Physics aware inference for the cryo-EM inverse problem: anisotropic network model heterogeneity, global pose and microscope defocus

Geoffrey Woollard[*]
gw@cs.ubc.ca

Shayan Shekarforoush[†]
shayan@cs.toronto.edu

Frank Wood[†]
fwood@cs.ubc.ca

Marcus A. Brubaker[‡]
mbrubake@yorku.ca

Khanh Dao Duc[§]
kdd@math.ubc.ca

Abstract

We propose a parametric forward model for single particle cryo-electron microscopy (cryo-EM), and employ stochastic variational inference to infer posterior distributions of the physically interpretable latent variables. Our cryo-EM forward model accounts for the biomolecular configuration (via spatial coordinates of pseudo-atoms, in contrast with traditional voxelized representations) the global pose, the effect of the microscope (contrast transfer function's defocus parameter). To account for conformational heterogeneity, we use the anisotropic network model (ANM). We perform experiments on synthetic data and show that the posterior of the scalar component along the lowest ANM mode and the angle of 2D in-plane pose can be jointly inferred with deep neural networks. We also perform Fourier frequency marching in the simulation and likelihood during training of the neural networks, as an annealing step.

1 Introduction

Single particle electron cryomicroscopy (cryo-EM) is a structural biology technique that gives detailed near-atomic-level information of biomolecules. Experimentalists labour for weeks, months, or even years to perfect experimental conditions that yield images which can be algorithmically processed to yield 3D structural insights. Here we propose an inference procedure that aims to determine the distribution of atomic heterogeneity from 2D images of single particles collected in a cryo-EM experiment, focusing on inferring interpretable parameters in a probabilistic framework.

Cryo-EM is a unique inverse imaging problem, in that there is a rich tradition of parametric equations in the forward model: for example, the three dimension structure of large biomolecules (typically folded proteins sometimes with nucleic acid and lipid) and the electron optics equations of the microscope. However, most cryo-EM data processing packages (1; 2; 3; 4; 5; 6) employed by practitioners transform raw 2D microscope images into one or more voxelized 3D maps that discretely represents a 3D scalar field numerically. This has been largely influenced by the historical context of data processing pipelines which grew out of a tradition of digital signal processing and computerized tomography (7; 8; 6). After map reconstruction practitioners then use another set of software tools to

[*]Department of Computer Science, University of British Columbia, Vancouver, British Columbia, Canada
[†]Department of Computer Science, University of Toronto, Toronto, Ontario
[‡]Department of Electrical Engineering and Computer Science, York University
Vector Institute, Toronto, Ontario, Canada
[§]Department of Mathematics, University of British Columbia, Vancouver, British Columbia, Canada

Machine Learning for Structural Biology Workshop, NeurIPS 2022.

fit atomic models into these 3D maps (9; 10; 11). Atomic models live in a coordinate based space, $\in \mathbb{R}^{3n_a}$, where n_a is the number of atoms; and the structural biology research community deposits them to the Protein Data Bank (12). To solve the "reconstruction problem" in cryo-EM one averages 2D images together into a 3D map of the Coulombic density. This siloed workflow is common (13) but at the same time problematic when does not quantify uncertainty propagation through the stages of computational workflows.

In contrast, recent work has focused on an atom or pseudo-atom encoding (14; 15; 16; 17) and some of this work has been compared and contrasted against voxelized treatments in a unifying framework in a recent review (18). Coordinated based approaches provides an opportunity to encode domain knowledge through physics based parametric forms, distribution types suited to their spaces (e.g. directional distributions for 2D pose), and distributional parameters coming from prior knowledge (e.g. CTF estimate, published atomic structures).

Here we propose a method to learn an ensemble of atomic structures directly from raw 2D cryo-EM measurements, and we perform inference on three latents: (1) continuous conformational heterogeneity (in contrast to discrete configurations), (2) the defocus of the point spread function of the microscope's objective lens, and equivalently its Fourier transform, the contrast transfer function (CTF), and (3) global 2D rotational pose. We consider the restricted case with a known fixed reference atomic model, and experiment on synthetic data with pose restricted to in plane 2D rotations as a proof of concept. Evaluation of the forward model, with all the cryo-EM specific deterministic computations, is rapid enough that it can be applied to each gradient step in gradient-based inference methods. We choose to approach this problem in a stochastic variational amortized inference setting (19; 20; 21). Our contributions are:

1. Scale the forward model to a large number of atoms in a large field of view ("box size"), with a fast approximate projection.
2. Perform inference on global 2D pose and conformational heterogeneity through deep encoder neural network architectures that map to a latent space that is physically interpretable through the forward model.
3. Employ frequency marching during training in the forward model and its likelihood, without having to retrain the inference neural networks.
4. Characterize the global rotation posterior with a projected normal mixture. (22).

2 Related Work

Over the past several years previous studies have employed a coordinate based representation (pseudo)atoms and perform inference with deep neural nets. EMAN2's deep Gaussian mixture model (14) uses an auto-encoder to represent continuous heterogeneity as a mixture of $N = 2000 - 3000$ Gaussian pseudo-atoms with learnable amplitude and location. Cryofold (15) represents the biomolecule as a set of coarse grained pseudo-atoms and learn parameters which control how intense and how spread out the Gaussian kernels are. They use a variational auto-encoder (VAE) to learn offsets to Gaussian centers and incorporate a prior that respects the polymeric nature of the folded protein. Rosenbaum *et al.* proposed a method that learns a conformational ensemble from synthetic cryo-EM measurements using a VAE approach (16) with a multilayer perceptrons (MLP) neural network architecture and all distributions are Gaussian. They learn the 3D pose and conformation of a coarse grained atomic representation, where each amino acid residue is represented by one Gaussian spherical density. They regularize the output of the conformational encoder and keep it close to the reference conformation with a backbone continuity loss. In contrast to these methods here the ANM employed in our forward model couples together heterogeneity of pseudo-atoms through the ANM component. We also sample from a projected normal distribution in the latent space for pose, and employ a mixture of projected normals in the variational posterior to account for uncertainty in the pose estimate.

Concurrent with this work Nashed *et al.*(17) demonstrate that the ANM modes capture the heterogeneity of adenylate kinase transitioning between open and closed conformations. A continuous trajectory was discreatly sampled in 50 states with a tool in a molecular viewer program to generate synthetic data. An autoencoder was used to estimate the up to 16 normal mode components, where the estimated values were used in a physics decoder. Other latent variables (rotation, CTF defocus, open conformation of atomic model) were provided and not inferred. The physics decoder represented the full atomic model with multiple Gaussians with atom-type specific parameters from established tabulated values. The elastic network model is computed on a subset of atoms, and then interpolated for the remaining atoms, avoiding the expensive diagonalization of the $3N \times 3N$ Hessian, where N is the number of atoms. Although very similar, the approach outlined here infers both conformational heterogenity and pose, uses a mixture of a directional distribution to parametrize the later in the variational posterior, and employs Fourier cropping. Admittedly, we infer in-plane 2D pose, employ only one ANM mode, use a reduced number of atoms, corrupt data with less noise, and generate synthetic data is directly from the forward model instead of interpolating between states, which would represent an alternate distribution than the model's prior.

3 Methods

3.1 Stochastic forward model (decoder/simulator)

Each observed image $\mathbf{Y}_i$ is simulated by a stochastic forward model; cf. Figure 1, Algorithm 1, and Eq. A11 in the appendix. A representative sample is shown in Figure 1. The forward model samples latents and maps them to the observation; it is also an interpretable physics decoder in stochastic variational inference, where the guide/encoder is responsible for sampling latents.

Algorithm 1 Generative (forward) model of image formation

Require: ref. conformation $\mathbf{m}_0 \in \mathbb{R}^{3n_a}$, pert. modes $(\mathbf{u}_m)m \in (R^{3n_a})^M$, pert. scales $\sigma_\alpha \in \mathbb{R}^M$, pseudo-atom radius $\sigma_a \in \mathbb{R}$, rotation param. $\mu_s \in \mathbb{R}^2$, defocus param. $\mu_z, \sigma_z \in \mathbb{R}$, detector noise param. $\sigma_n \in \mathbb{R}$, space grid $\mathbf{x} \in (\mathbb{R}^2)^{N\times N}$, freq. grid $\mathbf{k} \in (\mathbb{R}^2)^{N\times N}$, cropping size $K \le N \in \mathbb{N}$.

$\alpha \sim \mathcal{N}\left(\cdot \mid \mathbf{0},\ \mathrm{diag}(\sigma_\alpha)\right)$

$\mathbf{m} = \mathbf{m}_0 + \sum_m \alpha_m \mathbf{u}_m$ ▷ stoch. pseudo-atom conformation from ANM perturbations

$\mathbf{s} \sim \mathcal{PN}(\cdot \mid \mu_s)$

$\theta = \mathrm{atan2}(\mathbf{s})$

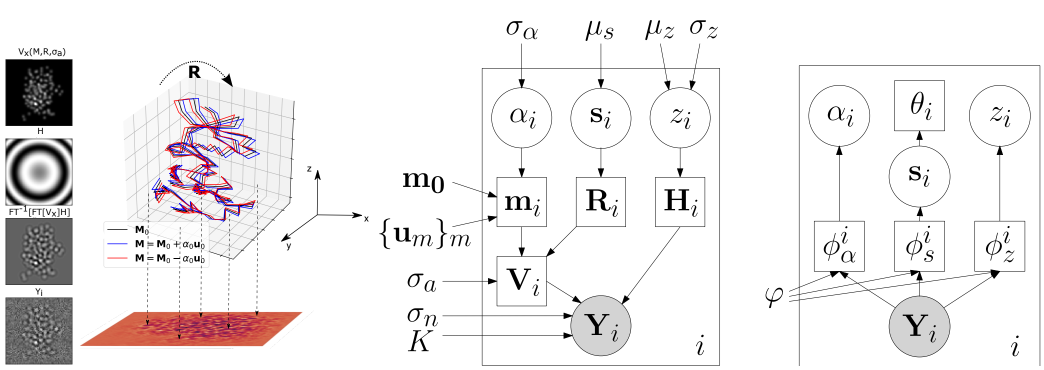
$\mathbf{R} = R_z(\theta)$ ▷ stoch. horizontal rotation matrix from unit vector

$\mathbf{V} = V_{\mathbf{x}}(\mathbf{m}, \mathbf{R}, \sigma_a)$ ▷ projected electron density field, cf. Eq. (11)

$z \sim \mathcal{N}(\mu_z, \sigma_z)$

$\mathbf{H} = \sin(-z \cdot |\mathbf{k}|^2)$ ▷ circular symmetric CTF; defocus in "natural units"

$\mathbf{Y} \sim \mathcal{N}\left(\cdot \mid \hat{\mathcal{F}}_{\mathbf{k}}^{-1}\left((\Pi_K \hat{\mathcal{F}}_{\mathbf{k}}\mathbf{V}) \odot \mathbf{H}\right),\ \sigma_n \cdot \frac{K}{N} \cdot \mathbb{1}\right)$ ▷ convolution & freq. cropping via FFT



(a) Cartoon of forward model. (b) Prior model. (c) Inference model.

Figure 1: **(a):** The reference atom positions are additively perturbed along the ANM mode component, rotated, and projected to 2D by integrating along the z-axis. The CTF is applied to the 2D projection, and Gaussian white noise is applied. The reference conformation $\mathbf{m}_0$ (in black), along with two states of $\mathbf{m}$, at $\pm\alpha_0$ (in blue and red) are shown. **(b)**: Graphical model of the stochastic physics simulator of cryo-EM image formation; cf. Algorithm 1. **(c)**: Graphical model of the inference model. The observed image is fed to three neural networks that independently predict distributional parameters, which are sampled from using a distribution of choice (Gaussian for defocus and ANM, and 2D Projected Normal for 2D in plane pose). See Algorithm 2 in the appendix for more detail.

3.2 Stochastic variational inference

We used the stochastic variational inference in the framework provided by the deep probabilistic programming language Pyro (`pyro.infer.SVI`) (20), which minimizes the evidence lower bound:

$$\text{ELBO} \equiv \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log p_\theta(\mathbf{x}, \mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x})\right] \tag{1}$$

In Eq 1 the observed single particle images are $\mathbf{x}$, latents in the graphical model are $\mathbf{z}$ (notation consistent with Pyro's documentation). The stochastic forward model is $p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z})$ is our forward model / physics based decoder. Prior domain knowledge of the the image formation process is included in required distributional parameters, and incorporates knowledge of the sample, the noise level, microscope fluctuations. The learned posterior $q_\phi(\mathbf{z}|\mathbf{x})$, is the variational distribution / encoder that consumes the measured data and maps it to latent space.

Our choices for the variational posterior / encoder are shown in Figure 1c and Algorithm 2 in the appendix, with the neural network architectures outlined and further details in Figure A1. Here we used three (one for each ANM, pose, and CTF defocus) CNN-MLP based neural networks with no conditioning or weight sharing, with architectures similar to those previously published in (23; 24; 25). We used a mixture of two projected normal distributions for the rotation, transforming the unit 2-vector ($\mathbf{s}$, belonging to the circle group S^1) to an angle that defines a rotation about the imaging axis (z-axis).

3.3 Training

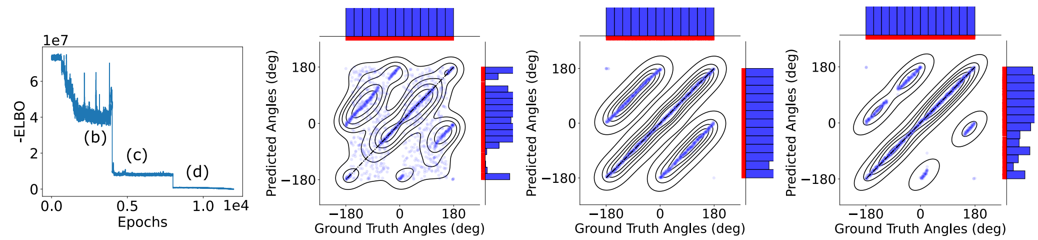
Training / optimizing the objective is done with `pyro.infer.SVI`, which takes a model, guide, optimizer, and loss as arguments. Here the `Trace_ELBO` loss is used (see Eq. 1). For the opimizer, we use `pyro.optim.ClippedAdam` optimizer with default parameters. During training, the neural networks in the guide always see the full resolution data, while in model, the observed data and simulation are cropped to a specified resolution K in Fourier space. The Gaussian white noise in the likelihood computation is adjusted by the boost in signal: i.e. $\sigma_n \to \sigma_n \times \frac{K}{N}$, where K is the maximal wave number after cropping and N is the full resolution wave number before cropping. The parameters in the guide are registered for optimization with `pyro.module` and no parameters in the prior model (e.g. μ_s) are optimized, although this is possible in Pyro. The `pyro.optim.ClippedAdam` optimizer with a learning rate of 0.001, a batch size of 500 is used, with $N_p = 2000$ training examples and 2000 independent and identically distributed (iid) testing examples. We noticed no difference in evaluation between testing and training data, indicating that the neural networks were not memorizing noise.

3.4 Data Set

We generated synthetic data from using the stochastic forward model of same biomolecule as in (16), Aurora A Kinase, with box size ($N = 64$ pixels), and pseudo-atoms positions from every second alpha carbon (PDB: 1OL5), for computational efficiency, and thus the protein is coarse grained as 133 pseudo-atoms by using every second alpha carbon backbone atom, which gives the general shape of the molecule. Unless otherwise noted, data was generated with the same parameters in the model (its prior): for the CTF defocus, $\mu_z = 50, \sigma_z = 3$ (in "natural units" of defocus; for the ANM modes, $\forall m,\ \mu_{\alpha_m} = 0$, $\sigma_{\alpha_0} = 3$, and $\forall m \neq 0,\ \ \sigma_{\alpha_m} = 0$, corresponding to a single mode; for the measurement noise $\sigma_n = 0.1$, corresponding to a signal to noise (signal variance / noise variance) of 2.6; for pose an in plane uniform prior of $\mu_{\mathbf{s}} = (0, 0)$ was used. For the deterministic projection a Gaussian spread of $\sigma_a = 0.8$ pixels was used and densities were truncated to within a $6\sigma_a \times 6\sigma_a$ pixel patch centred at each atom. See additional detail of the forward model in appendix A.3.

4 Experiments

We first established that the each latent (CTF defocus, 2D pose, ANM scalar) could be estimated while keeping the other two latents fixed at their ground truth values (data not shown). Inferring the pose was possible when there was a non-uniform prior on the pose, for example a standard deviation of $20 - 60°$. However inferring fully non-uniform pose became prohibitively difficult at full resolution, and we sought an alternative strategy instead of training for an excessively long time at full resolution.



(a) Training loss. (b) K=32 (c) K=48 (d) K=64

Figure 2: Effect of frequency marching on pose prediction. **(a)**: Training was done at Fourier cropping levels to 32 pixels (b), 48 pixels (c) and 64 pixels (full resolution) (d) for 4000 epochs each, for a total of $1.2 \cdot 10^4$ epochs. **(b-d)** : At the end of each training stage ((b):4000, (c): $2 \cdot 4000$, (d):$3 \cdot 4000$ epochs), samples for 2000 synthetic test set images were drawn from the respective posterior (checkpoint), and the joint and marginal distributions are shown, along with kernel density estimate contours showing iso-proportions of the density.

4.1 Frequency marching for pose prediction

Frequency marching is commonly employed in 2D or 3D array based representations of the potential. Inference begins at a low resolution stage using a small number of low frequency Fourier components, and this proceeds to incorporate higher frequency information as training proceeds. Pixel based ℓ_2 loss suitable for Gaussian white noise are sensitive to slight de-registration of pose (26), depending on the relative length scale between pixel size, width of pseudo-atoms (here σ_a), and the pose accuracy. This motivated us to take a closer look at how low pass filtering via Fourier cropping (Figure A2) could be employed during training.

After visualizing the likelihood (Figure A3), we noted how quickly the gradient signal vanishes, meaning samples too far from the ground truth value would have little to no gradient signal. To mitigate the vanishing gradient, we introduced a series of Fourier cropping levels during training. One can think of this as a transfer learning series, or piecewise SVI optimizations where the networks parameters start at some non-random state. Figure 2b-d shows a correlation of predicted to ground truth poses that improves as training progresses and the resolution used in the likelihood increases. Every time the Fourier cropping level (i.e. resolution) level changes, the training data is transformed to that resolution, and also during the simulation a corresponding Fourier crop is introduced. After cropping the noise remains white Gaussian, but adjusted by a factor from the cropping. This can be seen from the fact that Fourier cropping is a convolution with a constant top hat (i.e. step function) filter in real space, thereby summing nearby pixels. Thus the effect of Fourier cropping is to add iid Gaussian white noise, which translates to decreasing variance and boosting the signal to noise ratio. While three resolution levels are shown for reasons of clarity, any arbitrary "cropping schedule" could be employed such as a one Fourier wave number at a time. As training proceeds from $K = 32$ to $K = 48$ and finally $K = 64$ the correlation improves. Note that a line offset by $\pm 180°$ is apparent in the joint correlation plots, which is due to the $180°$ pseudo symmetry of the observed 2D views. As training proceeds at higher resolution the correlation from $180°$ pseudo-symmetry is mitigated. This supports a claim that using a mixture in the variational posterior for pose could help account for uncertainty arising from (pseudo)-symmetries of a biomolecule.

4.2 Inference of ANM and pose

We first worked up to doing joint inference on ANM and pose. In initial experiments we performed inference on single latents and pairs of latents, meaning the ground truth value was used in the posterior for the “missing" latents. While all single latents proved trainable in a small number of epochs, only the latent pairs of defocus-ANM and defocus-rotation were readily trainable. In contrast the ANM-pose latent pair proved difficult to train: the posterior sometimes collapsed to a large biased prediction (e.g. $\alpha_0 = 10$) or remained diffuse and uncorrelated. When we optimized the networks for both pose and ANM at the same time, i.e. at each gradient step the weights for the pose network and ANM network were updated, training failed.

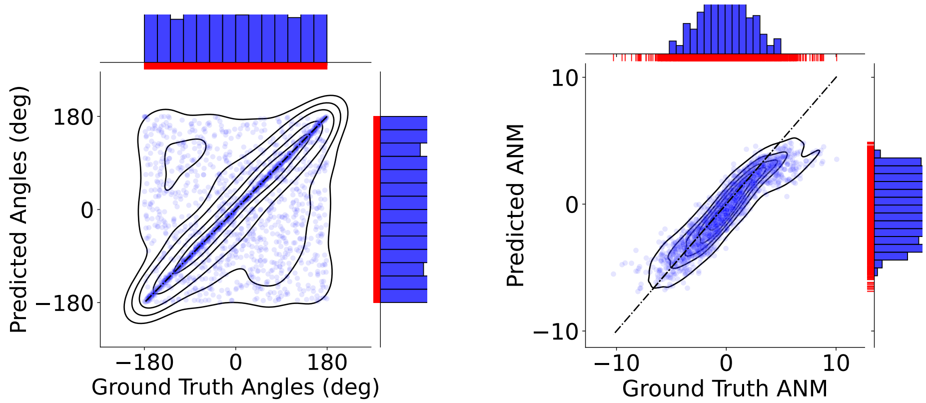


Figure 3: Joint inference of ANM and pose. After training the ground truth pose and ANM latents correlate with their sampled values from the posterior. Training samples and countours are as in Figure 2.

To overcome this, we decided to first optimize for pose, then fixed the pose weights and optimized for ANM. This approach, which was inspired by the training schedule in (27), made learning possible (Figure 3). This proved feasible for a uniform pose distribution and an ANM distribution ranging several pixels $\sigma_{\alpha_0} = 3$, with a defocus fixed at $z = 50$, and the Fourier cropping fixed at 32 pixels (half of the original level). The prediction and ground truth correlated with an average pose residual of $33°$, and average ANM residual of 1.0.

5 Conclusion and Future outlook

Our approach is inspired by simulation based inference and probabilistic programming literature (19; 28; 21), which perhaps can be put in contrast to approaches within the deep learning and computer vision, where general inference procedures are proposed that aim to be suitable without an intimate familiarity with the scientific domain, and in particular the domain specific mathematical modelling tradition.

Here we have shown a proof of principle of inference on synthetic cryo-EM images under a parametric forward model that is thereby physically interpretable. We used a two component mixture of the projected normal for sampling from in plane 2D pose. Extending to 3D pose is possible by sampling unit quaternions from the four dimensional form of the projected normal distribution, although it would require some more work to solve the issue with double cover. We employ a mixture of projected normals in the variational posterior, and this should allow us to inspect the mixture components that would be expected to arise from symmetries of the biomolecule's coordinates, and to marginalize over such symmetries.

We predicted the scalar on one ANM mode, and this can be extended to several low modes to express richer forms of heterogeneity. We used a mean field approximation, meaning that each network in the variational posterior takes as input the observed image, and does not condition on predicted latents for that observed image, and in future work we indent to relax this assumption. While the ANM is easily sampled from and scored, it is considered suitable for modelling low frequency fluctuations around a reference conformation, and does not immediately incorporate known prior information about bonds lengths and angles (e.g. in rings) and secondary structure elements, although there are extensions that incorporate shared rigidity among in a set of pseudo-atoms (29). Beyond ANMs, we are interested in modelling heterogeneity in a manner that can incorporate prior knowledge (sequence, secondary structure, polymeric nature of a biomolecule) and that furthermore has a distribution that is readily sampled from and scored.

5.1 Acknowledgements

GW thanks David Fleet for early discussions (circa 2019) about the ANM; Youssef Nashed, Frédéric Poitevin, Harshit Gupta, Michael Kagan, and Daniel Ratner for further discussions (circa 2021)

of employing ANMs in the context of end-to-end cryo-EM reconstruction; James Krieger for an overview of the implementation of the ANM in Prody (30), which inspired the implementation used here.

References

[1] 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] 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. *Nature Protocols*, 3(12):1941–1974, dec 2008.

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

[4] Ali Punjani, John L. Rubinstein, David J. Fleet, and Marcus A. Brubaker. CryoSPARC: Algorithms for rapid unsupervised cryo-EM structure determination. *Nature Methods*, 14(3):290–296, feb 2017.

[5] Timothy Grant, Alexis Rohou, and Nikolaus Grigorieff. cisTEM, user-friendly software for single-particle image processing. *eLife*, 7:1–24, mar 2018.

[6] Suvrajit Maji and Joachim Frank. What is in the black box? – A perspective on software in cryoelectron microscopy. *Biophysical Journal*, 120(20):4307–4311, 2021.

[7] Grant J. Jensen, editor. *Methods in Enzymology, volume 482: Cryo-EM, Part B: 3-D Reconstruction*. Academic Press, 2010.

[8] Amit Singer. Mathematics for Cryo-Electron Microscopy. In *Proceedings of the International Congress of Mathematicians (ICM 2018)*, pages 3995–4014. World Scientific, may 2019.

[9] P. Emsley, B. Lohkamp, W. G. Scott, and K. Cowtan. Features and development of Coot. *Acta Crystallographica Section D Biological Crystallography*, 66(4):486–501, apr 2010.

[10] Ana Casañal, Bernhard Lohkamp, and Paul Emsley. Current developments in Coot for macromolecular model building of Electron Cryo-microscopy and Crystallographic Data. *Protein Science*, 29(4):1055–1064, apr 2020.

[11] Dorothee Liebschner, Pavel V. Afonine, Matthew L. Baker, Gábor Bunkóczi, Vincent B. Chen, Tristan I. Croll, Bradley Hintze, Li-Wei Hung, Swati Jain, Airlie J. McCoy, Nigel W. Moriarty, Robert D. Oeffner, Billy K. Poon, Michael G. Prisant, Randy J. Read, Jane S. Richardson, David C. Richardson, Massimo D. Sammito, Oleg V. Sobolev, Duncan H. Stockwell, Thomas C. Terwilliger, Alexandre G. Urzhumtsev, Lizbeth L. Videau, Christopher J. Williams, and Paul D. Adams. Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallographica Section D Structural Biology*, 75(10):861–877, oct 2019.

[12] H. M. Berman, J Westbrook, Z Feng, G Gilliland, T N Bhat, H Weissig, I N Shindyalov, and P E Bourne. The Protein Data Bank. *Nucleic acids research*, 28(1):235–42, jan 2000.

[13] Robert M Glaeser, Eva Nogales, and Wah Chiu, editors. *Single-particle Cryo-EM of Biological Macromolecules*. IOP Publishing, may 2021.

[14] Muyuan Chen and Steven J. Ludtke. Deep learning-based mixed-dimensional Gaussian mixture model for characterizing variability in cryo-EM. *Nature Methods*, 18(8):930–936, aug 2021.

[15] Ellen D. Zhong, Adam Lerer, Joseph H. Davis, and Bonnie Berger. Exploring generative atomic models in cryo-EM reconstruction. *arXiv*, pages 1–13, jul 2021.

[16] 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*, pages 1–15, 2021.

[17] Youssef Nashed, Ariana Peck, Julien Martel, Axel Levy, Bongjin Koo, Gordon Wetzstein, Nina Miolane, Daniel Ratner, and Frédéric Poitevin. Heterogeneous reconstruction of deformable atomic models in Cryo-EM. 2022. `http://arxiv.org/abs/2209.15121`.

[18] Claire Donnat, Axel Levy, Frederic Poitevin, and Nina Miolane. Deep Generative Modeling for Volume Reconstruction in Cryo-Electron Microscopy. *arXiv*, pages 1–26, 2022. `https://arxiv.org/abs/2201.02867`.

[19] Jan-Willem van de Meent, Brooks Paige, Hongseok Yang, and Frank Wood. An Introduction to Probabilistic Programming. *arXiv*, pages 1–221, 2018.

[20] Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D. Goodman. Pyro: Deep Universal Probabilistic Programming. *Journal of Machine Learning Research*, 2018.

[21] Alexander Lavin, Hector Zenil, Brooks Paige, David Krakauer, Justin Gottschlich, Tim Mattson, Anima Anandkumar, Sanjay Choudry, Kamil Rocki, Atılım Güneş Baydin, Carina Prunkl, Brooks Paige, Olexandr Isayev, Erik Peterson, Peter L. McMahon, Jakob Macke, Kyle Cranmer, Jiaxin Zhang, Haruko Wainwright, Adi Hanuka, Manuela Veloso, Samuel Assefa, Stephan Zheng, and Avi Pfeffer. Simulation Intelligence: Towards a New Generation of Scientific Methods. *arXiv*, 2021.

[22] Daniel Hernandez-Stumpfhauser, F. Jay Breidt, and Mark J. van der Woerd. The general projected normal distribution of arbitrary dimension: Modeling and Bayesian inference. *Bayesian Analysis*, 12(1):113–133, 2017.

[23] Youssef S. G. Nashed, Frederic Poitevin, Harshit Gupta, Geoffrey Woollard, Michael Kagan, Chun Hong Yoon, and Daniel Ratner. CryoPoseNet: End-to-End Simultaneous Learning of Single-particle Orientation and 3D Map Reconstruction from Cryo-electron Microscopy Data. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, volume 1, pages 4049–4059. IEEE, oct 2021.

[24] Axel Levy, Frédéric Poitevin, Julien Martel, Youssef Nashed, Ariana Peck, Nina Miolane, Daniel Ratner, Mike Dunne, and Gordon Wetzstein. CryoAI: Amortized Inference of Poses for Ab Initio Reconstruction of 3D Molecular Volumes from Real Cryo-EM Images. *arXiv*, mar 2022. `http://arxiv.org/abs/2203.08138`.

[25] Ruyi Lian, Bingyao Huang, Liguo Wang, Qun Liu, Yuewei Lin, and Haibin Ling. End-to-end orientation estimation from 2D cryo-EM images. *Acta Crystallographica Section D Structural Biology*, 78(2):174–186, 2022.

[26] Rohan Rao, Amit Moscovich, and Amit Singer. Wasserstein K-Means for Clustering Tomographic Projections. *arXiv*, (2016):1–11, 2020.

[27] Ellen D Zhong, Adam Lerer, Joseph H Davis, and Bonnie Berger. CryoDRGN2 : Ab initio neural reconstruction of 3D protein structures from real cryo-EM images. *Iccv*, pages 4066–4075, 2021.

[28] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences of the United States of America*, 117(48):30055–30062, 2020.

[29] Florence Tama, Florent Xavier Gadea, Osni Marques, and Yves-Henri Sanejouand. Building-block approach for determining low-frequency normal modes of macromolecules. *Proteins: Structure, Function, and Genetics*, 41(1):1–7, oct 2000.

[30] She Zhang, James M Krieger, Yan Zhang, Cihan Kaya, Burak Kaynak, Karolina Mikulska-Ruminska, Pemra Doruker, Hongchun Li, and Ivet Bahar. ProDy 2.0: increased scale and scope after 10 years of protein dynamics modelling with Python. *Bioinformatics*, 37(20):3657–3659, oct 2021.

[31] Ken Dill, Robert L. Jernigan, and Ivet Bahar. *Protein Actions*. Garland Science, New York, NY : Garland Science, Taylor Francis Group, LLC, [2017] |, sep 2017.

[32] Benjamin A. Himes and Nikolaus Grigorieff. Cryo-TEM simulations of amorphous radiation-sensitive samples using multislice wave propagation. *BioRxiv*, page 6, 2021.

A Appendix

A.1 Inference

Algorithm 2 Inference (inverse) model, i.e., variational posterior per image

Require: image $\mathbf{Y}$, rotation mixture size $n \in \mathbb{N}$, neural networks $\phi_{z|\alpha|s}, m_{z|\alpha|s}, s_{z|\alpha|s}, \mathbf{w}_s$.

$z \sim \mathcal{N}(\cdot \mid m_z(\phi_z(\mathbf{Y})), s_z(\phi_z(\mathbf{Y})))$ ▷ proposal for defocus
$\alpha \sim \mathcal{N}(\cdot \mid \mathbf{m}_\alpha(\phi_\alpha(\mathbf{Y})), \mathbf{s}_\alpha(\phi_\alpha(\mathbf{Y})))$ ▷ proposal for ANM coefficients
$\mathbf{s} \sim \sum_{k=1}^{n} \mathbf{w}_s(\phi_s(\mathbf{Y}))_k \cdot \mathcal{PN}(\cdot \mid \mathbf{m}_s(\phi_s(\mathbf{Y}))_k)$ ▷ mixture proposal for rotation
$\theta = \text{atan2}(\mathbf{s})$

A.2 Neural network architecture

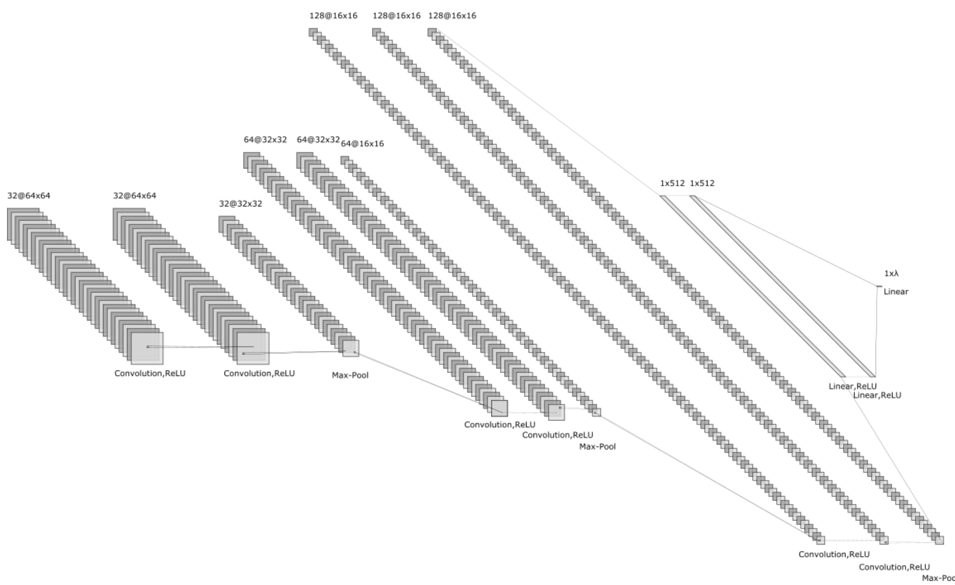


Figure A1: The neural network architectures is a series of three double convolutions with ReLU activations, (`Conv2d, ReLU, Conv2d, ReLU, MaxPool`) which is then flattened to an MLP two layers deep and 512 units wide with ReLU activations (`Linear,ReLU,Linear,ReLU`), to a final linear layer of size λ_{latent} to match the number of distribution parameters needed for the respective latent.

A.3 Forward Model

The forward model is outlined in Algorithm 1.

A.3.1 Conformational heterogeneity

The conformational heterogeneity can be understood from the perspective of an energy model physically inspired by each pseudo-atom being a "ball" attached by "springs" to other pseudo-atoms (31). The balls are centered at the positions of the pseudo-atom nuclei $\mathbf{m}_0 \in \mathbb{R}^{3n_a}$, where n_a is the number of pseudo-atom balls, and any new conformation $\mathbf{m} \in \mathbb{R}^{3n_a}$ has probability $\mathcal{P}(\mathbf{m})$, governed by energy $U(\mathbf{m}) \in \mathbb{R}$ and inverse temperature $\beta \in \mathbb{R}$.

$$\mathcal{P}(\mathbf{m}) = Z^{-1} \exp[-\beta U(\mathbf{m})] \tag{2}$$

$$U = U_{\text{anm}} = \frac{\gamma}{2} \sum_{ij} (r_{ij} - r_{0,ij})^2 \tag{3}$$

The anisotropic network model has energy U_{anm}, where r_{ij} is the distance between pseudo-atom pair ij in the sample $\mathbf{m}$, $r_{0,ij}$ is the corresponding reference distance in $\mathbf{m}_0$, and γ is a spring constant. The second derivative elements of the $3n_a \times 3n_a$ Hessian has a convenient analytical form with 3×3 symmetric ij submatrices given by

$$\mathbf{H}_{ij} = \frac{\gamma}{r_{ij}^2} \begin{bmatrix} x_{ij}^2 & x_{ij}y_{ij} & x_{ij}z_{ij} \\ x_{ij}y_{ij} & y_{ij}^2 & y_{ij}z_{ij} \\ x_{ij}z_{ij} & y_{ij}z_{ij} & z_{ij}^2 \end{bmatrix} \tag{4}$$

And the ii diagonal submatrices given by the row/column sum $\mathbf{H}_{ii} = \sum_{j \neq i} \mathbf{H}_{ij}$. The anisotropic network model is *anisotropic* in the sense that each xyz direction has its own Hessian component and can be different from other directions–hence *an*isotropic. The probability is then approximated by a second order Taylor expansion about a reference pseudo-atomic configuration $\mathbf{m}_0$, which is assumed to be at a local extremum point so the gradient term vanishes:

$$U(\mathbf{m}) = U(\mathbf{m}_0) - \frac{1}{2}(\mathbf{m} - \mathbf{m}_0)^T \mathbf{H}(\mathbf{m} - \mathbf{m}_0) \tag{5}$$

The eigendecomposition of $\mathbf{H} = \mathbf{U}\mathbf{\Lambda}^{\mathbf{-1}}\mathbf{U}^{\mathbf{T}}$, enables to project any pseudo-atom configuration $\mathbf{m}$ onto components of $\mathbf{U}$, $\mathbf{u}_m^T$, because $\alpha_m = \mathbf{u}_m^T(\mathbf{m} - \mathbf{m}_0)$. Thus $\mathbf{m}$ is a deterministic change of basis to the set $\{\alpha_m\}_1^{3n_a}$.

This exponential probability density function reduces the probability to a diagonal multivariate Gaussian through the orthogonality of the basis.

$$\begin{aligned} \mathcal{P}(\mathbf{m}) &= \mathcal{P}(\{\alpha_m\}) && (6) \\ &= (\det [2\pi \mathbf{\Lambda}])^{-1/2} \prod_m \exp -\beta \frac{\alpha_m^2}{\lambda_m} && (7) \\ &= \prod_m \mathcal{P}(\alpha_m) && (8) \end{aligned}$$

Thus a sample of pseudo-atomic positions is obtained by additively perturbing the mean pseudo-atomic positions $\mathbf{m_0} \in \mathbb{R}^{3n_a}$ by a perturbation vector, $\mathbf{u}_{\text{perturb}} = \sum_m \alpha_m \mathbf{u_m} \in \mathbb{R}^{3n_a}$, where each α_m is sampled from a Gaussian distribution, and each elastic network mode are fixed for constant $\mathbf{m_0}$.

$$\mathbf{m} = \mathbf{m_0} + \mathbf{u}_{\text{perturb}} \tag{9}$$

In the simplified model employed here, $\mathbf{u}_{\text{perturb}}$ is restricted the single lowest mode, and thus $\mathbf{m}$ is being sampling according to the distribution

$$\mathcal{P}(\mathbf{m}|\sigma_{\alpha_0}, \{\sigma_{\alpha_m} = 0\}_{m \neq 0}) = \mathcal{P}(\alpha_0|\alpha_1 = 0, ..., \alpha_{n_a-1} = 0) \propto \mathcal{N}(\alpha_m|0, \sigma_{\alpha_m} = \frac{\lambda_0}{2\beta}) \tag{10}$$

In brief, we explicitly compute the Hessian $\mathbf{H}$ of a reference conformation $\mathbf{m_0}$, compute its low mode eigenvectors and values, and keep this precomputed in memory. During stochastic simulation we sample a Gaussian scalar and additively perturb the reference conformation. Thus stochastic sampling of $\mathbf{m}$ is as fast as sampling Gaussians, scaling their corresponding eigenvectors, summing them to one eigenvector, and adding this perturbation to the reference conformation $\mathbf{m}_0$.

We use the PyTorch function `torch.linalg.eigh` to perform the eigendecomposition of the Hessian on the GPU.

A.3.2 Fourier Cropping

Fourier cropping was performed by sampling the CTF at full resolution (here $N = 64$ pixels), and then after the multiplication in Fourier space, transforming only the K central Fourier pixels back to real space. The effect of this can be visualized in Figure A2.

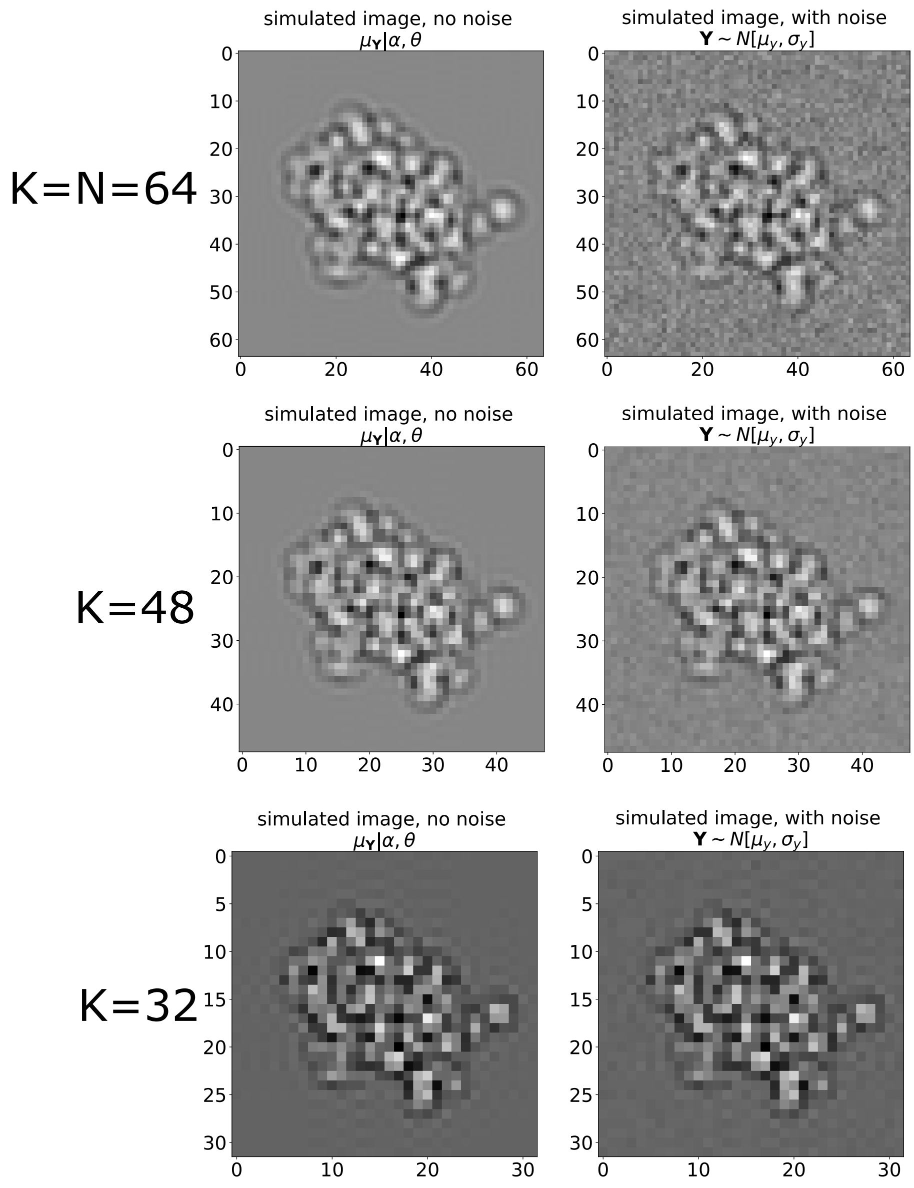


Figure A2: Three levels of Fourier cropping are shown. The noise is adjusted by $\sigma_n \to \frac{K}{N}\sigma_n$.

A.3.3 Projection

The 3D coordinates are projected down to, and sampled on a 2D array object of size $N \times N$. There is no sum done along the imaging (z) axis, because we take analytical advantage of the spherical shape of the isotropic Gaussian kernel, i.e. the convariance matrix of a single atom a atom is direction independent, $\Sigma_a^{-1} = \frac{I_3}{2\sigma_a^2}$. The projection is done by simply dropping the z coordinate after rotation ($\mathbf{Rm}_{(p)}$) as noted in (14; 32). We use the PyTorch data object `torch.sparse_coo_tensor` to implement this efficiently in batch on the GPU. The intensity at each pixel (i, j) is given by Eq. 11, whose parameters are further explained in Algorithm 1.

$$(V_{\mathbf{x}}(\mathbf{m}, \mathbf{R}, \sigma_a))_{i,j} = \sum_{p=1}^{n_a} \exp\left\{-\frac{\|\mathbf{x}_{i,j} - \mathbf{P}_{xy}\mathbf{Rm}_{(p)}\|^2}{2\sigma_a^2}\right\} \tag{11}$$

A.3.4 Likelihood

Under a white Gaussian likelihood, the probability falls off very quickly when the pose is out of registration.

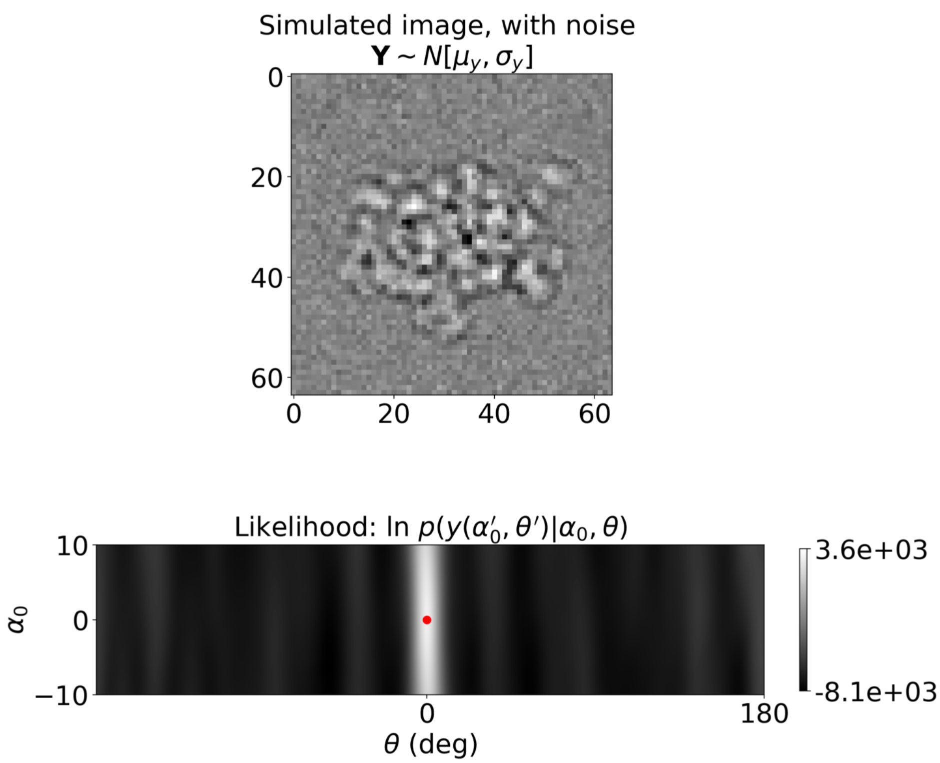


Figure A3: The likelihood of a was evaluated over the grid of latents: $\{\alpha_0\} \times \{\theta\} = (\{(\alpha_0, \theta) : \alpha_0 \in [-10, 10], \theta \in [-180, 180]\}$ for the shown noisy image, with $\alpha_0' = 0, \theta' = 0$ (red dot). The absolute scale of the likelihood is shown in the color bar in log_prob units.