Amortized Pose Estimation for X-Ray Single Particle Imaging

Jay Shenoy
Stanford University

Axel Levy
Stanford University

Frédéric Poitevin
SLAC National Accelerator Laboratory

Gordon Wetzstein
Stanford University

Abstract

X-ray single particle imaging (SPI) is a nascent technique that can capture the dynamics of biomolecules at room temperature. SPI experiments will one day collect tens of millions of images of the same molecule in order to overcome the weak scattering of individual proteins. Existing reconstruction algorithms will be unable to scale to datasets of this size because they perform computationally expensive search steps to estimate the orientation of the molecule in each image. In this work, we propose a reconstruction algorithm that amortizes the estimation of pose via an autoencoder framework. Our approach consists of a convolutional encoder that maps X-ray images to predicted poses and a physics-based decoder that implicitly fuses all the 2D scattering images into a volumetric representation of the molecule. We validate our method on 6 synthetic datasets of 2 distinct proteins, showing that for the largest datasets containing 5 million images, our technique can reconstruct the electron density in a single pass.

1 Introduction

Accurately determining the structures of biomolecules helps biologists better understand their function, which is important for modeling disease and developing new therapies. X-ray crystallography and cryo-electron microscopy (cryo-EM) are the leading methods routinely used to determine protein structures at atomic resolution. However, both methods limit the range of possible conformational states that can be captured – crystallization restricts protein movement by locking molecules in a lattice, while the process of freezing in cryo-EM prevents proteins from assuming states that lie at higher energies [1].

X-ray single particle imaging (SPI) is a nascent imaging technique that is being developed to circumvent some of the limitations of crystallography and cryo-EM. In X-ray SPI, individual aerosolized proteins are imaged at room temperature using an X-ray free-electron laser (XFEL). Because of the limited scattering cross-section of the imaged molecules, each image contains a small amount of signal, therefore millions of copies of the same protein need to be imaged at different orientations. The molecules are typically delivered to the imaging laser via aerosol jet, so each protein is captured at an unknown pose with respect to the detector. Reconstruction algorithms must therefore estimate the orientation of each image and merge all the images together into a 3D structure.

With free-electron lasers such as EuXFEL and LCLS-II expected to capture up to one million images per second [2], SPI reconstruction methods must be able to process datasets containing tens of millions of images. Unfortunately, existing algorithms for X-ray SPI [3, 4] do not scale well to large datasets, with pose estimation acting as the primary bottleneck. Recent work on reconstruction methods for cryo-electron microscopy (cryo-EM) has resolved this scaling issue through amortized

Machine Learning for Structural Biology Workshop, NeurIPS 2023.

inference [5, 6], a technique that avoids independently estimating the pose of each image. We first describe related work in the field of X-ray SPI and subsequently explain how recent developments in cryo-EM analysis inspire our own amortized method.

1.1 Reconstruction in X-Ray SPI

Current state-of-the-art reconstruction methods for X-ray SPI have found success in processing both real and synthetic datasets and are robust to low photon counts [3, 4, 7]. However, these techniques scale poorly for one of two reasons: either because (1) they perform a non-amortized pose estimation step that exhaustively searches the rotation group $SO(3)$ or (2) they compute dense similarity metrics between all pairs of images in the data.

One method of the former variety is the expand-maximize-compress (EMC) algorithm, which was initially developed by Loh and Elser [8] and later implemented in Dragonfly [3]. The EMC algorithm performs pose estimation in a probabilistic manner, iteratively maximizing the likelihood of the orientation of each image with respect to an estimate of the intensity volume until convergence. Multi-tiered iterative phasing (M-TIP) [4] is another algorithm that estimates orientations in a similar manner to EMC while also incorporating phase retrieval to bake real-space constraints into the optimization. Although EMC and M-TIP only require a constant number of iterations to converge, each iteration involves computing likelihoods for every possible pose across all the images. As such, the likelihood maximization step scales linearly in both the number of images as well as the space of all poses (a sampled subset of $SO(3)$), which becomes computationally expensive as the dataset grows in size. Another class of methods is based on the principle of manifold embedding using diffusion maps (DM). Unlike EMC and M-TIP, DM performs pose estimation by first learning a low-dimensional manifold from the space of images in the dataset and then mapping this manifold to the rotation group $SO(3)$ [9, 10]. Unfortunately, the DM algorithm scales poorly with dataset size because it requires computing distance metrics between all pairs of images in the dataset in order to construct a k-Nearest Neighbor Graph (k-NNG) [11].

1.2 Optimization-Based Amortized Inference in Cryo-EM

Recent work in reconstruction for cryo-EM has shown that end-to-end optimization with amortized inference can accelerate pose and conformation estimation. Rather than independently estimating the orientation of each image as done in EMC and M-TIP, amortized inference learns a function that maps images to these estimated variables. As a result, pose and conformation estimation scale relative to the complexity of the function parameterization instead of the size of the dataset. CryoDRGN [12] was the first technique to use amortized inference to determine the conformational states from experimental cryo-EM datasets with known poses, employing a neural implicit representation for structure determination. CryoDRGN2 [13] extended this method to incorporate pose estimation as well, but it utilizes a non-amortized exhaustive search strategy that slows down reconstruction. CryoAI [5] was the first method to successfully amortize pose estimation for homogeneous reconstruction on experimental data.

In this work, we introduce X-RAI, a method that amortizes the estimation of pose in a similar manner as cryoAI, employing a convolutional encoder and physics-based decoder for single particle reconstruction. Because our algorithm optimizes the encoder and decoder end-to-end via gradient descent, we avoid the expensive pose estimation step found in prior work. When the dataset is sufficiently large, amortization enables reconstruction to operate in an online fashion. Our method differs from cryoAI because its forward model is based on X-ray diffraction, an imaging technique that loses phase information and operates under fundamentally different physical principles than cryo-EM.

2 Methods

2.1 Image Formation Model

In X-ray SPI, each molecule being imaged possesses an electron density field that scatters the incident photons to form a diffraction image on the detector. The electron density can be formalized as a 3D function V that maps $\mathbb{R}^3$ to $\mathbb{R}$. According to the theory of X-ray diffraction [14], each image recorded

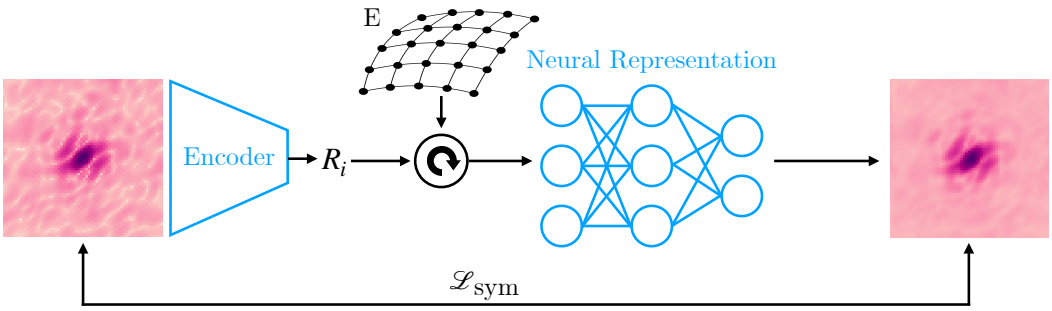

Figure 1: Visualization of X-RAI's reconstruction pipeline. The encoder takes in a Fourier amplitude image and predicts a rotation matrix R_i representing the molecule's orientation in the image. R_i rotates the coordinates of the Ewald sphere E, the result of which is used to query a neural implicit representation of the Fourier amplitude volume. This Ewald slice through the amplitude volume produces a noise-free estimate of the input amplitude image, which is then compared to the input image using a symmetric loss. The parameters optimized by the reconstruction procedure are shaded in blue, and the images are plotted on a log scale.

by the detector corresponds to a slice through the Fourier transform of V, where the coordinates of the slice lie on the Ewald sphere, a geometric object in Fourier space that is constructed based on the parameters of the imaging experiment. If we denote E as the coordinates along the Ewald sphere and $\hat{V} := F_{3D}[V]$ as the Fourier transform of V, then the resulting slice S can be formulated as:

$$S = \hat{V}(E) \tag{1}$$

Each detector image I_i captures a single molecule at a random orientation R_i with respect to the laser. Since rotation in the primal domain is equivalent to rotation in the Fourier domain, the resulting image can be constructed by simply rotating the coordinates of the Ewald slice as follows:

$$\hat{V}(R_i \cdot E) \tag{2}$$

Experimental detectors only capture the intensity and not the phase of the incident light. As such, the phase of the complex-valued $\hat{V}$ is lost, and the captured image corresponds to the squared magnitude of $\hat{V}$:

$$I_i = |\hat{V}(R_i \cdot E)|^2 = |\hat{V}|^2(R_i \cdot E) + \eta, \tag{3}$$

where η corresponds to additive noise (potentially signal dependent). Since the detector records discrete photon hits, the dominating source of noise in SPI experiments is Poissonian, particularly when the molecule is a weak-scatterer.

2.2 Overview of the Architecture

X-RAI employs an autoencoder architecture to reconstruct a diffraction volume from a set of X-ray diffraction images. Here, the diffraction volume refers to the real-valued function $|\hat{V}|^2$. Our pipeline, visualized in figure 1, starts by feeding each diffraction image to a convolutional encoder that outputs an estimate of the molecule's orientation. The coordinates of the Ewald slice are rotated by this orientation and are subsequently used to query a neural representation of the diffraction volume to yield a noise-free estimate of the input image. The input and output images are compared using a symmetric loss (first introduced in cryoAI [5]) that prevents spurious planar symmetries from arising during reconstruction, after which gradients are back-propagated to update both the encoder and decoder. The symmetric loss is reproduced below:

$$\mathcal{L}_{\text{sym}} = \sum_i \min\{||I_i - \Gamma(I_i)||^2, ||\mathcal{R}_\pi(I_i) - \Gamma(\mathcal{R}_\pi(I_i))||^2\}, \tag{4}$$

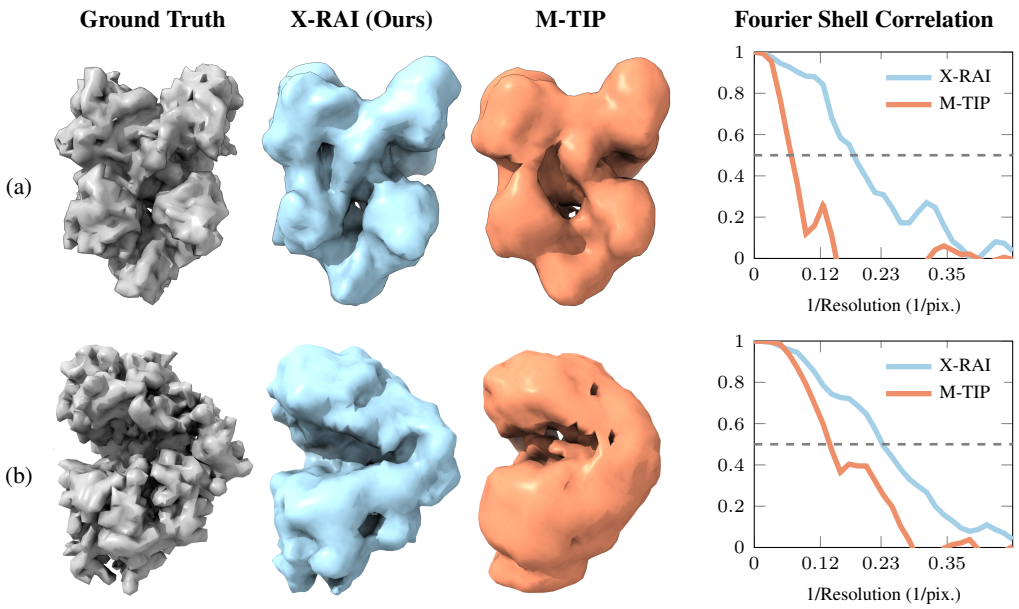

Figure 2: Comparison of the electron densities output by X-RAI and M-TIP against the ground truth volumes used for simulation. These reconstructions correspond to datasets containing 500K images with the following PDB codes: (a) 1O9K and (b) 2CEX_A. X-RAI achieves higher resolution than M-TIP for these datasets, which can be seen visually from the reconstruction quality as well as the FSC curves.

where Γ represents the operator corresponding to the autoencoder and $\mathcal{R}_\pi$ denotes a rotation of the input image by π radians. Here, the use of an L_2 loss assumes a Gaussian noise model as opposed to one that is Poissonian, which suffices for our experiments since the datasets we simulate are noise-free.

Pose Estimation. The encoder maps images to molecular poses using a convolutional neural network (CNN). First, we take the square root of each input diffraction image to produce a Fourier amplitude image that is fed as input to the CNN. In practice, we find that transforming the images in this way aids reconstruction by boosting the signal at higher frequencies, which naturally drops off in the Fourier spectra of natural images. The amplitude images are then low-pass filtered into a Gaussian pyramid before being input to a convolutional neural network inspired by VGG16 [15]. The output of this CNN is a feature vector that is subsequently fed as input to a fully-connected neural network that outputs an estimate of the pose, parameterized as a six-dimensional vector in $S^2 \times S^2$ [16].

Physics-Based Decoder. The decoder maps the pose estimate produced by the encoder to a noise-free estimate of the Fourier amplitude image. This mapping is performed by querying a neural implicit representation of the Fourier amplitude volume (the square root of the diffraction volume) with the Ewald slice coordinates E rotated by the estimated pose. Neural representations such as NeRF [17] and SIREN [18] have found broad success in signal representation for computer vision, and cryoDRGN [12] was the first method to employ such a representation for protein reconstruction in cryo-EM. To represent the Fourier amplitude, we utilize a variant of SIREN called FourierNet that is tailored to represent Fourier spectra [5]. The amplitude image produced by the decoder is compared against the input amplitude image using the symmetric loss in equation 4.

3 Results

We use Skopi [19], a software package for simulating X-ray SPI experiments, to generate datasets for two model proteins (PDB: 1O9K [20] and 2CEX_A [21]) at three different dataset sizes: 50,000,

Dataset	Method	Time	Resolution (pixels) ↓	Train Error (Med/MSE) ↓	Test Error (Med/MSE) ↓
1O9K (Train: 50K)	X-RAI	8:16h	4.73	1.5 / 2.2	1.5 / 2.2
	M-TIP	0:35h	7.20	9.2 / 16.4	9.5 / 16.4
1O9K (Train: 500K)	X-RAI	8:15h	**4.09**	**1.3 / 1.7**	**1.3 / 1.7**
	M-TIP	4:44h	6.90	9.4 / 16.5	10.3 / 17.1
1O9K (Train: 5M)	X-RAI	8:13h	4.41	2.0 / 2.7	1.9 / 2.6
	M-TIP	> 24h	—	—	—
2CEX_A (Train: 50K)	X-RAI	8:16h	5.15	2.0 / 3.2	2.1 / 3.3
	M-TIP	0:29h	11.52	53.4 / 72.6	53.1 / 74.2
2CEX_A (Train: 500K)	X-RAI	8:15h	**4.30**	1.5 / 1.9	**1.4 / 1.8**
	M-TIP	4:35h	7.32	23.1 / 62.5	23.6 / 62.34
2CEX_A (Train: 5M)	X-RAI	8:14h	4.87	**1.4 / 1.8**	**1.4 / 1.8**
	M-TIP	> 24h	—	—	—

Table 1: Comparison of reconstruction accuracy for X-RAI and M-TIP on datasets of various sizes. Training and test pose error are reported in degrees, and the best resolutions and pose errors for each of the two proteins are highlighted in **bold**. X-RAI outperforms M-TIP across all datasets, maintaining a constant runtime for 50K, 500K, and 5M images. M-TIP converges to a solution faster for 50K and 500K images but times out after 24 hours for 5M images.

500,000, and 5,000,000 images (hereafter referred to as 50K, 500K, and 5M, respectively). A distinct autoencoder model is optimized for each dataset via stochastic gradient descent, and we adjust the number of training epochs based on the dataset size such that each reconstruction performs the same number of gradient update steps. The model is implemented in PyTorch [22] and trained on a single Tesla A100 GPU for each experiment. We report the resolution of each reconstruction, as determined by a Fourier shell correlation with a 0.5 cutoff, as well as the accuracy of the estimated poses in degrees. Each dataset is also processed using M-TIP [4], which serves as a baseline for comparison. More details regarding data generation, training, and evaluation can be found in appendix A.1.

The results of these reconstructions are reported in figure 2 and table 1. X-RAI is able to reconstruct all six datasets to within 6 pixels of resolution, and its accuracy remains relatively equal across all dataset sizes, which is to be expected in the noise-free setting as even the smallest dataset of $50,000$ images sufficiently covers $SO(3)$. Our method outperforms M-TIP across all datasets, although M-TIP is able to converge to a solution faster than X-RAI for 50K and 500K images. M-TIP times out after 24 hours when processing datasets with 5M images, behavior that is explained further in appendix A.2. Furthermore, it struggles to reconstruct the datasets corresponding to 2CEX_A even with 500K images, with orientation errors exceeding 20 degrees.

Notably, X-RAI's reconstruction time remains constant relative to the size of the dataset, demonstrating the efficacy of the encoder at amortizing pose estimation. Moreover, the pose accuracy on the test set remains high, showing that the encoder is able to learn and not memorize the statistics of the input diffraction images. Our method reconstructs both datasets containing 5 million images in an online fashion, processing the data sequentially in batches of 64.

4 Discussion

Reconstruction algorithms for X-ray SPI experiments will need to scale to datasets containing tens of millions of images that will be produced by next-generation XFELs. Existing techniques scale poorly with dataset size because they estimate the orientation of each image independently. In this work, we propose an amortized approach to pose estimation that uses an autoencoder framework to reconstruct the electron density. We validate our method on synthetic, noise-free datasets containing up to 5 million images, demonstrating online reconstruction for the largest datasets. The main limitation of our method is that it has not been validated on datasets containing realistic levels of signal and noise. In the future, we intend to test our method on synthetic datasets containing noise as well as experimental datasets captured by real free-electron lasers.

References

[1] Johan Bielecki, Filipe RNC Maia, and Adrian P Mancuso. "Perspectives on single particle imaging with x rays at the advent of high repetition rate x-ray free electron laser sources". In: *Structural Dynamics* 7.4 (2020).

[2] RW Schoenlein et al. "The linac coherent light source: recent developments and future plans". In: *Applied Sciences* 7.8 (2017), p. 850.

[3] Kartik Ayyer et al. "*Dragonfly*: an implementation of the expand–maximize–compress algorithm for single-particle imaging". In: *Journal of Applied Crystallography* 49.4 (June 2016), pp. 1320–1335. DOI: 10.1107/s1600576716008165. URL: https://doi.org/10.1107/s1600576716008165.

[4] Jeffrey J. Donatelli, James A. Sethian, and Peter H. Zwart. "Reconstruction from limited single-particle diffraction data via simultaneous determination of state, orientation, intensity, and phase". In: *Proceedings of the National Academy of Sciences* 114.28 (June 2017), pp. 7222–7227. DOI: 10.1073/pnas.1708217114. URL: https://doi.org/10.1073/pnas.1708217114.

[5] Axel Levy et al. "CryoAI: Amortized Inference of Poses for Ab Initio Reconstruction of 3D Molecular Volumes from Real Cryo-EM Images". In: *Lecture Notes in Computer Science*. Springer Nature Switzerland, 2022, pp. 540–557. DOI: 10.1007/978-3-031-19803-8_32. URL: https://doi.org/10.1007/978-3-031-19803-8_32.

[6] Axel Levy et al. "Amoritized Inference for Heterogeneous Reconstruction in Cryo-EM". In: *Proc. NeurIPS*. 2022.

[7] Ahmad Hosseinizadeh et al. "Conformational landscape of a virus by single-particle X-ray scattering". In: *Nature Methods* 14.9 (Aug. 2017), pp. 877–881. DOI: 10.1038/nmeth.4395. URL: https://doi.org/10.1038/nmeth.4395.

[8] Ne-Te Duane Loh and Veit Elser. "Reconstruction algorithm for single-particle diffraction imaging experiments". In: *Physical Review E* 80.2 (Aug. 2009). DOI: 10.1103/physreve.80.026705. URL: https://doi.org/10.1103/physreve.80.026705.

[9] Dimitrios Giannakis, Peter Schwander, and Abbas Ourmazd. "The symmetries of image formation by scattering I Theoretical framework". In: *Optics Express* 20.12 (May 2012), p. 12799. DOI: 10.1364/oe.20.012799. URL: https://doi.org/10.1364/oe.20.012799.

[10] Peter Schwander et al. "The symmetries of image formation by scattering II Applications". In: *Optics Express* 20.12 (May 2012), p. 12827. DOI: 10.1364/oe.20.012827. URL: https://doi.org/10.1364/oe.20.012827.

[11] Ali Dashti, Ivan Komarov, and Roshan M. D'Souza. "Efficient Computation of k-Nearest Neighbour Graphs for Large High-Dimensional Data Sets on GPU Clusters". In: *PLoS ONE* 8.9 (Sept. 2013). Ed. by Attila Gursoy, e74113. DOI: 10.1371/journal.pone.0074113. URL: https://doi.org/10.1371/journal.pone.0074113.

[12] Ellen D. Zhong et al. "CryoDRGN: reconstruction of heterogeneous cryo-EM structures using neural networks". In: *Nature Methods* 18.2 (Feb. 2021), pp. 176–185. DOI: 10.1038/s41592-020-01049-4. URL: https://doi.org/10.1038/s41592-020-01049-4.

[13] Ellen D. Zhong et al. "CryoDRGN2: Ab initio neural reconstruction of 3D protein structures from real cryo-EM images". In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2021. DOI: 10.1109/iccv48922.2021.00403. URL: https://doi.org/10.1109/iccv48922.2021.00403.

[14] PP Ewald. "Introduction to the dynamical theory of X-ray diffraction". In: *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography* 25.1 (1969), pp. 103–108.

[15] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[16] Yi Zhou et al. "On the continuity of rotation representations in neural networks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 5745–5753.

[17] Ben Mildenhall et al. "Nerf: Representing scenes as neural radiance fields for view synthesis". In: *Communications of the ACM* 65.1 (2021), pp. 99–106.

[18] Vincent Sitzmann et al. "Implicit neural representations with periodic activation functions". In: *Advances in neural information processing systems* 33 (2020), pp. 7462–7473.

[19] Ariana Peck et al. "Skopi: a simulation package for diffractive imaging of noncrystalline biomolecules". In: *Journal of Applied Crystallography* 55.4 (2022), pp. 1002–1010.

[20] Bing Xiao et al. "Crystal structure of the retinoblastoma tumor suppressor protein bound to E2F and the molecular basis of its regulation". In: *Proceedings of the National Academy of Sciences* 100.5 (2003), pp. 2363–2368.

[21] Axel Muller et al. "Conservation of structure and mechanism in primary and secondary transporters exemplified by SiaP, a sialic acid binding virulence factor from Haemophilus influenzae". In: *Journal of Biological Chemistry* 281.31 (2006), pp. 22212–22222.

[22] Adam Paszke et al. "Pytorch: An imperative style, high-performance deep learning library". In: *Advances in neural information processing systems* 32 (2019).

[23] Ariana Peck. "cmtip". In: (2022). URL: `https://github.com/apeck12/cmtip`.

A Appendix

A.1 Data Generation and Training

The detector used for simulation is a square with a side length of 100 mm and has 128×128 pixels. The simulated laser has a beam energy of 4.6 keV and incident fluence of 10^{12} photons per pulse. We use a random set of orientations to generate the images, which for the sake of simplicity are simulated to be noise-free. For each protein, we also generate a test set of 10,000 images in order to evaluate the accuracy of the encoder after reconstruction.

We optimize a distinct autoencoder model for each dataset via stochastic gradient descent, feeding the images to the pipeline in batches of 64. Each image is scaled to contain a total intensity summed over all pixels of at least $20,000$ and is converted to an amplitude image via a square root operation before being input to the encoder. The number of training epochs is determined such that each reconstruction pipeline performs the same number of gradient updates. More specifically, the 50K datasets run for 100 epochs, the 500K datasets run for 10 epochs, and the 5M datasets run for 1 epoch. We perform 5 independent reconstructions using X-RAI and 10 reconstructions with M-TIP, reporting the results of the best reconstruction for each method as determined by the pose accuracy on the training data. The Fourier amplitude volume output by each reconstruction pipeline is phased using alternating steps of the error reduction and hybrid input-output algorithms, and we take the inverse Fourier transform of the resulting volume to retrieve the electron density V.

We rely on two metrics to evaluate reconstruction accuracy – volumetric resolution and pose error. For the former, we compute the resolution of the output electron density by aligning the estimated volume with the ground truth molecular volume used for simulation. We report the resolution as the Fourier shell correlation between the two volumes using a 0.5 cutoff. To compute the accuracy of the poses estimated by the encoder, we first align the orientation estimates with the ground truth orientations used during simulation by searching for a rotation matrix that minimizes the average view-direction error between the true and estimated poses. Because the Ewald geometry used in our experiments has minimal curvature and the Fourier amplitude is centrosymmetric, all the diffraction images possess near-in-plane rotational symmetry by π radians. Thus, the ground truth pose and its in-plane rotation by π radians are both valid orientation estimates for a given image. We design a pose error metric that accounts for this in-plane ambiguity by projecting the in-plane components of the aligned orientation estimates to lie in the interval $[0, \pi]$ and then computing the angle of the rotation matrix that aligns each projected estimate with the corresponding ground truth pose.

A.2 M-TIP Implementation

We use an unofficial implementation of M-TIP [23] that is not optimized for performance. One artifact of this implementation is that it attempts to load the entire dataset into memory at once, causing the reconstruction to time out after 24 hours for datasets containing 5M images.

Running M-TIP for more iterations or increasing the sampling density of its orientation search over $SO(3)$ could improve the algorithm's performance at the cost of increased runtime, but we do not test alternate settings in our evaluation.