Improving Ab-Initio Cryo-EM Reconstruction with Semi-Amortized Pose Inference

Shayan Shekarforoush[1, 2]
shayan@cs.toronto.edu

David B. Lindell[1, 2]
lindell@cs.toronto.edu

Marcus A. Brubaker[1, 2, 3, 4]
mab@eecs.yorku.ca

David J. Fleet[1, 2, 4]
fleet@cs.toronto.edu

[1]University of Toronto [2]Vector Institute [3]York University [4]Google DeepMind

Abstract

Electron cryomicroscopy (cryo-EM) is a popular experimental technique to recover the 3D structure of macromolecular complexes, such as proteins, using extremely noisy images that contain particles posed in unknown orientations. We propose a new semi-amortized approach to the *ab-initio* reconstruction problem. In early stages, when uncertainty is high, poses are estimated using auto-encoding, followed by auto-decoding as uncertainty decreases. A multi-head encoder is adopted for amortization to infer multiple plausible poses for each image, encouraging exploration of pose space, while flexible auto-decoding iteratively update poses per-image using stochastic gradient descent. Empirical results on synthetic datasets demonstrate that our method is able to handle multi-modal pose distributions, and the use of auto-decoding yields faster and more accurate pose convergence compared to baselines. We also show that on experimental data our approach achieves reconstruction with higher resolution than the current state-of-the-art.

1 Introduction

During a cryo-EM experiment, a *particle stack* of 10^4–10^7 images of a target bio-molecule are acquired by an electron microscope, from which the goal is to reconstruct the unknown 3D structure [1]. This *ab-initio* reconstruction task presents some challenges. First, the pose of the particle in each observation is unknown and needs to be estimated. Second, to prevent radiation damage, the electron exposure is limited leading to a poor signal-to-noise ratio (SNR). This obscures high-resolution details in images, complicating pose and structure estimation. Third, structures are often non-rigid, thus accounting for structural variability is crucial to achieve high-resolution reconstruction [2, 3, 4, 5].

Recent advancements in method development have centered on deep learning. CryoDRGN [4] introduced an image encoder-volume decoder architecture to model continuous heterogeneity with known poses. CryoDRGNv2 [6] improved on using a hierarchical pose search, comprising grid search followed by branch-and-bound (BnB), akin to cryoSPARC [7]. On the other hand, cryoPoseNet [8] and cryoAI [9] use amortized inference through CNNs to efficiently estimate poses without orientation matching. However, these amortized methods may struggle to capture multi-modal posterior distributions in early reconstruction stages, and their reliance on a global encoder can slow pose convergence, making them less accurate than approaches like cryoSPARC [7] and cryoDRGNv2 [6] with explicit per-image pose search. For a more comprehensive review of prior work, see Supp. A.

We present a novel method for ab-initio homogeneous reconstruction that handles multi-modal pose distributions with a tailored encoder and accelerates pose optimization using semi-amortization [10]. Unlike cryoPoseNet and cryoAI, which produce one or two pose estimates, we adopt a multi-head

Machine Learning for Structural Biology Workshop, NeurIPS 2024.

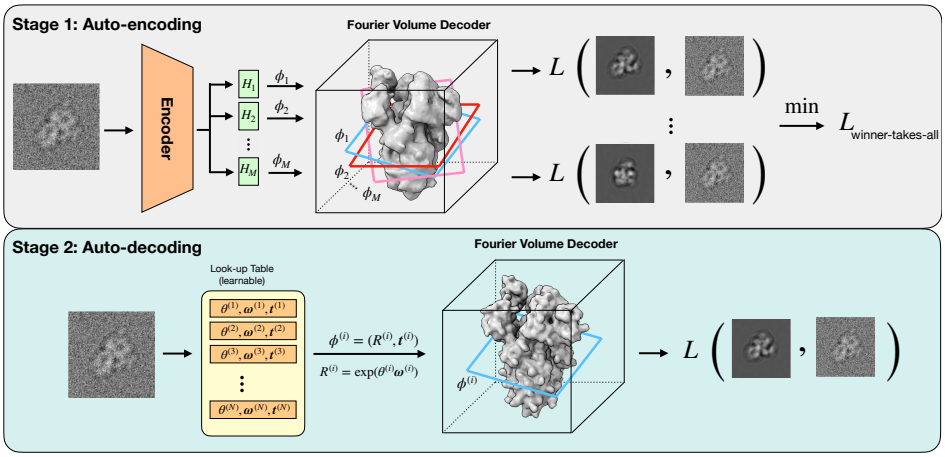

Figure 1: Our semi-amortized method consists of two stages: (i) an auto-encoding stage where a multi-head encoder maps the input image to the pose candidate set $\{\phi_1, \dots \phi_M\}$, followed by computing projections by slicing through the volume decoder in Fourier space based on the predicted pose set. The projection with the minimum error is used in the final loss. (ii) an auto-decoding stage where pose parameters are stored in axis-angle representation per-image. The same volume decoder is used to obtain projections, and the reconstruction loss is computed for a single projection.

CNN-based encoder to predict multiple candidate poses for each input image. This design accounts for pose uncertainty and encourages pose exploration during early stages of reconstruction. For structure decoding, instead of computationally expensive implicit models, as in cryoDRGN and cryoAI, we use an explicit parameterization enabling faster 3D reconstruction. For training, inspired by multi-choice learning [11, 12, 13], we use a "winner-takes-all" loss in which the decoder is queried to obtain 2D projection for all predicted poses, and the one with the lowest reconstruction error is selected to determine the loss. Also, as higher resolution details emerge, the pose posterior becomes uni-modal, allowing the pose search to focus on the most likely mode. At this point, we propose to switch to auto-decoding where poses are iteratively refined using stochastic gradient descent (SGD). This explicit per-image pose optimization yields accelerated convergence compared to solely relying on amortized inference, which depends on potentially sub-optimal encoder predictions.

Through comprehensive analysis on synthetic datasets, we validate that semi-amortized inference noticeably accelerates the convergence of poses and our multi-head encoder can handle multiple modes in the pose posterior. Moreover, our method quantitatively and qualitatively outperforms cryoSPARC [7], cryoAI [9], and cryoDRGN [6] on a real experimental dataset.

2 Methodology

We propose a semi-amortized approach to ab-initio cryo-EM reconstruction by mixing pose auto-encoding and auto-decoding. Initially, we adopt amortized inference (Fig. 1, auto-encoding) where a multi-head encoder outputs a set of pose guesses to handle the pose uncertainty. Once the pose posterior becomes less uncertain, we circumvent sub-optimal encoder predictions by switching to direct optimization (Fig. 1, auto-decoding) yielding arguably more accurate poses. We couple our pose estimation module with an explicit volumetric decoder representing the 3D structure in the Hartley space [4]. The explicit model enables faster evaluation of projections compared to implicit neural representations [14, 15, 16], significantly reducing the reconstruction time (see Supp. D). We provide the detailed mathematical formulation of the cryo-EM reconstruction problem in Supp. B.

2.1 Multi-choice Auto-encoding

Due to low SNR and near-symmetries in biological structures, there exist several equally-plausible poses for each image early in reconstruction, rendering naive optimization or search methods prone to local minima. To account for uncertainty, we build upon cryoAI [9] and extend the encoder to return multiple plausible poses. Formally, given the image $I_i \in \mathbb{R}^{H \times W}$, M poses are obtained as,

$$\phi_i = (R_{i,j}, t_{i,j}) = H_{\theta_j}(F_i)\ , \quad R_{i,j} \in SO(3)\ , t_{i,j} \in \mathbb{R}^2 \tag{1}$$

where image-specific intermediate features $F_i \in \mathbb{R}^{C\times H\times W}$ are extracted by VGG16 [17], and then supplied to M separate fully-connected predictor heads, $H_{\theta_j}, 1 \leq j \leq M$, yielding the pose set $\{\phi_{i,1}, \ldots, \phi_{i,M}\}$. To optimize the encoder-decoder, inspired by multi-choice learning [11, 12, 13], we use the "winner-takes-all" loss. The negative log-likelihood of the observed Fourier image $\hat{I}_i$ conditioned on the 3D Fourier volume estimate $\hat{V}$ with pre-computed CTF $\hat{g}_i$ and predicted pose $\phi_{i,j}$ is given as

$$\mathcal{L}_{i,j} = -\frac{1}{2\sigma^2}\sum_{\omega}[\hat{g}_i \cdot (\hat{\mathcal{P}}[R_{i,j}, t_{i,j}]\hat{V})(\omega) - \hat{I}_i(\omega)]^2 \,, \tag{2}$$

where $\hat{\mathcal{P}}$ is the slicing operator and σ^2 is the noise variance. The minimum is then selected as the final loss for the corresponding image, i.e., $\mathcal{L}_i = \min_j \mathcal{L}_{i,j}$. Interestingly, this loss encourages each head to specialize in pose estimation over localized regions with minimal overlap across heads (see Supp. F). Also, cryoAI [9] can be viewed a special case of this formulation; it assigns two poses to each image by input augmentation, and selects the best one with a symmetrized loss. In contrast, our approach augments the output of the encoder with multiple heads, each providing a pose estimate.

2.2 Switching from Auto-encoding to Auto-decoding

Unlike in early stages when high uncertainty encourages pose exploration, in later stages, as higher frequency details of the structure emerge, the variance of the pose posterior tends to decrease, becoming unimodal. At this point, the gap between the amortized and variational posterior is mainly determined by the error in the pose estimate (predicted mean), prioritizing accuracy over exploration. However, a feed-forward network, as a globally parameterized function of input images, may be too restrictive with limited prediction accuracy, rendering amortization as a barrier to further refinement of the 3D structure. In prior work [10, 18, 19, 20], a similar issue called the *amortization gap* has been discussed which measures the KL-divergence between the true and predicted variational posteriors.

To address this issue, we adopt a semi-amortized inference scheme [10] comprising two stages. First, the encoder predicts a set of pose candidates using a multi-head architecture. In the second stage, rather than amortized inference, pose parameters are directly optimized for each image using SGD. To initialize poses for the i-th image, we choose the one with the lowest reconstruction loss from the set of candidates $\{\phi_{i,1}, \ldots, \phi_{i,M}\}$, namely $\phi_i^* = \phi_{i,s}$ such that $s = \arg\min_j \mathcal{L}_{i,j}$. Subsequently, the pose and structure are optimized by coordinate descent using the negative log-likelihood as the objective function. Please see Supp. C for more details on pose optimization.

3 Experiments

Datasets. We compare our method with amortized methods of cryoAI [9] and cryoDRGN [6], as well as cryoSPARC [7] on both synthetic and real datasets. We simulate two synthetic datasets using PDB deposited atomic models of heat shock protein (HSP) [21] (1.5 Å) and pre-catalytic spliceosome [22] (4.33 Å), each comprising $N = 50,000$ noisy CTF-corrupted projections of size $L = 128$ with SNR $= 0.1$. To simplify the study, it is assumed synthetic particles are centered and we omit estimating translation parameters. We also adopt as a real benchmark of 80S experimental dataset [23] (EMPIAR-10028) containing 105,247 images of length $L = 360$ with pixel size 1.34 Å. Following prior work [6, 9], we downsample the images to $L = 128$ (3.76 Å) and randomly split the data into two halves and run the reconstruction methods independently on each.

Implementation details. During auto-encoding, we use encoders with $M = 7$ and $M = 15$ heads for reconstruction on synthetic and real datasets, respectively, with Adam [24] to optimize encoder and decoder with learning rates 0.0001 and 0.05. Once switched to direct optimization (after 7 epochs for synthetic and 15 epochs for real data), we reduce the decoder learning rate to 0.02 and allocate a new optimizer for pose parameters with learning rate 0.05. We use a batch size of 64 and train for the same number of epochs (20 for synthetic and 30 for real data). We use the public cryoAI codebase, cryoSPARC v4.4.0 [7] with default settings, and run the CryoDRGN in homogenous ab-initio setting.

3.1 Results

We first qualitatively compare reconstructions obtained by our method with cryoAI, cryoDRGN, and cryoSPARC (Fig. 2, left). We found that cryoAI fails to estimate translation parameters for 80S dataset. Hence, for this method, we preprocess images to be well-centered, whereas our method and

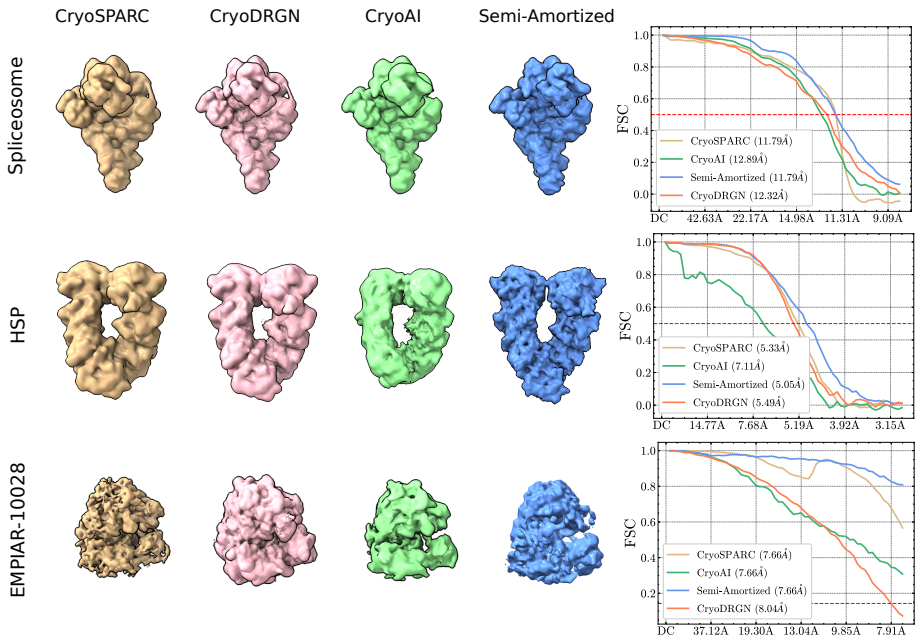

Figure 2: Qualitative and quantitative comparison of our semi-amortized method with cryoAI [9], cryoDRGN [6] and cryoSPARC [7]. **(Left)** Final 3D reconstructions on two synthetic datasets and one experimental data are depicted using ChimeraX [25]. **(Right)** FSC curves are visualized for quantitative comparison. The red dashed lines show the standard threshold levels of 0.5 and 0.143 to report the resolution (in angstroms) for synthetic and real data, respectively. Our method achieves higher resolution on the Spliceosome and HSP datasets, and it is competitive with the state of the art on EMPIAR-10028 datasets. See additional results on Spike Protein in Supp. E.

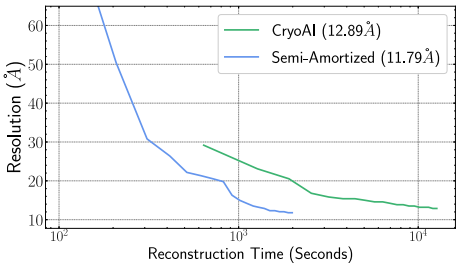

Figure 3: Resolution as a function of log time.

Method	HSP	Spliceosome
CryoSPARC [7]	6.23 / 1.05	1.41 / 1.36
CryoAI [9]	45.83 / 61.86	2.85 / 2.61
Ours	**3.27 / 0.97**	**0.68 / 0.61**

Table 1: Estimated rotation accuracy quantified as mean/median errors in units of degrees. CryoAI has greater error on the HSP dataset due to convergence to local minima, resulting in inaccurate pose estimates.

others are fed with off-centered particles. Both our method and cryoSPARC capture high-frequency details of the 3D structure on all datasets, whereas reconstructions by amortized methods, cryoAI and cryoDRGN, are inferior on HSP and 80S datasets. In particular, on HSP, cryoAI gets stuck in local minima as it fails to handle high uncertainty in poses caused by symmetries in this structure.

For quantitative comparison, we plot the gold-standard Fourier Shell Correlation (FSC) [26] (Fig. 2, right). FSC obtained by our method outperforms cryoSPARC as well as amortized methods of cryoAI and cryoDRGN on all datasets. Also, our method achieves higher or competitive resolution compared to others. We also report the mean and median errors in estimated poses on synthetic datasets in Table 1, showing that our method outperforms others. Moreover, the resolution-time plot in Fig. 3 shows that our semi-amortized method achieves a high-resolution reconstruction significantly faster than cryoAI. Our semi-amortization scheme accelerates the improvement in the resolution and the explicit decoder is more computationally efficient than an implicit MLP. In Supp. D we provide a detailed ablation study that shows our method is $\sim$ 6x faster and uses $\sim$ 5x less memory compared to cryoAI.

3.2 Semi-Amortized vs. Fully Amortized

To show the advantage of auto-decoding, we compare our semi-amortized method with a fully-amortized baseline on the Spliceosome dataset. After 7 epochs of auto-encoding, we split the reconstruction into two: one continues with pose encoding, while the other switches to direct pose

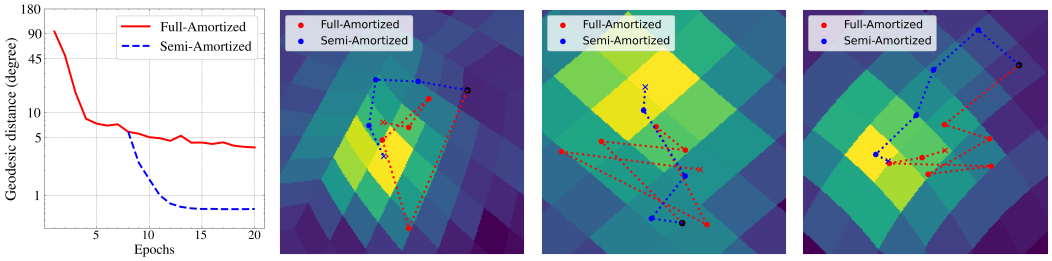

Figure 4: Comparison of fully- vs. semi-amortized methods in pose optimization on the Spliceosome dataset. **(Left)** We plot the mean pose error versus optimization epoch. Switching from amortized inference to direct optimization using our method **(blue)** leads to faster pose convergence compared to fully-amortized inference **(red)**. **(Right)** We visualize the approximate log posterior for three particles as a view-direction distribution on a unit sphere. The neighborhood of the mode of interest is visualized using Gnomonic projection. The **black** dot marks the initial point of optimization.

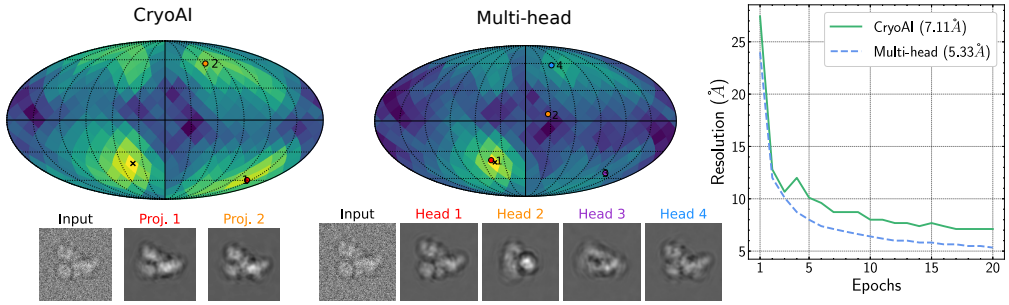

Figure 5: Comparison of the performance of our multi-head encoder ($M = 4$) with the cryoAI encoder on the challenging HSP [21] dataset. **(Left)** The approximate log posterior of view direction is visualized on the unit sphere with highlighted areas showing modes of the distribution. CryoAI and multi-head encoders provide two and four pose estimates, respectively, which are marked with colored dots on the sphere, with the corresponding projections illustrated below. CryoAI fails to find the correct mode while our method is able cover multiple modes. **(Right)** With our multi-head encoder, the reconstruction converges to a much higher resolution compared to cryoAI.

optimization. The mean pose error (Fig. 4, left), reveals that with transition to direct optimization, pose error drops rapidly. However, the fully-amortized baseline shows slow convergence, highlighting the superiority of auto-decoding in later reconstruction stages. We further investigate the trajectory of pose estimates overlaid on the optimization landscape (Fig. 4, right). Poses from direct optimization (blue dots) show stable convergence to the optimal point, while those from the encoder (red dots) often oscillate. Auto-decoding is more flexible and achieves more stable convergence than auto-encoding with potentially sub-optimal pose prediction. See Supp. G for more examples.

3.3 Multi-Modal Pose Posterior

Lastly, we examine the performance of our multi-head encoder vs. cryoAI encoder in handling the pose uncertainty on HSP dataset. We run our method with $M = 4$ heads. In the provided example, Fig. 5 left (see Supp. H for additional examples), the approximate posterior distribution over view direction is visualized for both cryoAI and the multi-head encoder. Our multi-head encoder identifies the correct mode while cryoAI selects an incorrect one. Our encoder also captures other posterior modes, offering improved exploration of pose space compared to cryoAI's limited predictions. This also leads to faster convergence and higher-resolution reconstructions (Fig. 5, right). In Supp. F, we investigate how each head specializes in pose prediction for non-overlapping regions of $SO(3)$.

4 Conclusion

In this paper, we propose a new semi-amortized approach to ab-initio cryo-EM reconstruction. We develop a multi-head encoder to estimate a set of plausible candidates to handle pose uncertainty. As the uncertainty is reduced, we switch to auto-decoding which iteratively refines poses using SGD per-image. Our results show that the multi-head encoder is able to capture multiple modes of the pose

distribution, and our flexible auto-decoding accelerates convergence of poses and reconstructions. Our method outperforms cryoAI on experimental data and achieves competitive results with cryoSPARC.

Limitations and Future work. We assume that the 3D structure is rigid while it is often flexible and deform within the sample. As a direction for future work, our semi-amortized method with multi-head encoder can be extended to include heterogeneity as a latent variables as well. Moreover, developing a well-defined heuristic to decide on relative length of two stages of auto-encoding and auto-decoding is an interesting direction to explore in the future.

References

[1] Amit Singer and Fred J Sigworth. Computational methods for single-particle electron cryomicroscopy. *Annual Review of Biomedical Data Science*, 3:163–190, 2020.

[2] Sjors HW Scheres, Haixiao Gao, Mikel Valle, Gabor T Herman, Paul PB Eggermont, Joachim Frank, and Jose-Maria Carazo. Disentangling conformational states of macromolecules in 3d-em through likelihood optimization. *Nature Methods*, 4(1):27–29, 2007.

[3] Roy R Lederman, Joakim Andén, and Amit Singer. Hyper-molecules: On the representation and recovery of dynamical structures, with application to flexible macro-molecular structures in cryo-em. *arXiv preprint arXiv:1907.01589*, 2019.

[4] Ellen D Zhong, Tristan Bepler, Bonnie Berger, and Joseph H Davis. CryoDRGN: reconstruction of heterogeneous cryo-EM structures using neural networks. *Nature Methods*, 18(2):176–185, 2021.

[5] Ali Punjani and David J Fleet. 3DFlex: Determining structure and motion of flexible proteins from cryo-EM. *Nature Mthods*, 20(6):860–870, 2023.

[6] Ellen D Zhong, Adam Lerer, Joseph H Davis, and Bonnie Berger. CryoDRGN2: Ab initio neural reconstruction of 3d protein structures from real cryo-EM images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4066–4075, 2021.

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

[8] Youssef SG Nashed, Frédéric 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4066–4076, 2021.

[9] 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. In *European Conference on Computer Vision*, pages 540–557. Springer, 2022.

[10] Yoon Kim, Sam Wiseman, Andrew Miller, David Sontag, and Alexander Rush. Semi-amortized variational autoencoders. In *International Conference on Machine Learning*, pages 2678–2687. PMLR, 2018.

[11] Abner Guzman-Rivera, Dhruv Batra, and Pushmeet Kohli. Multiple choice learning: Learning to produce multiple structured outputs. *Advances in Neural Information Processing Systems*, 25, 2012.

[12] Stefan Lee, Senthil Purushwalkam Shiva Prakash, Michael Cogswell, Viresh Ranjan, David Crandall, and Dhruv Batra. Stochastic multiple choice learning for training diverse deep ensembles. *Advances in Neural Information Processing Systems*, 29, 2016.

[13] Christian Rupprecht, Iro Laina, Robert DiPietro, Maximilian Baust, Federico Tombari, Nassir Navab, and Gregory D Hager. Learning in an uncertain world: Representing ambiguity through multiple hypotheses. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3591–3600, 2017.

[14] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems*, 33:7462–7473, 2020.

[15] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems*, 33:7537–7547, 2020.

[16] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[17] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[18] Devon Hjelm, Russ R Salakhutdinov, Kyunghyun Cho, Nebojsa Jojic, Vince Calhoun, and Junyoung Chung. Iterative refinement of the approximate posterior for directed belief networks. *Advances in Neural Information Processing Systems*, 29, 2016.

[19] Rahul Krishnan, Dawen Liang, and Matthew Hoffman. On the challenges of learning with inference networks on sparse, high-dimensional data. In *International Conference on Artificial Intelligence and Statistics*, pages 143–151. PMLR, 2018.

[20] Joe Marino, Yisong Yue, and Stephan Mandt. Iterative amortized inference. In *International Conference on Machine Learning*, pages 3403–3412. PMLR, 2018.

[21] D Goodsell. PDB-101 Molecule of the Month: Hsp90. 2008.

[22] Clemens Plaschka, Pei-Chun Lin, and Kiyoshi Nagai. Structure of a pre-catalytic spliceosome. *Nature*, 546(7660):617–621, 2017.

[23] Wilson Wong, Xiao-chen Bai, Alan Brown, Israel S Fernandez, Eric Hanssen, Melanie Condron, Yan Hong Tan, Jake Baum, and Sjors HW Scheres. Cryo-EM structure of the Plasmodium falciparum 80S ribosome bound to the anti-protozoan drug emetine. *Elife*, 3:e03080, 2014.

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[25] Thomas D Goddard, Conrad C Huang, Elaine C Meng, Eric F Pettersen, Gregory S Couch, John H Morris, and Thomas E Ferrin. UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Science*, 27(1):14–25, 2018.

[26] Marin Van Heel and Michael Schatz. Fourier shell correlation threshold criteria. *Journal of Structural Biology*, 151(3):250–262, 2005.

[27] Axel Levy, Gordon Wetzstein, Julien NP Martel, Frederic Poitevin, and Ellen Zhong. Amortized inference for heterogeneous reconstruction in cryo-EM. *Advances in Neural Information Processing Systems*, 35:13038–13049, 2022.

[28] 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, 2010.

[29] 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, 2017.

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

[31] 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, 1994.

[32] 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, 1996.

[33] Sjors HW Scheres. A Bayesian view on cryo-EM structure determination. *Journal of Molecular Biology*, 415(2):406–418, 2012.

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

[35] Eugene L Lawler and David E Wood. Branch-and-bound methods: A survey. *Operations Research*, 14(4):699–719, 1966.

[36] Dan Rosenbaum, Marta Garnelo, Michal Zielinski, Charlie Beattie, Ellen Clancy, Andrea Huber, Pushmeet Kohli, Andrew W Senior, John Jumper, Carl Doersch, et al. Inferring a continuous distribution of atom coordinates from cryo-EM images using VAEs. *arXiv preprint arXiv:2106.14108*, 2021.

[37] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[38] Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, Yasemin Altun, and Yoram Singer. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6(9), 2005.

[39] Stefan Lee, Senthil Purushwalkam, Michael Cogswell, David Crandall, and Dhruv Batra. Why m heads are better than one: Training a diverse ensemble of deep networks. *arXiv preprint arXiv:1511.06314*, 2015.

[40] Kimin Lee, Changho Hwang, KyoungSoo Park, and Jinwoo Shin. Confident multiple choice learning. In *International Conference on Machine Learning*, pages 2014–2023. PMLR, 2017.

[41] Kai Tian, Yi Xu, Shuigeng Zhou, and Jihong Guan. Versatile multiple choice learning and its application to vision computing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6349–6357, 2019.

[42] Victor Letzelter, Mathieu Fontaine, Mickaël Chen, Patrick Pérez, Slim Essid, and Gael Richard. Resilient multiple choice learning: A learned scoring scheme with application to audio scene analysis. *Advances in Neural Information Processing Systems*, 36, 2024.

[43] Simon Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R Ledsam, Klaus Maier-Hein, SM Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic U-Net for segmentation of ambiguous images. *Advances in Neural Information Processing Systems*, 31, 2018.

[44] Eddy Ilg, Ozgun Cicek, Silvio Galesso, Aaron Klein, Osama Makansi, Frank Hutter, and Thomas Brox. Uncertainty estimates and multi-hypotheses networks for optical flow. In *European Conference on Computer Vision*, pages 652–667, 2018.

[45] Ye Yuan and Kris Kitani. Diverse trajectory forecasting with determinantal point processes. *arXiv preprint arXiv:1907.04967*, 2019.

[46] Benjamin Biggs, David Novotny, Sebastien Ehrhardt, Hanbyul Joo, Ben Graham, and Andrea Vedaldi. 3D Multi-bodies: Fitting sets of plausible 3d human models to ambiguous image data. *Advances in Neural Information Processing Systems*, 33:20496–20507, 2020.

[47] Earl J Kirkland. *Advanced computing in electron microscopy*, volume 12. Springer, 1998.

[48] Ronald N Bracewell. Strip integration in radio astronomy. *Australian Journal of Physics*, 9(2):198–217, 1956.

[49] Todd K Moon. The expectation-maximization algorithm. *IEEE Signal processing magazine*, 13(6):47–60, 1996.

[50] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[51] Julian Panetta. Optimizing over SO(3). Technical report, University of California, Davis, 2018.

[52] F Sebastian Grassia. Practical parameterization of rotations using the exponential map. *Journal of Graphics Tools*, 3(3):29–48, 1998.

[53] Alexandra C Walls, Young-Jun Park, M Alejandra Tortorici, Abigail Wall, Andrew T McGuire, and David Veesler. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell*, 181(2):281–292, 2020.

[54] Krzysztof M Gorski, Eric Hivon, Anthony J Banday, Benjamin D Wandelt, Frode K Hansen, Mstvos Reinecke, and Matthia Bartelmann. HEALPix: A framework for high-resolution discretization and fast analysis of data distributed on the sphere. *The Astrophysical Journal*, 622(2):759, 2005.

Supplementary Material

Improving Ab-Initio Cryo-EM Reconstruction with Semi-Amortized Pose Inference

A Related work

Cryo-EM reconstruction. Methods for cryo-EM reconstruction can be categorized as either homogeneous or heterogeneous. Homogeneous techniques [7, 8, 9] assume a rigid structure while heterogeneous ones [4, 6, 5, 27] allow for conformational variation. We focus on homogeneous reconstruction, but our optimization framework could be extended to heterogeneous data as well.

Early reconstruction techniques rely on common-lines [28, 29, 30] or projection mapping [31, 32] to select optimal poses. Other works [33, 34] frame the reconstruction problem in the context of maximum a posteriori (MAP) estimation, and jointly reconstructs poses and structure via expectation maximization (EM). We compare our approach to cryoSPARC [7], a state-of-the-art method that uses stochastic gradient descent and a branch-and-bound search [35] for ab initio reconstruction and pose estimation. Like these methods, our auto-decoding stage directly optimizes pose of every image.

More recently, amortized inference techniques have been proposed for pose estimation [36, 8, 9, 27]. These techniques avoid explicit per-image pose optimization; instead, they train an auto-encoder or variational one [37] to associate each particle image with a predicted pose [8]. One challenge is that the auto-encoders can become stuck in local optima during training [8]. To address this issue, cryoAI [9] produces two pose estimates per image coupled with a symmetrized loss function that penalizes the best one. We build on this concept by adopting a multi-head neural architecture as the encoder to output multiple plausible pose candidates and avoid local optima.

Multi-choice learning (MCL). Inspired by scenarios where a set of hypotheses needs to be generated to account for uncertainty in the prediction task, MCL [11] was introduced in a supervised setup to learn multiple structured-outputs with SSVMs [38]. Their motivating question was: *can we learn to produce a set of plausible hypotheses?* To address this, they define an"oracle" loss in which only *the most accurate* output pays the penalty. This loss is minimized even if there is only a single accurate prediction in the set. The early follow-up work [39, 12] uses the same loss to learn a deep CNN ensemble composed of M heads with a shared backbone network. Importantly, they show that the ensemble-mean loss hurts prediction diversity across different heads, while training with the "oracle" loss yields specialized heads. Variations have since been proposed to mitigate hypothesis collapse or overconfidence issues in MCL by modifying the loss or applying learnable probabilistic scoring schemes [40, 41, 42, 13]. MCL has been used to mitigate the ambiguity in several tasks including image segmentation [43], optical-flow estimation [44], trajectory forecasting [45], human pose and shape estimation [13, 46]. In our work, we use the "oracle" loss in context of auto-encoder which supervises the pose encoder indirectly through projections provided by the decoder.

B Problem Definition

The image formation model for cryo-EM is often well-approximated using the weak-phase object model [47]. This model assumes the 3D structure is an unknown density map, $V : \mathbb{R}^3 \rightarrow \mathbb{R}_{\geq 0}$, represented under a canonical orientation. Cryo-EM images, $\{I_i\}_{i=1}^N$, are approximated as orthographic projections of the 3D map that are oriented and shifted by unknown rotation $R_i \in SO(3)$ and in-plane translation $t_i = (t_x, t_y) \in \mathbb{R}^2$. Formally,

$$I_i(x, y) = [g_i \star (S_{t_i} \mathcal{P}_{R_i} V)](x, y) + n(x, y) \, , \tag{3}$$

where $\mathcal{P}_{R_i}(\cdot)$ is the linear operator computing the integral along the optical axis, z, over the input density map rotated by R_i, and S is the shift operator. The projection is convolved with the image-specific point-spread function (PSF), g_i, and corrupted by additive noise n. It is common to assume that n follows a zero-mean white (or colored) Gaussian distribution.

By the Fourier slice theorem [48], the Fourier transform of a projection is equal to a central slice through the density map's 3D Fourier spectrum. Consequently,

$$\hat{I}_i(\omega_x, \omega_y) = \hat{g}_i \hat{S}_{t_i} (\hat{\mathcal{P}}_{R_i} \hat{V})(\omega_x, \omega_y) + \hat{n}(\omega_x, \omega_y) \, , \tag{4}$$

where $\hat{I}$ and $\hat{V}$ denote the 2D and 3D Fourier transforms of the image and the density map. The slice perpendicular to the projection is computed by $(\hat{\mathcal{P}}_{R_i}\hat{V})$. The translation by S_{t_i} becomes a phase shift operator $\hat{S}_{t_i}$, and convolution with g_i is equivalent to element-wise multiplication with, $\hat{g}_i$, the contrast transfer function (CTF). The noise $\hat{n}$ remains zero-mean Gaussian.

Under this model, given the structure $\hat{V}$, the negative log-likelihood of observing image $\hat{I}_i$ with noise variance σ^2 and poses (R_i, t_i) is

$$\mathcal{L} = -\frac{1}{2\sigma^2} \sum_{\omega_x,\omega_y} [\hat{g}_i \hat{S}_{t_i}(\hat{\mathcal{P}}_{R_i}\hat{V})(\omega_x,\omega_y) - \hat{I}_i(\omega_x,\omega_y)]^2 \,. \tag{5}$$

Ab-initio reconstruction methods [7, 8, 9, 6] solve jointly for the unknown structure $\hat{V}$ and poses (R_i, t_i). They often follow an Expectation-Maximization (EM) [49, 33] procedure in which the E-step aligns images with the structure yielding pose estimates (R_i, t_i), and then in the M-step the volume $\hat{V}$ is updated by minimizing the negative log-likelihood in Eq. 5. Since errors in pose estimates lead to blurry reconstructions, accurate pose estimates are crucial to finding high-resolution structures. As discussed above, poses are either optimized through search and projection matching [7, 6] or estimated by an encoder network [9, 8, 27].

C Pose Optimization

For the auto-encoding stage, we follow cryoAI and design the convolutional backbone to output the six-dimensional representation commonly referred to as $S2S2$. To compute the rotation matrix from this representation, the 6D vector is split into two 3D vectors and normalized, denoted as $v_1, v_2 \in \mathbb{R}^3$. We then compute the cross product between them, $v_3 = v_1 \times v_2$, yielding a new unit vector. Selecting v_1 and v_3 as the first and third columns of the target rotation matrix, we compute $\tilde{v}_2 = v_3 \times v_1$, which is another unit vector orthogonal to both v_1 and v_3. Then, $R = [v_1, \tilde{v_2}, v_3]$.

Once switched to direct optimization, we change parameterization to axis-angle representation. During auto-decoding, we alternate between five iterations of pose SGD updates and one iteration of volume update. To update poses, we keep the volume fixed and optimize for the negative log-likelihood (Eq. 5) with respect to the pose parameters. We define the new pose estimate based on the current one as follows:

$$R_{t+1} = R_\delta R_t \tag{6}$$

where R_δ is an infinitesimal rotation matrix perturbing the current estimate. The perturbation matrix R_δ is parameterized by axis-angle representation. By a single vector $\boldsymbol{\omega} \in \mathbb{R}^3$, one can represent both the axis $||\boldsymbol{\omega}||$ and the angle $0 < \frac{\boldsymbol{\omega}}{||\boldsymbol{\omega}||} < \pi$ for any given rotation. Using Rodrigues formula, the perturbation matrix $R_\delta(\boldsymbol{\omega})$ can be parameterized as a function of $\boldsymbol{\omega}$. To find the optimal $\boldsymbol{\omega}$, one can initialize it with zero vector, and then use automatic differentiation [50] in pytorch to compute the gradient with respect to $\boldsymbol{\omega}$ and make updates using Adam [24]. However, a naive implementation of the function $R_\delta(\boldsymbol{\omega})$ would lead to numerically unstable calculations of the partial derivative $\frac{\partial R_\delta}{\partial \boldsymbol{\omega}}$. In fact, there is a singularity at the zero vector, and the partial derivative involves terms that are unstable around the origin. Formally, the derivative of i-th column of the rotation matrix R_δ with respect to the vector $\boldsymbol{\omega}$ is [51, 52],

$$\begin{aligned}
\frac{\partial R_\delta^{(i)}}{\partial \boldsymbol{\omega}} = &- \left(\mathbf{e}^{(i)} \otimes \boldsymbol{\omega} + [\mathbf{e}^{(i)}]_\times\right) \frac{\sin(\|\boldsymbol{\omega}\|)}{\|\boldsymbol{\omega}\|} + [(\boldsymbol{\omega} \cdot \mathbf{e}^{(i)})I + \boldsymbol{\omega} \otimes \mathbf{e}^{(i)}] \left(\frac{1 - \cos(\|\boldsymbol{\omega}\|)}{\|\boldsymbol{\omega}\|^2}\right) \\
&+ (\boldsymbol{\omega} \otimes \boldsymbol{\omega}) \left((\boldsymbol{\omega} \cdot \mathbf{e}^{(i)}) \frac{2\cos(\|\boldsymbol{\omega}\|) - 2 + \|\boldsymbol{\omega}\| \sin(\|\boldsymbol{\omega}\|)}{\|\boldsymbol{\omega}\|^4}\right) \\
&+ [(\boldsymbol{\omega} \times \mathbf{e}^{(i)}) \otimes \boldsymbol{\omega}] \frac{\|\boldsymbol{\omega}\| \cos(\|\boldsymbol{\omega}\|) - \sin(\|\boldsymbol{\omega}\|)}{\|\boldsymbol{\omega}\|^3} \,.
\end{aligned}$$

where $\otimes$ and $\times$ are tensor and cross products, respectively. $\mathbf{e}^{(i)}$ is the i-th standard basis in 3D and $[\mathbf{v}]_\times$ denotes the cross product matrix for the vector $\mathbf{v}$. In all four terms, there are scalars such as $\frac{\sin(\|\boldsymbol{\omega}\|)}{\|\boldsymbol{\omega}\|}$ or $\frac{1-\cos(\|\boldsymbol{\omega}\|)}{\|\boldsymbol{\omega}\|^2}$ that evaluate to $\frac{0}{0}$ at zero angle $\boldsymbol{\omega} = 0$. Similar to [51], for $||\boldsymbol{\omega}|| \ll 1$, we

Table 2: The reconstruction time per epoch, GPU memory, and number of parameters for different methods averaged across three runs. GPU memory is recorded separately for the encoding and decoding modules. Also, two numbers provided for semi-amortized method corresponds to auto-encoding and auto-decoding stages, respectively. In CryoAI-explicit, the implicit decoder of CryoAI is replaced with an explicit decoder. Since we store rotations in 3D representation during direct optimization, the number of parameters of pose module is $N \times 3$ with N denoting number of images in millions.

Model	Time (s)	GPU Mem. (GB)		# Params (M)
		Encoding	Decoding	
Semi-Amortized	99.76, 81.61	1.17, -	2.22, 0.35	13.01, $4.29 + N \times 3$
Fully-Amortized	99.75	1.17	2.22	13.01
CryoAI-explicit	142.40	2.32	0.75	9.05
CryoAI	622.39	2.78	14.05	5.09

substitute these terms with their numerically robust Taylor expansion, for instance,

$$\frac{\sin(\|\boldsymbol{\omega}\|)}{\|\boldsymbol{\omega}\|} = 1 - \frac{||\boldsymbol{\omega}||^2}{6} + O(||\boldsymbol{\omega}||^4) \,,$$
$$\frac{1-\cos(\|\boldsymbol{\omega}\|)}{\|\boldsymbol{\omega}\|^2} = \frac{1}{2} - \frac{||\boldsymbol{\omega}||^2}{24} + O(||\boldsymbol{\omega}||^4) \,.$$

We implement a differentiable and numerically stable version of the function $R_\delta(\boldsymbol{\omega})$ in pytorch and use it in our pose estimation module.

D Structure Decoder and Ablation Study

Recent works [9, 4] use coordinate networks [15, 16, 14] to implicitly model the Fourier representation of the 3D structure. Instead, we couple the pose estimation module with an explicit parameterization of the structure in the Fourier domain. The explicit representation is less computationally expensive than an MLP to evaluate and update. Also, this choice is motivated by the fact the implicit decoder needs to be queried multiple times for each image with the multi-head encoder.

We parameterize the volume using the Hartley representation [4]. The Fourier and Hartley transforms, respectively denoted as $F(\omega)$ and $H(\omega)$, are related as

$$H(\omega) = \mathcal{R}[F(\omega)] - \mathcal{I}[F(\omega)], \tag{7}$$

where ω denotes the frequency coordinate and $\mathcal{R}$ and $\mathcal{I}$ are the real and imaginary part, respectively. The Hartley representation is real-valued, and so more memory efficient to use than storing complex-valued Fourier coefficients. To account for high dynamic range of the Hartley coefficients, we assume the Hartley field is decomposed into mantissa, $m(\omega)$ and exponent $e(\omega)$ fields [9] as,

$$H(\omega) = m(\omega) \times \exp(e(\omega)). \tag{8}$$

This decomposition restricts the range of values for $m(\omega)$ and $e(\omega)$ and makes the reconstruction less sensitive to the initialization of the field.

We perform an ablation study on the decoder, detailed in Table 2, as well as a comparison with baselines in terms of reconstruction time per epoch, GPU memory usage, and number of parameters. As our method consists of two stages, we report numbers for each stage separately. In the first stage, the encoder has H=7 heads, while in the second stage, it is replaced with a pose module with size depending on the number of particles (N). To facilitate comparison, we include another baseline, cryoAI-explicit, in which the implicit decoder is replaced by an explicit one as in our method. The explicit and implicit decoders have 4.29 and 0.33 million parameters, respectively. Despite a larger decoder, our method is 6x faster and uses 5x less memory compared to cryoAI. Importantly, swapping the implicit decoder with an explicit one (cryoAI-explicit) significantly drops time and memory, indicating that the implicit decoder is a major computational and memory bottleneck. Yet, cryoAI-explicit uses 2x more GPU memory for encoding and is 1.5x slower than our method as it performs early input augmentation and runs the entire encoder twice per image. Our method, by

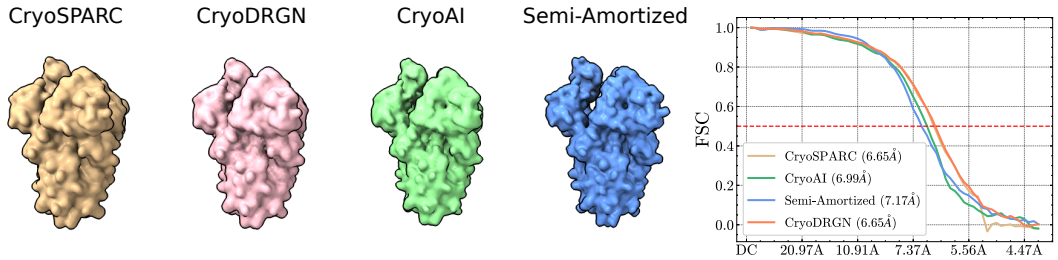

Figure 6: Qualitative and quantitative comparison of reconstructions of our semi-amortized method with cryoAI [9], cryoDRGN [6] and cryoSPARC [7] on Spike dataset. **(Left)** Final 3D reconstructions are depicted using ChimeraX [25]. **(Right)** FSC curves are visualized for quantitative comparison. The red dashed lines show the standard threshold levels of 0.5 and 0.143 to report the resolution (in Angstrom) for synthetic and real data, respectively.

augmenting the encoder head, saves memory and time during pose encoding. Finally, the amortized baseline, which uses an explicit decoder coupled with a multi-head encoder (H=7), uses more memory in decoding (negligible vs implicit decoder) and runs slower than the direct optimization stage of the semi-amortized method.

E Additional Results on Spike Protein

We further compare our method with others on an additional synthetic dataset based on SARS-CoV-2 Spike protein [53] (2.13 Å). We follow the same procedure to simulate projections PDB deposited atomic models and create a dataset of $N = 50,000$ noisy CTF-corrupted projections of size $L = 128$ with SNR $= 0.1$. The qualitative comparison of reconstructions obtained by our method with cryoAI, cryoDRGN, and cryoSPARC is provided in the left of Fig. **??**. For quantitative comparison, the gold-standard Fourier Shell Correlation (FSC) is visualized (Fig. 2, right).

We also inspect the benefit of semi-amortization for Spike dataset (Fig. 7, left). Similar to results obtained by Spliceosome and HSP, as our method switches to direct pose optimization, the error in pose drops quickly, whereas the fully-amortized baseline exhibits slow convergence. This clearly shows the superiority of auto-decoding compared to auto-encoding during the later stages of optimization.

In Fig. 7, we also visualize and compare the trajectory of pose estimates of auto-decoding and auto-encoding approaches. Poses obtained by auto-decoding (blue dots) show stable convergence to the optimal point (highlighted area) whereas those inferred by auto-encoding (red dots) frequently oscillate. In fact, the encoder is a globally parameterized function which might be too restrictive, yielding sub-optimal pose predictions. Therefore, poses inferred in an amortized fashion might fail to consistently converge to the optimal point. On other hand, direct optimization during auto-decoding is intuitively more flexible as it is performed separately and locally for each image, exhibiting more stable convergence.

F Specialization of Encoder Heads

A natural question about the multi-head encoder is: how each head does take part in pose encoding process? To address this, using the synthetic datasets, we conduct an experiment with our multi-head architecture ($M = 4$) and visualize the performance of each head on different regions of $SO(3)$ space. In particular, as before, we define a uniform grid on the unit sphere using HEALPix [54] and assign images to their corresponding cells based on the view-direction. Now, for all images end up in the same cell, we compute the average rotation error and visualize it separately for each head. As shown in Fig. 8, all heads actively participate in pose estimation and they are able to specialize in prediction of poses for images with certain view-direction. A similar result has been provided in prior work on MCL [39, 12], to show that minimizing the error made by the best prediction ("oracle" loss) encourages diversity in deep ensembles. In our problem, by optimizing a "winner-takes-all" loss, the whole burden of pose estimation is no longer on a single network but it gets divided between multiple heads as separate predictors.

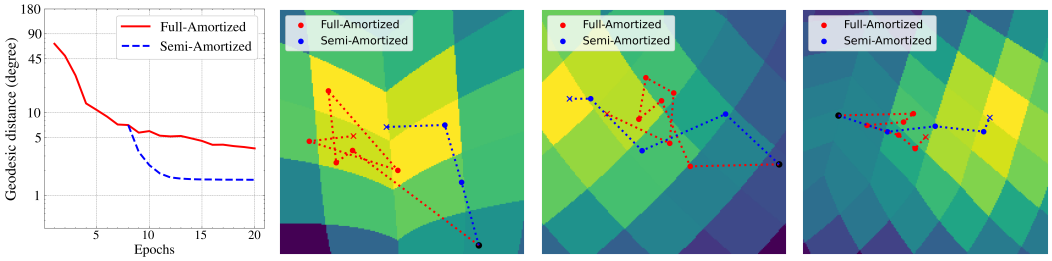

Figure 7: Comparison of fully- vs. semi-amortized methods in pose optimization on Spike dataset. **(Left)** The mean geodesic distance between predicted poses and ground-truths is shown across epochs. Switching from amortized inference to direct optimization by our method **(blue)** leads to faster pose convergence compared to fully-amortized inference **(red)**. **(Right)** For qualitative comparison, the approximate log posterior for three particles is visualized as a view-direction distribution on a unit sphere. After Gnomonic projection, the neighborhood of the mode of interest is visualized. **Black** dot cross the initial point of optimization.

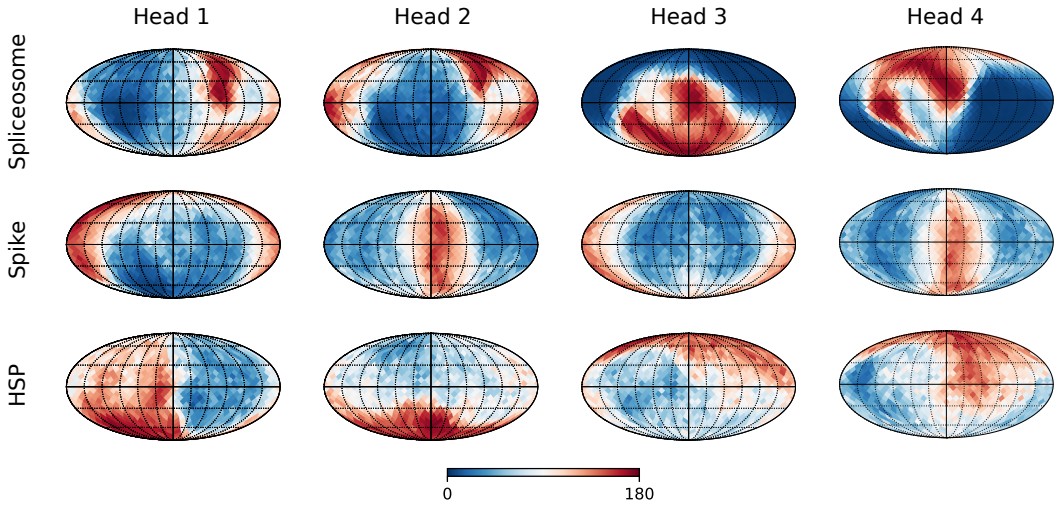

Figure 8: Average rotation error visualized over the unit sphere for different heads of our multi-head pose encoder ($M = 4$). The sphere is uniformly divided into cells using HEALPix [54] and based on their ground-truth view-direction, images are assigned to corresponding cells. For each cell, the average rotation error is visualized, showing diverse behavior of different heads across the space. Blue and red colors show low and high error regions, respectively. Error ranges from zero to 180 degrees.

G Semi-Amortized vs. Fully-Amortized Convergence

To validate the advantages of direct pose optimization in our semi-amortized method, we further show more qualitative examples of paths taken by pose estimates over the optimization landscape during reconstruction in Fig. 9. For both methods, optimization start from the same point marked by **black** dot in the vicinity of the distribution mode. It will then continue in paths colored in **blue** and **red** for semi-amortized and full-amortized methods, respectively. We observe in all examples that iterative updates by stochastic gradient descent demonstrate a stable convergence toward the optima while poses obtained by amortized inference show unstable behavior around the mode.

H Visualizations of the Multi-Modal Pose Posterior

Through more examples (Fig. 10), we demonstrate that cryoAI fails to handle ambiguity in pose estimation on HSP dataset. The visualization shows that pose estimates by cryoAI become stuck in incorrect modes whereas our pose encoder with multi-head architecture is able to return a pose candidate that captures the correct mode.

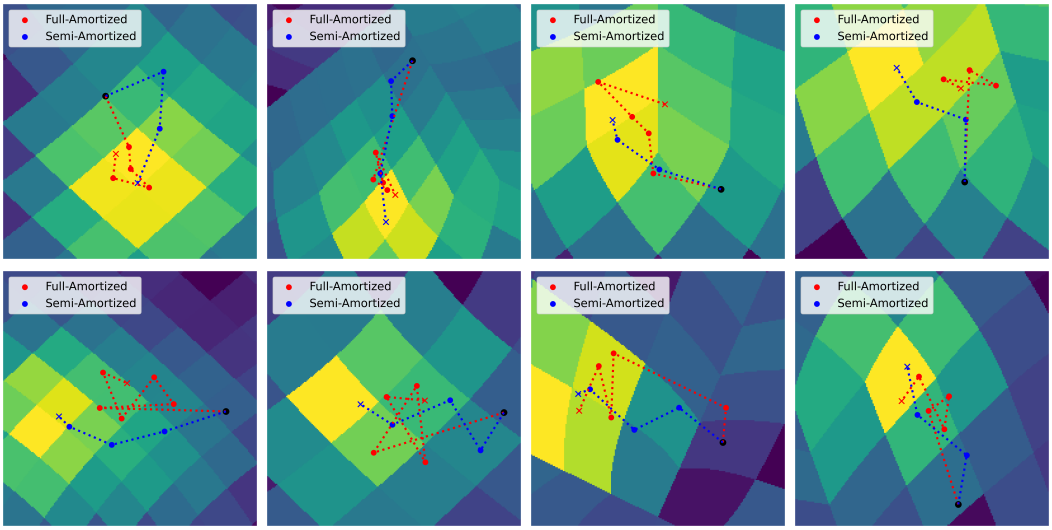

Figure 9: Using four examples per dataset, the behavior of fully-amortized and semi-amortized pose inference methods are compared. Two rows correspond to Spike and Spliceosome datasets, respectively. Each plot shows the approximate log pose posterior, marginalized over in-plane rotations represented as a heat map on a uniform grid over the unit sphere S^2. Gnomonic projection to 2D is also applied, followed by zooming on the proximity of the mode of interest. **Black** dot is the starting point while **blue** dots and **red** dots show poses estimated by fully- and semi-amortized methods, respectively.

I Videos

In the supplementary package, using ChimeraX [25] we provide videos that show reconstructions obtained by semi-amortized method, cryoAI and cryoSPARC on all synthetic and experimental datasets. In these 3D visualizations, we rotate the structure to show the resolved structure from different views.

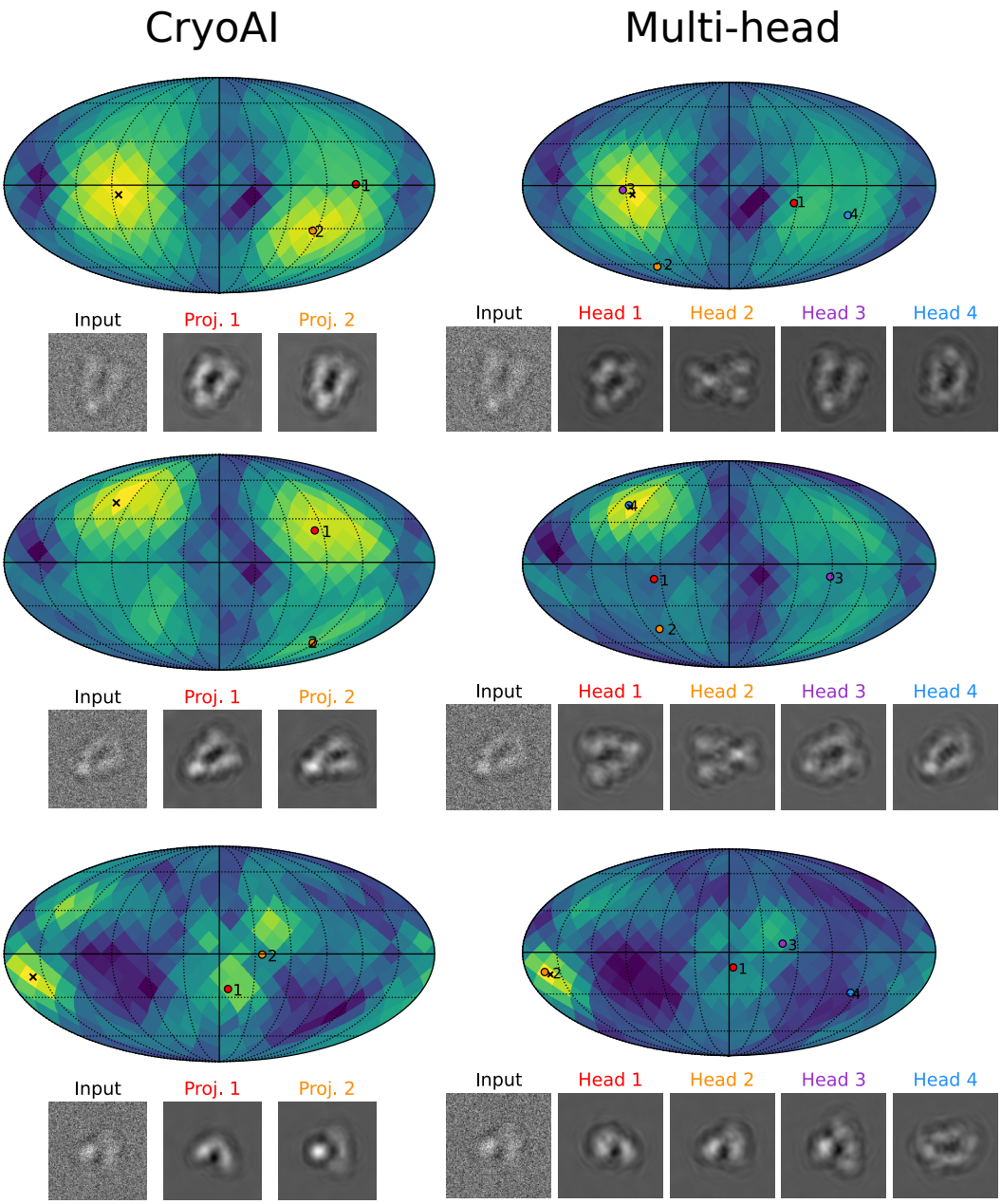

Figure 10: The approximate log posterior of view-direction visualized on the unit sphere with highlighted areas showing modes of the distribution. CryoAI [9] and our multi-head encoders provide two and four pose estimates, respectively, which are marked with colored dots on the sphere (the order of poses is arbitrary). The corresponding projections are also illustrated. CryoAI cannot identify the correct mode of pose distribution.