. For a more in depth proof of those formulations the reader is referred to these publications [8, 17]. In all the subsequent experiments we adopt s2s2 as the chosen orientation parameterization.

Comparison to Automatic differentiation. To provide points of comparisons with previous work, we considered the following two scenarios. First, we monitored the performance of our approach when poses were known (see *tomo* curve in Fig.2). In this case, as expected, the network converges almost instantly to what we consider a lower bound for the loss. We also considered the case where the poses and map were solved through stochastic gradient descent, similar to [31]. In this case, the loss hardly decreases over a few thousand steps of optimization, mainly due to attempting to solve for orientation and structure simultaneously, which is difficult without careful initialization of the solved variables or alternating the updates between them. Additionally, as mentioned earlier, the number of orientation variables to be solved by AD scales with the dataset size.

Reconstruction quality. To measure the reconstruction quality of the estimated map $\mathbf{x}$, we compute the Fourier shell correlation (FSC) every 200 steps between the current and the ground truth map (see Fig.3-A). In the first few hundred steps, the FSC curve intersects the 0.5 cutoff at resolutions worse than 3.2 Å and converges rapidly to ∼1.8 Å which is close to the Nyquist-Shannon limit of 1.6 Å.

Effect of noise on reconstruction. We tested the ability of our method to handle realistic noise typically encountered in cryoEM data. Figure 3-B summarizes our findings. As expected, adding noise is detrimental to the quality of the reconstruction. Yet, even at realistic value of -10 dB we see that the effective resolution of the reconstruction is still better than 2.8 Å. Visual inspection of the reconstructed maps (see Fig.4) is consistent with this observation.

Effect of CTF. Finally, we tested the ability of our method to reconstruct images that were both degraded by added noise (0 dB) and by CTF corruption (dataset G). The FSC curve of the resulting reconstruction is very similar to the one obtained without CTF corruption (see Fig.3-C). However, the shape differs slightly, with higher correlation in the highest resolution shells and lower correlation in the medium-resolution range. Visual inspection of the resulting map is consistent with this observation (see Fig.4).

Limits of the current implementation. In order to disentangle the various factors that could lead to a deterioration of the FSC curve in response to added noise or CTF corruption, as observed in Fig.3, we carried out the reconstruction of Datasets A and G with known orientations. The

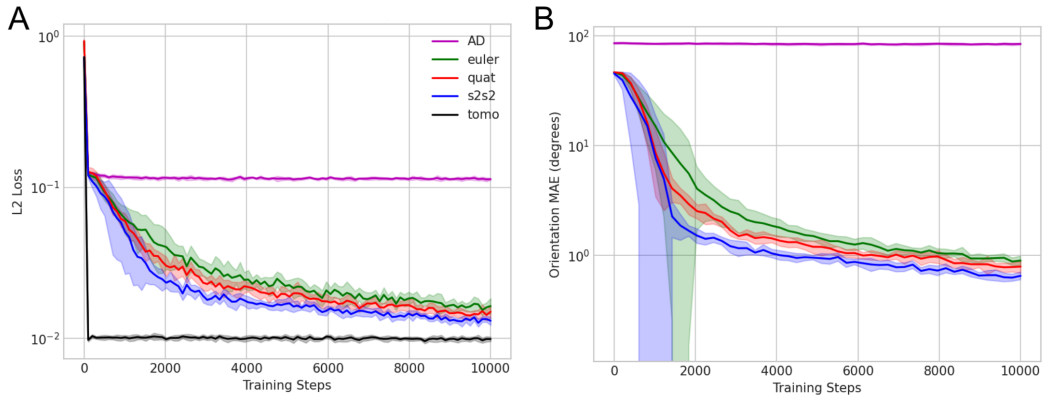

Figure 2. **SO(3) parameterization and convergence**. **(A)** Projection 2D L2 loss convergence for various representations of SO(3): Euler angles (euler), quaternions (quat) and s2s2. Each corresponds to 10 independent reconstruction runs (average and standard deviation). For comparison, the loss curves obtained through automatic differentiation (AD) or with known poses (tomo) are also shown. **(B)** Evolution of the mean absolute error (MAE) over the poses.

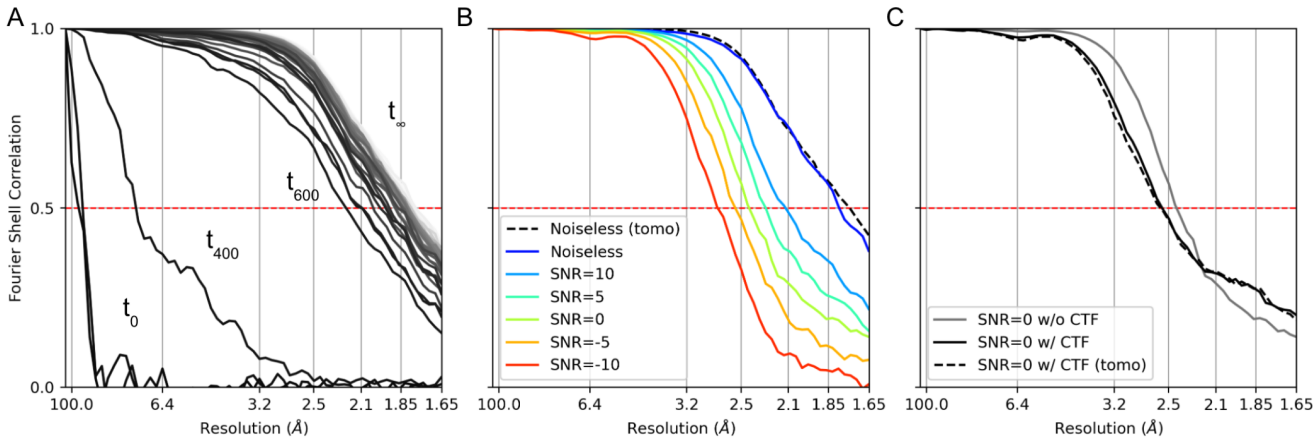

Figure 3. **Fourier-Shell Correlation to ground truth**. The dotted red line indicates the FSC threshold used to estimate the map resolution. **(A)** FSC curves at regular intervals, from black to white, during optimization in the absence of noise (dataset A). **(B)** FSC curves for final models at various noise levels (datasets B-F), in the absence of CTF corruption. Signal-to-Noise Ratio (SNR) is given in dB. The *tomo* baseline represented with a dotted line corresponds to reconstrution with known poses. **(B)** FSC curves for final models at zero SNR with (dataset D) and without CTF (dataset G) corruption of the dataset. The *tomo* baseline represented with a dotted line corresponds to reconstrution with known poses

resulting FSC curves are denoted "tomo" in Fig.3-B,C. Interestingly, the resulting reconstructions are very close to their counterpart where the orientations had to be learned, suggesting that pose estimation is not a limitation in our framework. This is also supported by the mean absolute error measured on the pose estimates, summarized in Table 1 where in the worst case angles are off by a few degrees. This result is expected mainly due to the simplicity of the implemented forward model. The current forward model relies on tri-linear interpolation to extract a 2D slice from the 3D Fourier volume, which can cause projection artifacts. Ideally, appropriate interpolation kernels and Fourier gridding/resampling techniques [36] should be employed to reduce such artifacts. We are currently developing more realistic image formation models that can be directly incorporated in the presented framework.

5. Conclusion

In this paper we presented a new method for solving the pose estimation problem in cryoEM using an autoencoder architecture employing a differentiable forward model decoder. We showed, on simulated data, that the method is able to reconstruct high-resolution 3D map from simulated data in the presence of noise or CTF corruption and without any prior structural knowledge. We have recently become aware of contemporaneous work using a VAE for estima-

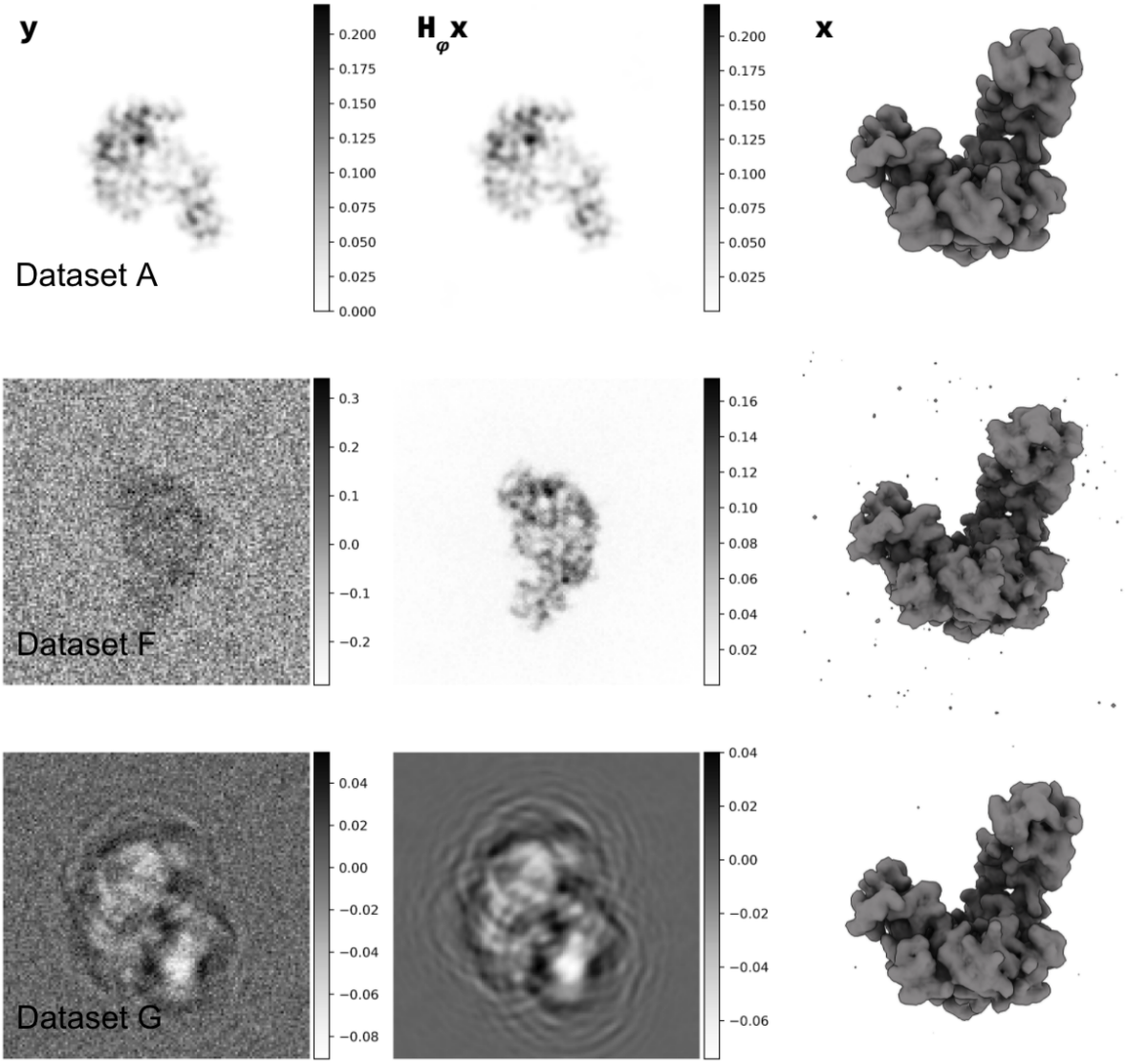

Figure 4. **Reconstructions**. The left column shows samples from their respective datasets (A-top, F-center, G-bottom). The center column shows the final reconstructed image. The right column shows the reconstructed map at the same contour level.

tion of pose and conformation[32]. While this approach also uses NNs to estimate poses, it shares with [54, 30] the limitation that it relies on a consensus structure initialization, the quality of which might impact the ability of the methods to reliably learn poses and conformations, in particular for highly flexible systems. We hypothesize that the methods presented here would provide a solution to this issue. Because the computational cost is independent of the number of images, we also envision this method to be the basis of a new pipeline for the large datasets expected to be needed to resolve continuous conformations.

Acknowledgments

We thank Wah Chiu, Khanh Dao Duc, Mark Hunter, TJ Lane, Julien Martel, Nina Miolane, and Gordon Wetzstein for numerous discussions that helped shape this project. We also thank Takanori Nakane for his suggestions that helped strengthen our results presentation. This work was supported by the U.S. Department of Energy, under DOE Contract No. DE-AC02-76SF00515. We acknowledge the use of the computational resources at the SLAC Shared Scientific Data Facility (SDF). MK is supported by the US Department of Energy (DOE) under grant DE-AC02-76SF00515, and by the SLAC Panofsky Fellowship.

References

[1] A celebration of structural biology. *Nature Methods*, 18(5):427–427, May 2021. 2

[2] Timothy S. Baker and R. Holland Cheng. A Model-Based Approach for Determining Orientations of Biological Macromolecules Imaged by Cryoelectron Microscopy. *Journal of Structural Biology*, 116(1):120–130, Jan. 1996. 2

[3] Tristan Bepler, Kotaro Kelley, Alex J. Noble, and Bonnie Berger. Topaz-Denoise: general deep denoising models for cryoEM and cryoET. *Nat Commun*, 11(1):5208, Oct. 2020. Number: 1 Publisher: Nature Publishing Group. 2

[4] Tristan Bepler, Andrew Morin, Micah Rapp, Julia Brasch, Lawrence Shapiro, Alex J. Noble, and Bonnie Berger. Positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs. *Nat Methods*, 16(11):1153–1160, Nov. 2019. Number: 11 Publisher: Nature Publishing Group. 2

[5] Navid Borhani, Eirini Kakkava, Christophe Moser, and Demetri Psaltis. Learning to see through multimode fibers. *Optica*, 5(8):960–966, Aug 2018. 2

[6] J. M. de la Rosa-Trevín, A. Quintana, L. del Cano, A. Zaldívar, I. Foche, J. Gutiérrez, J. Gómez-Blanco, J. Burguet-Castell, J. Cuenca-Alba, V. Abrishami, J. Vargas, J. Otón, G. Sharov, J. L. Vilas, J. Navas, P. Conesa, M. Kazemi, R. Marabini, C. O. S. Sorzano, and J. M. Carazo. Scipion: A software framework toward integration, reproducibility and validation in 3D electron microscopy. *Journal of Structural Biology*, 195(1):93–99, July 2016. 2

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data Via the *EM* Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, Sept. 1977. 2

[8] Luca Falorsi, Pim de Haan, Tim R Davidson, Nicola De Cao, Maurice Weiler, Patrick Forré, and Taco S Cohen. Explorations in homeomorphic variational auto-encoding. *arXiv preprint arXiv:1807.04689*, 2018. 6

[9] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. *arXiv:1406.2661 [cs, stat]*, June 2014. arXiv: 1406.2661. 2

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

[11] Nikolaus Grigorieff. FREALIGN: High-resolution refinement of single particle structures. *Journal of Structural Biology*, 157(1):117–125, Jan. 2007. 2

[12] Harshit Gupta, Michael T. McCann, Laurène Donati, and Michael Unser. CryoGAN: A New Reconstruction Paradigm for Single-Particle Cryo-EM via Deep Adversarial Learning. *bioRxiv*, page 2020.03.20.001016, July 2020. Publisher: Cold Spring Harbor Laboratory Section: New Results. 2, 3

[13] Harshit Gupta, Thong H. Phan, Jaejun Yoo, and Michael Unser. Multi-CryoGAN: Reconstruction of Continuous Conformations in Cryo-EM Using Generative Adversarial Networks. In Adrien Bartoli and Andrea Fusiello, editors, *Computer Vision – ECCV 2020 Workshops*, Lecture Notes in Computer Science, pages 429–444, Cham, 2020. Springer International Publishing. 2, 3

[14] Michael Hohn, Grant Tang, Grant Goodyear, P. R. Baldwin, Zhong Huang, Pawel A. Penczek, Chao Yang, Robert M. Glaeser, Paul D. Adams, and Steven J. Ludtke. SPARX, a new environment for Cryo-EM image processing. *Journal of Structural Biology*, 157(1):47–55, Jan. 2007. 2

[15] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*, Jan. 2017. arXiv: 1412.6980. 6

[16] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. *arXiv:1312.6114 [cs, stat]*, May 2014. arXiv: 1312.6114. 2

[17] Shoshichi Kobayashi and Katsumi Nomizu. *Foundations of differential geometry*, volume 1. New York, London, 1963. 6

[18] Werner Kühlbrandt. The Resolution Revolution. *Science*, 343(6178):1443–1444, Mar. 2014. Publisher: American Association for the Advancement of Science Section: Perspective. 2

[19] Shuai Li, Mo Deng, Justin Lee, Ayan Sinha, and George Barbastathis. Imaging through glass diffusers using densely connected convolutional networks. *Optica*, 5(7):803–813, Jul 2018. 2

[20] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, Dec. 2017. 2

[21] S.P. Mallick, S. Agarwal, D.J. Kriegman, S.J. Belongie, B. Carragher, and C.S. Potter. Structure and View Estimation for Tomographic Reconstruction: A Bayesian Approach. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2253–2260, June 2006. ISSN: 1063-6919. 2

[22] Michael T. McCann, Kyong Hwan Jin, and Michael Unser. Convolutional Neural Networks for Inverse Problems in Imaging: A Review. *IEEE Signal Processing Magazine*, 34(6):85–95, Nov. 2017. Conference Name: IEEE Signal Processing Magazine. 2

[23] Nina Miolane, Frédéric Poitevin, Yee-Ting Li, and Susan Holmes. Estimation of Orientation and Camera Parameters from Cryo-Electron Microscopy Images with Variational Autoencoders and Generative Adversarial Networks. *arXiv:1911.08121 [cs, eess, q-bio, stat]*, May 2021. arXiv: 1911.08121. 2

[24] C. W. Müller, G. J. Schlauderer, J. Reinstein, and G. E. Schulz. Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding. *Structure*, 4(2):147–156, Feb. 1996. 5

[25] Takanori Nakane, Abhay Kotecha, Andrija Sente, Greg McMullan, Simonas Masiulis, Patricia M. G. E. Brown, Ioana T. Grigoras, Lina Malinauskaite, Tomas Malinauskas, Jonas Miehling, Tomasz Uchański, Lingbo Yu, Dimple Karia, Evgeniya V. Pechnikova, Erwin de Jong, Jeroen

Keizer, Maarten Bischoff, Jamie McCormack, Peter Tiemeijer, Steven W. Hardwick, Dimitri Y. Chirgadze, Garib Murshudov, A. Radu Aricescu, and Sjors H. W. Scheres. Single-particle cryo-EM at atomic resolution. *Nature*, 587(7832):152–156, Nov. 2020. Number: 7832 Publisher: Nature Publishing Group. 2

[26] Thanh Nguyen, Yujia Xue, Yunzhe Li, Lei Tian, and George Nehmetallah. Deep learning approach for fourier ptychography microscopy. *Opt. Express*, 26(20):26470–26484, Oct 2018. 2

[27] Abbas Ourmazd. Cryo-EM, XFELs and the structure conundrum in structural biology. *Nat Methods*, 16(10):941–944, Oct. 2019. Number: 10 Publisher: Nature Publishing Group. 2

[28] Pawel A. Penczek, Robert A. Grassucci, and Joachim Frank. The ribosome at improved resolution: New techniques for merging and orientation refinement in 3D cryo-electron microscopy of biological particles. *Ultramicroscopy*, 53(3):251–270, Mar. 1994. 2

[29] Gabi Pragier and Yoel Shkolnisky. A common lines approach for ab-initio modeling of cyclically-symmetric molecules. *Inverse Problems*, 35(12):124005, Dec. 2019. arXiv: 1901.10888. 2

[30] Ali Punjani and David J. Fleet. 3D Flexible Refinement: Structure and Motion of Flexible Proteins from Cryo-EM. *bioRxiv*, page 2021.04.22.440893, Apr. 2021. Publisher: Cold Spring Harbor Laboratory Section: New Results. 2, 3, 8

[31] Ali Punjani, John L. Rubinstein, David J. Fleet, and Marcus A. Brubaker. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat Methods*, 14(3):290–296, Mar. 2017. Number: 3 Publisher: Nature Publishing Group. 2, 6

[32] Dan Rosenbaum, Marta Garnelo, Michal Zielinski, Charlie Beattie, Ellen Clancy, Andrea Huber, Pushmeet Kohli, Andrew W. Senior, John Jumper, Carl Doersch, S. M. Ali Eslami, Olaf Ronneberger, and Jonas Adler. Inferring a Continuous Distribution of Atom Coordinates from Cryo-EM Images using VAEs. *arXiv:2106.14108 [cs, eess]*, June 2021. arXiv: 2106.14108. 2, 8

[33] H. Rullgård, L.-G. Öfverstedt, S. Masich, B. Daneholt, and O. Öktem. Simulation of transmission electron microscope images of biological specimens. *Journal of Microscopy*, 243(3):234–256, 2011. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2818.2011.03497.x. 5

[34] Sjors H.W. Scheres. A Bayesian View on Cryo-EM Structure Determination. *J Mol Biol*, 415(2-4):406–418, Jan. 2012. 2

[35] Sjors H. W. Scheres. RELION: Implementation of a Bayesian approach to cryo-EM structure determination. *Journal of Structural Biology*, 180(3):519–530, Dec. 2012. 2

[36] Hermann Schomberg and Jan Timmer. The gridding method for image reconstruction by fourier transformation. *IEEE transactions on medical imaging*, 14(3):596–607, 1995. 7

[37] Tanvir R. Shaikh, Haixiao Gao, William T. Baxter, Francisco J. Asturias, Nicolas Boisset, Ardean Leith, and Joachim Frank. SPIDER image processing for single-particle reconstruction of biological macromolecules from electron micrographs. *Nat Protoc*, 3(12):1941–1974, Dec. 2008. Number: 12 Publisher: Nature Publishing Group. 2

[38] O. V. Shcherbakov, I. N. Zhdanov, and Ya. A. Lushin. A convolutional autoencoder as a generative model of images for problems of distinguishing attributes and restoring images in missing regions. *J. Opt. Technol.*, 82(8):528–532, Aug 2015. 3

[39] Tomoyoshi Shimobaba, Yutaka Endo, Ryuji Hirayama, Yuki Nagahama, Takayuki Takahashi, Takashi Nishitsuji, Takashi Kakue, Atsushi Shiraki, Naoki Takada, Nobuyuki Masuda, and Tomoyoshi Ito. Autoencoder-based holographic image restoration. *Appl. Opt.*, 56(13):F27–F30, May 2017. 3

[40] F. J. Sigworth. A Maximum-Likelihood Approach to Single-Particle Image Refinement. *Journal of Structural Biology*, 122(3):328–339, Jan. 1998. 2

[41] Amit Singer, Ronald R. Coifman, Fred J. Sigworth, David W. Chester, and Yoel Shkolnisky. Detecting consistent common lines in cryo-EM by voting. *Journal of Structural Biology*, 169(3):312–322, Mar. 2010. 2

[42] Ayan Sinha, Justin Lee, Shuai Li, and George Barbastathis. Lensless computational imaging through deep learning. *Optica*, 4(9):1117–1125, Sep 2017. 2

[43] C. O. S. Sorzano, R. Marabini, J. Velázquez-Muriel, J. R. Bilbao-Castro, S. H. W. Scheres, J. M. Carazo, and A. Pascual-Montano. XMIPP: a new generation of an open-source image processing package for electron microscopy. *Journal of Structural Biology*, 148(2):194–204, Nov. 2004. 2

[44] Guang Tang, Liwei Peng, Philip R. Baldwin, Deepinder S. Mann, Wen Jiang, Ian Rees, and Steven J. Ludtke. EMAN2: An extensible image processing suite for electron microscopy. *Journal of Structural Biology*, 157(1):38–46, Jan. 2007. 2

[45] Dimitry Tegunov and Patrick Cramer. Real-time cryo-electron microscopy data preprocessing with Warp. *Nat Methods*, 16(11):1146–1152, Nov. 2019. Number: 11 Publisher: Nature Publishing Group. 2

[46] Karen Ullrich, Rianne van den Berg, Marcus Brubaker, David Fleet, and Max Welling. Differentiable probabilistic models of scientific imaging with the Fourier slice theorem. *arXiv:1906.07582 [cs, eess, stat]*, June 2019. arXiv: 1906.07582. 3

[47] B. K. Vainshtein and A. B. Goncharov. Determination of the spatial orientation of arbitrarily arranged identical particles of unknown structure from their projections. *Soviet Physics Doklady*, 31:278, Apr. 1986. 2

[48] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010. 3

[49] Thorsten Wagner, Felipe Merino, Markus Stabrin, Toshio Moriya, Claudia Antoni, Amir Apelbaum, Philine Hagel, Oleg Sitsel, Tobias Raisch, Daniel Prumbaum, Dennis Quentin, Daniel Roderer, Sebastian Tacke, Birte Siebolds,

Evelyn Schubert, Tanvir R. Shaikh, Pascal Lill, Christos Gatsogiannis, and Stefan Raunser. SPHIRE-crYOLO is a fast and accurate fully automated particle picker for cryo-EM. *Commun Biol*, 2(1):1–13, June 2019. Number: 1 Publisher: Nature Publishing Group. 2

[50] Feng Wang, Huichao Gong, Gaochao Liu, Meijing Li, Chuangye Yan, Tian Xia, Xueming Li, and Jianyang Zeng. DeepPicker: A deep learning approach for fully automated particle picking in cryo-EM. *Journal of Structural Biology*, 195(3):325–336, Sept. 2016. 2

[51] Lanhui Wang, Amit Singer, and Zaiwen Wen. Orientation Determination of Cryo-EM Images Using Least Unsquared Deviations. *SIAM J. Imaging Sci.*, 6(4):2450–2483, Jan. 2013. Publisher: Society for Industrial and Applied Mathematics. 2

[52] Fangshu Yang, Thanh-an Pham, Harshit Gupta, Michael Unser, and Jianwei Ma. Deep-learning projector for optical diffraction tomography. *Optics express*, 28(3):3905–3921, 2020. 2

[53] Mona Zehni, Laurène Donati, Emmanuel Soubies, Zhizhen J. Zhao, and Michael Unser. Joint Angular Refinement and Reconstruction for Single-Particle Cryo-EM. *IEEE Trans. on Image Process.*, 29:6151–6163, 2020. arXiv: 2003.10062. 2

[54] Ellen D. Zhong, Tristan Bepler, Bonnie Berger, and Joseph H. Davis. CryoDRGN: reconstruction of heterogeneous cryo-EM structures using neural networks. *Nat Methods*, 18(2):176–185, Feb. 2021. Number: 2 Publisher: Nature Publishing Group. 2, 3, 8

[55] Yanan Zhu, Qi Ouyang, and Youdong Mao. A deep convolutional neural network approach to single-particle recognition in cryo-electron microscopy. *BMC Bioinformatics*, 18(1):348, July 2017. 2