# CryoSPIN: 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

Cryo-EM is an increasingly popular method for determining the atomic resolution 3D structure of macromolecular complexes (eg, proteins) from noisy 2D images captured by an electron microscope. The computational task is to reconstruct the 3D density of the particle, along with 3D pose of the particle in each 2D image, for which the posterior pose distribution is highly multi-modal. Recent developments in cryo-EM have focused on deep learning for which amortized inference has been used to predict pose. Here, we address key problems with this approach, and propose a new semi-amortized method, cryoSPIN, in which reconstruction begins with amortized inference and then switches to a form of auto-decoding to refine poses locally using stochastic gradient descent. Through evaluation on synthetic datasets, we demonstrate that cryoSPIN is able to handle multi-modal pose distributions during the amortized inference stage, while the later, more flexible stage of direct pose optimization yields faster and more accurate convergence of poses compared to baselines. On experimental data, we show that cryoSPIN outperforms the state-of-the-art cryoAI in speed and reconstruction quality. Project Webpage.

## 1 Introduction

Single particle electron cryo-microscopy (cryo-EM) has gained popularity among structural biologists as a powerful experimental method for determining the 3D structure of macromolecular complexes, such as proteins and viruses, unlocking our understanding of biological function at the scale of the cell. Thanks to pioneering advances in hardware and data processing techniques, cryo-EM has enabled reconstruction of challenging structures at atomic or near-atomic resolution [1].

During a cryo-EM experiment, $10^4$–$10^7$ particle images, each containing an instance of the target bio-molecule, are acquired using a transmission electron microscope. From that *particle stack*, the goal is to reconstruct the unknown 3D density map [2]. This *ab-initio* reconstruction task presents some challenges. First, the 3D pose (orientation and position) of the particle in each image is unknown, and must be inferred along with the 3D structure. Second, low electron exposures are used to limit radiation damage to the particles (e.g., see Fig. 1). But this reduces the signal-to-noise ratio (SNR), obscuring high-resolution image detail and hindering pose and structure estimation. Third, bio-molecules are typically non-rigid and exhibit structural variations within a sample. Hence, for such heterogeneous datasets, it is crucial to account for the conformational variability in order to obtain high-resolution reconstruction [3–6].

38th Conference on Neural Information Processing Systems (NeurIPS 2024).


https://doi.org/10.52202/079017-1773

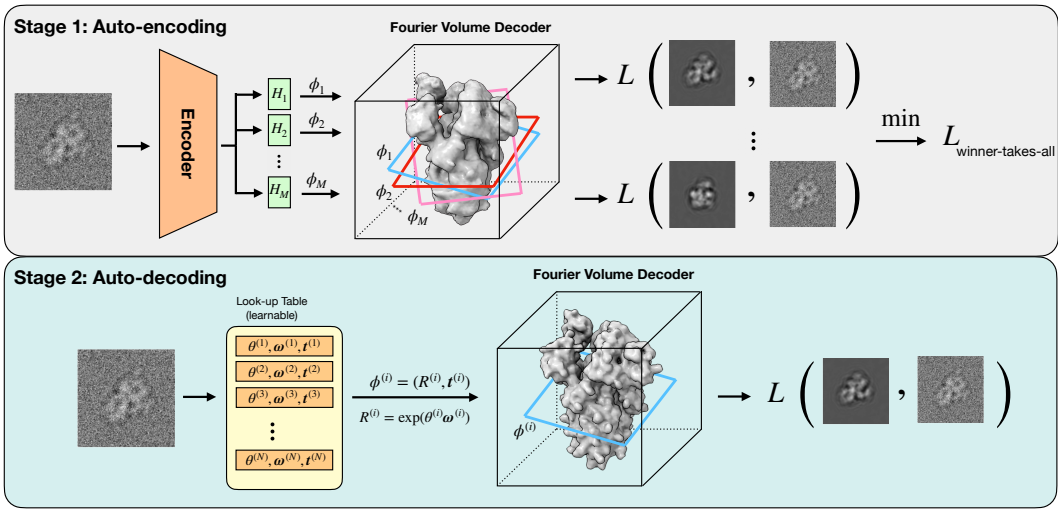


Figure 1: CryoSPIN consists of two stages: (i) an auto-encoding stage where an image encoder equipped with multiple heads maps the input image to the pose candidate set $\{\phi_1, \ldots \phi_M\}$, followed by computing projections by slicing through the volume decoder in Fourier space based on the pose set. These projections are compared with the input image and the one with the minimum error is used. (ii) An auto-decoding stage where pose parameters in axis-angle representation are stored for all images. The same volume decoder is used to obtain projections, and the reconstruction loss is computed for a single projection.

The development of methods in cryo-EM has recently focused on deep learning. CryoDRGN [5] proposed an image encoder-volume decoder architecture to model continuous heterogeneity, but the poses were assumed to be known. CryoDRGNv2 [7] introduced hierarchical pose search, comprising grid search followed by a form of branch-and-bound (BnB) optimization, akin to cryoSPARC [8]. More recently, cryoPoseNet [9] and cryoAI [10] introduced amortized pose inference, to avoid the expense of orientation matching for each particle image. They employ a convolutional neural network (CNN) to map an input image to a pose estimate. Nevertheless, although efficient, amortized inference may not accurately represent multi-modal posterior pose distributions, which occur frequently early in reconstruction before the structure is estimated well. As such, compared to previous methods like cryoSPARC [8] and cryoDRGNv2 [7] that consider many pose candidates, amortized methods sometimes fail to identify the correct mode. Moreover, amortization confines the pose search to a global parametric function (the encoder) which can lead to slow convergence.

Here, we introduce a new approach, called cryoSPIN, to ab-initio homogeneous reconstruction that is able to handle multi-modal pose distributions with a tailored encoder, and accelerates pose optimization with semi-amortization [11]. First, we perform amortized pose inference through a shared multi-choice encoder that maps each input image to multiple pose estimates (Fig. 1, top). In contrast to cryoPoseNet [9] and cryoAI [10], which use pose encoders that produce one or two estimates, we attach multiple pose predictor heads to a shared CNN feature extractor to predict multiple plausible poses for each image. By design, our multi-output encoder is able to account for pose uncertainty and encourage exploration of pose space during the initial stages of reconstruction. For the 3D decoder, unlike cryoDRGN and cryoAI that use MLPs with computationally expensive feed-forward implicit networks, we adopt an explicit parameterization to further accelerate the reconstruction. To train the encoder-decoder architecture, we apply a "winner-takes-all" loss in which the 3D decoder is queried to obtain a 3D-to-2D projection for each predicted pose. Inspired by the loss function in Multi-Choice Learning [12–14], we select the projection with the lowest image reconstruction error to determine the loss.

During the course of optimization, as higher resolution details are resolved, the pose posterior becomes predominantly uni-modal. Thus, the pose search can be narrowed down to a neighborhood containing the most likely mode. In this stage, we propose to transition from auto-encoding to auto-decoding (Fig. 1, bottom). In particular, for each image, we choose the pose with the lowest reconstruction loss, iteratively refine it with stochastic gradient descent (SGD), and continue pose optimization

and reconstruction in an alternating fashion. This direct, per-image optimization procedure achieves faster convergence to more accurate poses than amortized inference alone, which relies on potentially sub-optimal predictions of the encoder as a parametric function.

We evaluate cryoSPIN in comparison to state-of-the-art methods, cryoSPARC [8], cryoAI [10], and cryoDRGN [7] on both synthetic and real experimental datasets. Using synthetic data, we find that semi-amortized inference noticeably accelerates the convergence of poses, and yields 3D reconstructions at similar or higher resolutions than baseline methods. We also show that our multi-choice encoder empirically accounts for multiple modes in the pose posterior. Finally, we apply cryoSPIN to an experimental dataset and obtain reconstructions with higher resolution than those from cryoAI, while matching the resolution of cryoSPARC reconstructions.

To summarize, we make following contributions.

- Building upon cryoAI, we develop a new encoder based on a multi-head architecture to return multiple plausible candidates to further mitigate pose ambiguity.
- We train the encoder coupled with an explicit volume decoder using a *winner-take-all loss*, which penalizes the pose hypothesis with least reconstruction error.
- Beyond the architecture and objective, we introduce *semi-amortized pose inference*, which begins with feed-forward amortized pose inference, followed by direct, per-particle pose optimization.
- In ab-initio reconstruction, our semi-amortized method, cryoSPIN, is faster than amortized method of cryoAI and achieves higher-resolution on synthetic and experimental datasets.

## 2 Related work

**Cryo-EM reconstruction.** Methods for cryo-EM reconstruction can be categorized as either homogeneous or heterogeneous. Homogeneous techniques [8–10] assume a rigid structure while heterogeneous ones [5, 7, 6, 15] 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 [16–18] or projection mapping [19, 20] to select optimal poses. Other works [21, 22] frame the reconstruction problem in the context of maximum a posteriori (MAP) estimation, and jointly reconstruct pose and structure via expectation maximization (EM). We compare our approach to cryoSPARC [8], a state-of-the-art method that uses stochastic gradient descent and a branch-and-bound search [23] 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 [24, 9, 10, 15]. These techniques avoid explicit per-particle pose optimization; instead, they train an auto-encoder or variational one [25] to associate each particle image with a predicted pose [9]. One challenge is that the auto-encoders can become stuck in local optima during training [9]. To address this issue, cryoAI [10] 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 [12] was introduced in a supervised setup to learn multiple structured-outputs with SSVMs [26]. 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 [27, 13] 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 [28–30, 14]. MCL has been used to mitigate the ambiguity in several tasks including image segmentation [31], optical-flow estimation [32], trajectory forecasting [33], human pose and shape estimation [14, 34]. 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.


https://doi.org/10.52202/079017-1773

## 3 Problem Definition

Image formation in cryo-EM is well-approximated using the weak-phase object model [35]. Under this model, the structure of interest is an unknown 3D density map, $V : \mathbb{R}^3 \to \mathbb{R}_{\geq 0}$, represented in some canonical coordinate frame. Cryo-EM images, $\{I_i\}_{i=1}^N$, are approximated as orthographic projections of the 3D map under some unknown coordinate transformation. We denote the unknown orientation by a rotation $R_i \in SO(3)$ and the unknown position in terms of an in-plane translation $t_i = (t_x, t_y) \in \mathbb{R}^2$. Formally,

$$I_i(x, y) = [g_i \star (S_{t_i} \mathcal{P}_{R_i} V)](x, y) + n(x, y) \, , \tag{1}$$

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 [36], the Fourier transform of a projection is equal to a central slice through the density map's 3D Fourier spectrum. That is,

$$\hat{I}_i(\omega_x, \omega_y) = \hat{g}_i \hat{S}_{t_i} (\hat{\mathcal{P}}_{R_i} \hat{V})(\omega_x, \omega_y) + \hat{n}(\omega_x, \omega_y) \, , \tag{2}$$

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 $\sigma^2$ and pose $(R_i, t_i)$ is

$$\mathcal{L} = -\frac{1}{2\sigma^2} \sum_{\omega_x, \omega_y} [\hat{g}_i \hat{S}_{t_i} (\hat{\mathcal{P}}_{R_i} \hat{V})(\omega_x, \omega_y) - \hat{I}_i(\omega_x, \omega_y)]^2 \, . \tag{3}$$

Ab-initio reconstruction methods [8–10, 7] solve jointly for the unknown structure $\hat{V}$ and per-particle poses $(R_i, t_i)$. They often follow an Expectation-Maximization (EM) [37, 21] 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. 3. 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 [8, 7] or estimated by an encoder network [10, 9, 15].

## 4 Methodology

We introduce cryoSPIN, a two-staged approach to ab-initio cryo-EM reconstruction, combining auto-encoding and auto-decoding. We start with amortized inference (Fig. 1, auto-encoding) where a shared encoder equipped with multiple heads provide a set of pose guesses. Multiple guesses enable multimodal inference in the initial stages of reconstruction when the 3D structure is poorly resolved. As reconstruction improves, the pose posterior becomes less uncertain, at which point we switch to direct pose optimization (Fig. 1, auto-decoding). The former enables handling the pose uncertainty early in reconstruction while the latter circumvents sub-optimality of encoder network leading to arguably more accurate poses and faster convergence. We couple our pose estimation module with an explicit volumetric decoder representing the 3D structure in Hartley space [5]. Our explicit model enables faster evaluation of projections compared to multiple passes through implicit neural representations [38–40]. The following sections discuss these components in detail.

### 4.1 Multi-choice encoder

With randomly initialized or a low-resolution reconstruction, the pose posterior contains multiple modes. Also, due to high levels of noise in images and near-symmetries in biological structures, pose estimation inherently involves high uncertainty. As a result, there exist several plausible poses for each image, and naive optimization or search is prone to local minima.

To account for uncertainty in pose estimation, we build upon cryoAI [10] and extend the encoder to return multiple candidate poses $\phi$. Formally, given the image $I_i \in \mathbb{R}^{H\times W}$, $M$ pairs of rotations and translations are computed as,

$$\phi_{i,j} \;=\; (R_{i,j}, t_{i,j}) \;=\; H_{\theta_j}(F_i), \quad 1 \leq j \leq M \tag{4}$$

where $R_{i,j} \in SO(3)$ and $t_{i,j} \in \mathbb{R}^2$ . The image-specific intermediate features, $F_i = \text{VGG}(I_i) \in \mathbb{R}^{C\times H\times W}$ are extracted by a shared convolutional backbone based on the VGG16 architecture [41], which are then supplied to $M$ separate pose predictor heads, $H_{\theta_j}, 1 \leq j \leq M$, yielding poses $\phi_{i,j}$. The heads are composed of fully-connected layers with distinct learnable weights, while the backbone is shared across all heads. Using multiple heads facilitates pose exploration and reduces uncertainty in the pose estimation process. Moreover, as $\theta_j$ are randomly initialized, different heads are free to specialize on subsets of pose space so that they collectively produce a set of likely poses $\{\phi_{i,1}, \ldots, \phi_{i,M}\}$.

Inspired by multi-choice learning [12–14], we optimize the encoder-decoder using a "winner-take-all" loss. For each image $I_i$ and predicted pose $\phi_{i,j}$, the negative log-likelihood, $\mathcal{L}_{i,j}$ in (Eq. 3), is computed, and the minimum is selected as the final loss for the corresponding image, i.e.,

$$\mathcal{L}_i = \min_j \mathcal{L}_{i,j} \; . \tag{5}$$

Minimizing this loss requires only one of the predicted poses to be accurate. Interestingly, we find that this approach produces heads that specialize to somewhat disjoint regions of pose space (see Supp. C), consistent with previous work showing that it improves diversity in prediction tasks [13, 27]. This 'divide-and-conquer' approach facilitates pose space exploration, which is crucial in order to avoid local minima during the early stages of reconstruction where uncertainty in the 3D structure is significant. CryoAI [10] can be viewed a special case of this formulation; it assigns two poses to each image as a consequence of input augmentation, and selects the best one with a symmetrized loss. In contrast, our approach augments the output of the encoder network with multiple heads, each providing a pose estimate.

### 4.2 Switching from auto-encoding to auto-decoding

As optimization progresses and fine-grained details of the 3D structure are resolved, the posterior pose variance tends to decrease, with the posterior approaching a unimodal distribution. At this point, the gap between the amortized and variational posterior distributions 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 the input image, may be limited in the accuracy of its predictions, and hence the approximations inherent in amortized inference become significant sources of error in the refinement of the 3D structure. In prior work [11, 42–44], a related issue, called the *amortization gap*, is characterized by the KL-divergence between the true and predicted (amortized) variational posteriors.

To address this issue, we adopt a semi-amortized inference scheme [11] 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 stochastic gradient descent. 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

$$s = \arg\min_j \mathcal{L}_{i,j} \; . \tag{6}$$

Subsequently, the pose and structure are optimized by coordinate descent using the negative log-likelihood (Eq. 3) as the objective function. See Supp. A for more details on rotation optimization.

### 4.3 Explicit representation as volume decoder

Recent works [10, 5] use coordinate networks [39, 40, 38] to implicitly model the Fourier representation of the 3D structure. Instead, we couple the pose estimation module with an explicit parameterization (3D voxel array) 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.


https://doi.org/10.52202/079017-1773

We parameterize the volume using the Hartley representation [5]. 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 $\omega$ 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 [10] 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.

## 5 Experiments

In what follows, we consider both synthetic and real datasets to empirically compare our semi-amortized ab-initio reconstruction method with state-of-the-art methods, cryoAI [10] and cryoDRGN [7] which use amortized inference for reconstruction, and cryoSPARC [8] which performs pose search separately for each particle image.

**Synthetic data.** We generate synthetic datasets by simulating the image formation process formalized in Sec. 3 using atomic models deposited in the Protein Data Bank (PDB). We compute ground-truth density maps of size $128^3$ for a heat shock protein (HSP) [45] (1.5 Å), pre-catalytic spliceosome [46] (4.33 Å), and SARS-CoV-2 spike protein [47] (2.13 Å). Then, $N = 50,000$ projections of size $L = 128$ are generated by randomly rotating and projecting the density map along the canonical z-axis. To simplify the analysis and interpretation of results (e.g. Figs. 4, 5), we consider synthetic particles are centered, and therefore omit the estimation of 2D translation and allowing us to visualize the pose posterior more easily. Finally, a random CTF is applied in the Fourier space and Gaussian noise is added yielding SNR $= 0.1$.

**Experimental data.** As a widely adopted experimental benchmark, we use the 80S ribosome dataset (EMPIAR-10028 [48]), containing 105,247 particle images of linear box size $L = 360$ with pixel size 1.34 Å. Following cryoAI and cryoDRGN, we downsample the images to $L = 128$ (3.76 Å) using cryoSPARC software [8], and randomly split the data into two halves and run the reconstruction methods independently on each (to enable Fourier Shell Correlation as a quantitative measure of the resolution of the 3D reconstruction). Unlike synthetic data, particles are not well-centered so it is required to estimate translation parameters as well as rotations. This is achieved by augmenting each head to return translation parameters as well.

**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 [49] 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 $0.05$. To be consistent with cryoAI, we use a batch size of $64$ and train for the same number of epochs (20 for synthetic, 30 for real data). We use the public cryoAI codebase, run cryoSPARC v4.4.0 [8] with default settings, and run cryoDRGN homogenous ab-initio job. Methods are implemented in Pytorch [50]. Experiments are run on a single NVIDIA A40 GPU.

### 5.1 Results

We first qualitatively compare final reconstructions of cryoSPIN with cryoAI [10], cryoDRGN [7], and cryoSPARC [8] on synthetic and real datasets (Fig. 2, left). We observe that cryoAI gets stuck in local minima when evaluated on experimental data (eg, 80S ribosome). In particular, we found that in the presence of unknown planar shift, cryoAI fails to estimate effective translation parameters, but after centering all particles via pre-processing, we could reproduce the cryoAI results. Hence, the provided results for cryoAI in Fig. 2 are obtained after this pre-processing step, whereas our method and others are given off-centered experimental images. Both cryoSPIN and cryoSPARC capture high-frequency details of the 3D structure on all datasets, whereas reconstructions by amortized

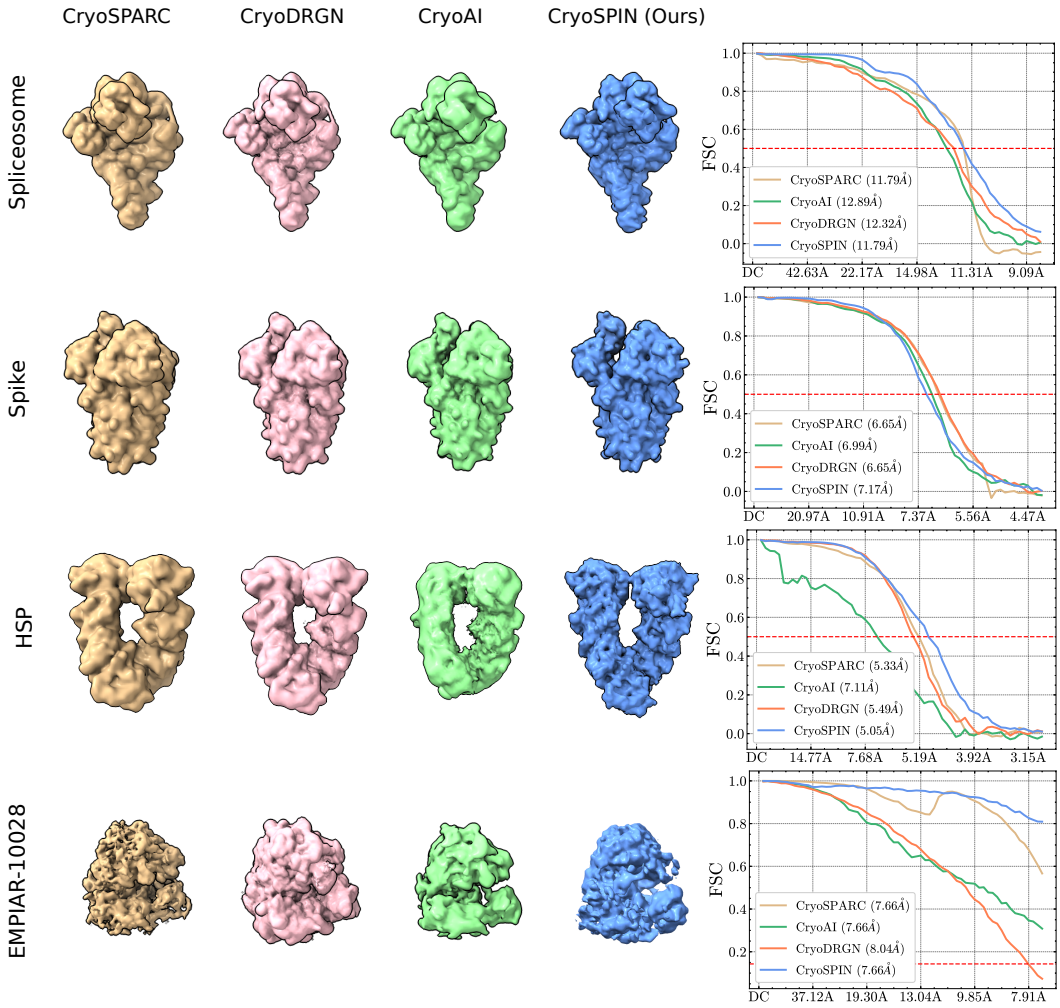


Figure 2: Qualitative and quantitative comparison of reconstructions obtained by our proposed semi-amortized method, cryoSPIN, with cryoAI [10], cryoDRGN [7] and cryoSPARC [8] We note that cryoAI often becomes stuck in local minima when particles are not centered, so for cryoAI we translate the input images to correct for the spatial offset. **(Left)** Final 3D reconstructions on three synthetic datasets and one experimental dataset (EMPIAR-10028) are depicted using ChimeraX [51]. **(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. CryoSPIN achieves higher resolution on the Spliceosome and HSP datasets, and it is competitive with the state of the art on the Spike and EMPIAR-10028 datasets.

methods, cryoAI and cryoDRGN, are inferior on the 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 visualize the Fourier Shell Correlation (FSC) [52] between the reconstruction and the ground truth (Fig. 2, right). FSC is the gold-standard metric measuring the normalised cross-correlation coefficient between two 3D Fourier volumes along shells of increasing radius. Higher FSC implies more accurate reconstruction. On the Spliceosome, HSP and 80S experimental datasets, our reconstruction outperforms amortized methods of cryoAI and cryoDRGN. Also, our method outperforms cryoSPARC on Spliceosome, HSP and 80S, while achieving competitive FSC on Spike. We use the standard 0.5 and 0.143 cutoff thresholds to report the resolution for synthetic and real data, respectively. CryoSPIN achieves higher or competitive resolution compared to others on all datasets.


https://doi.org/10.52202/079017-1773

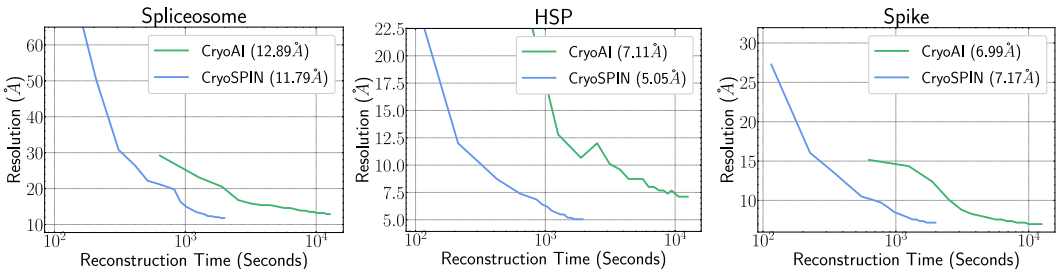


Figure 3: 3D resolution as a function of log time for different methods. These plots show the semi-amortized method is significantly faster than cryoAI.

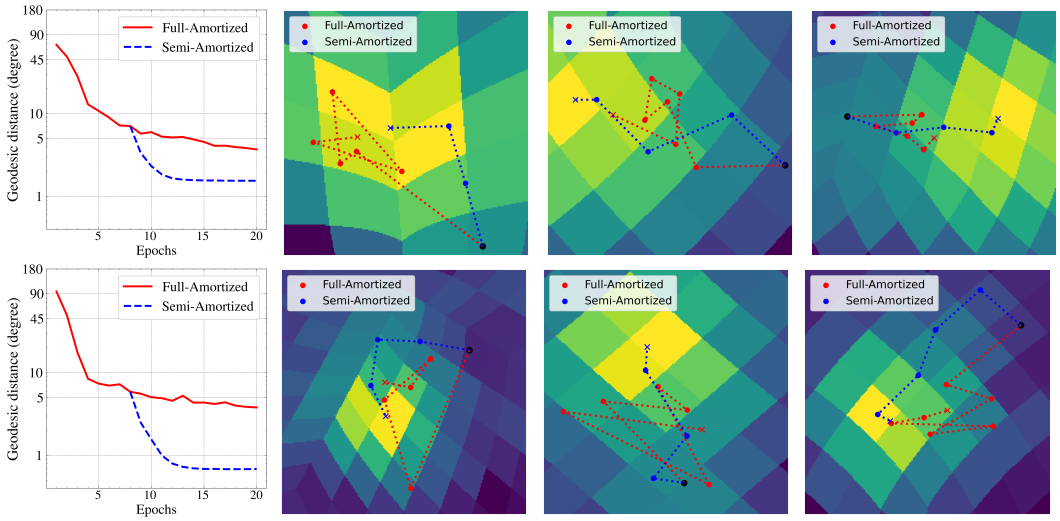


Figure 4: Quantitative and qualitative comparison of fully- vs. semi-amortized methods in pose optimization for Spike (top) and Spliceosome (bottom) datasets. **(Left)** Mean geodesic distance between the predicted pose and ground-truth is visualized at different epochs. By switching from amortized inference to direct optimization, semi-amortized method **(blue)** enjoys accelerated pose convergence compared to amortized inference **(red)**. **(Right)** To visualize pose inference, images depict the approximate log posterior for three particles (marginalized over in-plane rotations) as view-direction distribution over a HEALPix [53] uniform grid on a unit sphere $S^2$. After Gnomonic projection to 2D, we show the neighborhood of the mode of interest. **Black** dots depict starting points. **Blue** and **red** dots are pose estimates from fully- and semi-amortized methods.

In Fig. 3, we visualize resolution-time plots, showing that cryoSPIN gets to high-resolution reconstructions significantly faster than cryoAI. In fact, semi-amortization in cryoSPIN accelerates the improvement in the resolution and with our explicit structure decoder, we circumvent expensive MLP evaluations and train faster without any degradation to the final reconstruction quality as shown in Fig. 2. In Supp. B, through a detailed comparison with cryoAI, we show cryoSPIN is $\sim$ 6x faster and uses $\sim$ 5x less memory compared to cryoAI. Finally, we report the mean and median errors in estimated poses on synthetic datasets in Table 1, showing that our method outperforms cryoSPARC and cryoAI. High errors by cryoAI on HSP dataset shows that it gets stuck in local minima and fails to accurately estimate poses.

Table 1: Mean/median errors of estimated rotations in units of degrees evaluated on synthetic datasets.

| Method | HSP | Spike | Spliceosome |
|---|---|---|---|
| CryoSPARC [8] | 6.23 / 1.05 | 1.61 / 1.51 | 1.41 / 1.36 |
| CryoAI [10] | 45.83 / 61.86 | 2.52 / 2.29 | 2.85 / 2.61 |
| CryoSPIN (Ours) | **3.27 / 0.97** | **1.54 / 0.90** | **0.68 / 0.61** |

https://doi.org/10.52202/079017-1773

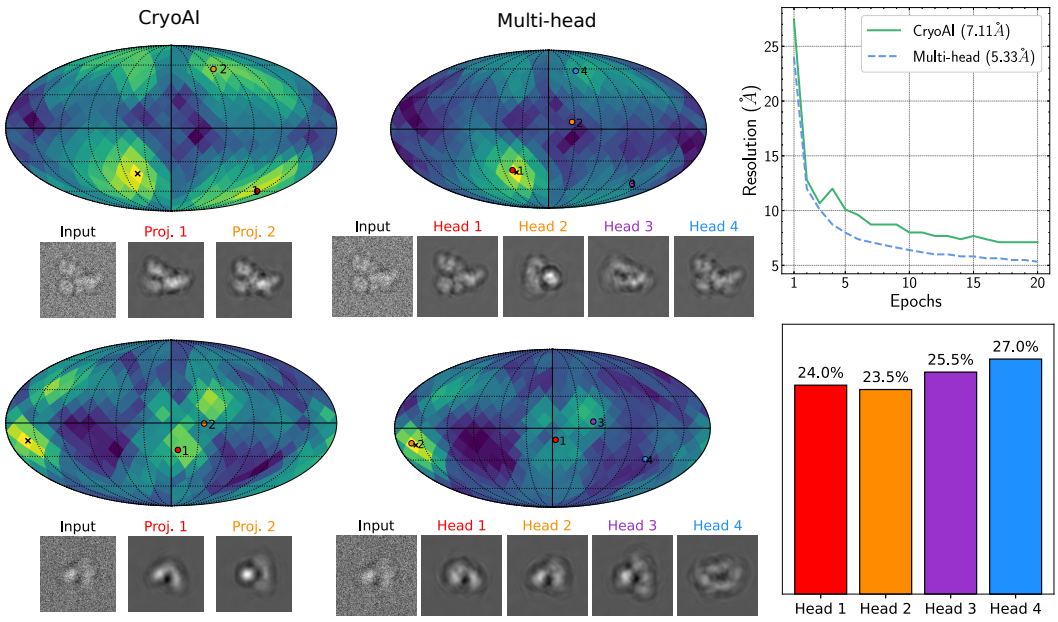


Figure 5: Comparing performance of our multi-head pose encoder ($M = 4$) with cryoAI pose encoder on the challenging HSP [45] dataset. **(Left)** The approximate log posterior of view direction is visualized on the unit sphere with highlighted areas showing modes of the distribution. CryoAI and multi-head encoders provide two and four pose estimates, respectively, which are marked with colored dots on the sphere (the order of poses is arbitrary). Below the sphere, the corresponding projections are illustrated. CryoAI fails to find the correct mode while our method is able cover multiple modes. **(Top, right)** With our multi-head encoder, the reconstruction converges to a much higher resolution compared to cryoAI. **(Bottom, right)** Percentage of images assigned to each head is visualized as a bar plot confirming that all heads participate in pose estimation.

### 5.2 Semi-Amortized vs. Fully-Amortized

To showcase the benefit of the auto-decoding stage, we compare our semi-amortized method with a fully-amortized baseline on spike and spliceosome datasets. Starting with auto-encoding, we branch the running reconstruction into two after 7 epochs: the first continues using the encoder while the second switches to direct pose optimization. We plot the average pose error with respect to the ground-truth at different epochs in Fig. 4 (left). When using the multi-head encoder, the pose with the least reconstruction error is selected among the candidates. Interestingly, 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.

Next, we analyze the evolution of pose estimates over the optimization landscape depicted as a heat map (Fig. 4, right). Qualitatively, poses obtained by direct optimization (blue dots) show stable convergence to the optimal point (highlighted area) while those inferred by the encoder (red dots) frequently oscillate. Indeed, the encoder in amortized inference is a globally parameterized function which might be too restrictive, yielding sub-optimal pose parameters. Therefore, poses inferred in an amortized fashion might fail to consistently converge to the optimal point. On other hand, direct optimization is intuitively more flexible as it is performed separately and locally for each image, exhibiting more stable convergence. See Supp. D for more examples.

### 5.3 Multi-Modal Pose Posterior

Lastly, we examine how well cryoSPIN multi-head encoder performs vs. cryoAI encoder in terms of handling the uncertainty in the pose on the challenging dataset of HSP [45]. To simplify the visualization, we run our method with $M = 4$ heads in this experiment. We first inspect the behavior of encoders for two example cryo-EM images (Fig. 5, left, see Supp. E for more examples). In each row, the approximate posterior distribution over view direction is visualized for both cryoAI and


https://doi.org/10.52202/079017-1773

multi-head encoders given the input image and reconstruction. Our multi-head encoder returns a set of plausible candidates, while cryoAI obtains two pose estimates by data augmentation. In both examples, the multi-head encoder identifies the correct mode while cryoAI selects the incorrect one. Note that the pose set predicted by the multi-head encoder contains other posterior modes as well. Intuitively, cryoAI encoder with two pose predictions cannot explore the pose space as much as our multi-head encoder. Thus, the pose posterior in cryoAI remains uncertain and multi-modal during reconstruction. This also contributes to slower convergence to higher-resolution reconstructions as depicted in the top left of Fig. 5. Finally, we visualize the percentage of images assigned to each head by our "winner-takes-all" loss. The relatively even spread of assignments to different heads shows that the multi-head encoder does not suffer hypotheses collapse [30], namely all heads actively participate in the pose prediction. In Supp. C, we investigate how each head specializes in pose prediction for non-overlapping regions of $SO(3)$.

## 6 Conclusion, Limitations and Future Work

In this paper, we introduce cryoSPIN, a new semi-amortized approach to ab initio cryo-EM reconstruction. We develop a new multi-head encoder that estimates a set of plausible candidate pose to handle the uncertainty. As the uncertainty is reduced, we transition to an auto-decoding stage where poses are iteratively optimized using SGD for each image. Our results show that our multi-head encoder is able to capture multiple modes of the pose distribution, and our flexible direct optimization enables accelerated convergence of poses and reconstructions. CryoSPIN outperforms cryoAI on experimental data and achieves competitive results with cryoSPARC.

**Limitations and future work.** In this work, we assume that the 3D structure is rigid while many biomolecules are flexible in practice, exhibiting different conformations. Our results show that, even in this relatively constrained homogeneous case, there is a significant gap between amortized inference and our semi-amortized approach. We expect that the benefits of semi-amortized inference will also apply to such heterogeneous cases. In particular, our proposed multi-head encoder should be applicable to the estimation of the multimodal posterior over heterogeneity latent parameters as well as pose. Yet, there are significant challenges in heterogeneous reconstruction, such as extra uncertainty due to heterogeneity and conflation of latent spaces, making it outside the scope of our analysis on semi-amortized inference.

The development of principled criteria for switching from amortized to auto-decoding is also an interesting direction for future research. Our experiments with a range of values show that after a sufficient number of epochs, the reconstruction error becomes insensitive to the switching time. Our intuition is that when switching too early, the pose posterior may have significant uncertainty, and the pose estimates may be too far from the correct basin of attraction to afford robust convergence. Therefore, it is better to switch late than early. However, a well-defined heuristic for the optimal switch time is lacking.

**Societal impact.** Structure determination with Cryo-EM for macromolecules is one of the most exciting, fundamental areas in structural biology, and is already having with vast social impact. Cryo-EM was used to determine the structure of the SARS-COV2 spike protein, the discovery of its pre-fusion conformation, and the evaluation of medical counter-measures. It compliments advances in methods like alpha-fold for structure prediction, transforming our understanding of the process of life at the scale of the cell, and the design and development of new therapeutics.

## Acknowledgments and Disclosure of Funding

This research was supported in part by the Province of Ontario, the Government of Canada, through NSERC, CIFAR, and the Canada First Research Excellence Fund for the Vision: Science to Applications (VISTA) programme, and by companies sponsoring the Vector Institute.

## References

[1] Werner Kühlbrandt. The resolution revolution. *Science*, 343(6178):1443–1444, 2014. 1

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

[3] 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. 1

[4] 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.

[5] 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. 2, 3, 4, 5, 6

[6] Ali Punjani and David J Fleet. 3DFlex: Determining structure and motion of flexible proteins from cryo-EM. *Nature Mthods*, 20(6):860–870, 2023. 1, 3

[7] 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. 2, 3, 4, 6, 7

[8] 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. 2, 3, 4, 6, 7, 8

[9] 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. 2, 3, 4

[10] 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. 2, 3, 4, 5, 6, 7, 8, 18

[11] 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. 2, 5

[12] 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. 2, 3, 5

[13] 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. 3, 5, 16

[14] 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. 2, 3, 5

[15] 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. 3, 4

[16] 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. 3

[17] 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.

[18] Gabi Pragier and Yoel Shkolnisky. A common lines approach for ab initio modeling of cyclically symmetric molecules. *Inverse Problems*, 35(12):124005, 2019. 3

 https://doi.org/10.52202/079017-1773

[19] 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. 3

[20] 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. 3

[21] Sjors HW Scheres. A Bayesian view on cryo-EM structure determination. *Journal of Molecular Biology*, 415(2):406–418, 2012. 3, 4

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

[23] Eugene L Lawler and David E Wood. Branch-and-bound methods: A survey. *Operations Research*, 14(4):699–719, 1966. 3

[24] 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. 3

[25] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[26] 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. 3

[27] 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. 3, 5, 16

[28] Kimin Lee, Changho Hwang, KyoungSoo Park, and Jinwoo Shin. Confident multiple choice learning. In *International Conference on Machine Learning*, pages 2014–2023. PMLR, 2017. 3

[29] 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.

[30] 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. 3, 10

[31] 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. 3

[32] 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. 3

[33] Ye Yuan and Kris Kitani. Diverse trajectory forecasting with determinantal point processes. *arXiv preprint arXiv:1907.04967*, 2019. 3

[34] 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. 3

[35] Earl J Kirkland. *Advanced computing in electron microscopy*, volume 12. Springer, 1998. 4

[36] Ronald N Bracewell. Strip integration in radio astronomy. *Australian Journal of Physics*, 9(2):198–217, 1956. 4

[37] Todd K Moon. The expectation-maximization algorithm. *IEEE Signal processing magazine*, 13(6):47–60, 1996. 4

[38] 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. 4, 5

[39] 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. 5

[40] 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. 4, 5

[41] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[42] 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. 5

[43] 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.

[44] Joe Marino, Yisong Yue, and Stephan Mandt. Iterative amortized inference. In *International Conference on Machine Learning*, pages 3403–3412. PMLR, 2018. 5

[45] D Goodsell. PDB-101 Molecule of the Month: Hsp90. 2008. 6, 9

[46] Clemens Plaschka, Pei-Chun Lin, and Kiyoshi Nagai. Structure of a pre-catalytic spliceosome. *Nature*, 546(7660):617–621, 2017. 6

[47] 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. 6

[48] 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. 6

[49] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6, 15

[50] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019. 6

[51] 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. 7

[52] Marin Van Heel and Michael Schatz. Fourier shell correlation threshold criteria. *Journal of Structural Biology*, 151(3):250–262, 2005. 7

[53] 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. 8, 16, 17

 https://doi.org/10.52202/079017-1773

[54] 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. 15

[55] Julian Panetta. Optimizing over SO(3). Technical report, University of California, Davis, 2018. 15

[56] F Sebastian Grassia. Practical parameterization of rotations using the exponential map. *Journal of Graphics Tools*, 3(3):29–48, 1998. 15

# Supplementary Material

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

## A 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. 3) with respect to the pose parameters. We define the new pose estimate based on the current one as follows:

$$R_{t+1} = R_\delta R_t \tag{9}$$

where $R_\delta$ is an infinitesimal rotation matrix perturbing the current estimate. The perturbation matrix $R_\delta$ 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 [54] in pytorch to compute the gradient with respect to $\boldsymbol{\omega}$ and make updates using Adam [49]. 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 $R_\delta$ with respect to the vector $\boldsymbol{\omega}$ is [55, 56],

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

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 [55], for $||\boldsymbol{\omega}|| \ll 1$, we substitute these terms with their numerically robust Taylor expansion, for instance,

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

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.

## B Ablation Study on Decoder

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


https://doi.org/10.52202/079017-1773

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 |

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 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.

## C 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 [53] 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. 6, 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 [27, 13], 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.

## D More on Semi-Amortized vs. Fully-Amortized

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. 7. 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.

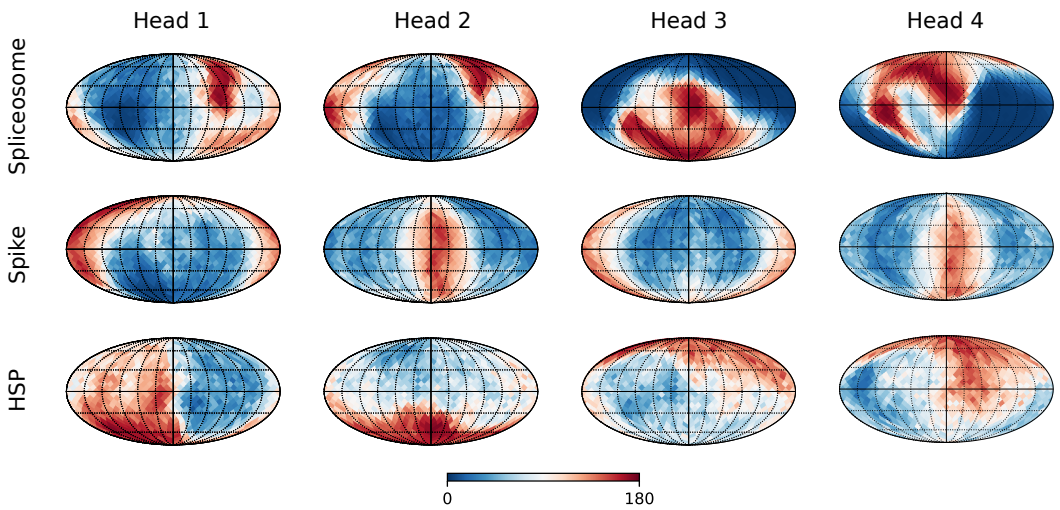


Figure 6: Average rotation error is visualized over the sphere for different heads of the multi-head pose encoder ($M = 4$). The sphere is uniformly divided into cells using HEALPix [53]; images are assigned to cells based on ground-truth view-direction. For each cell, the average rotation error is visualized, 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.

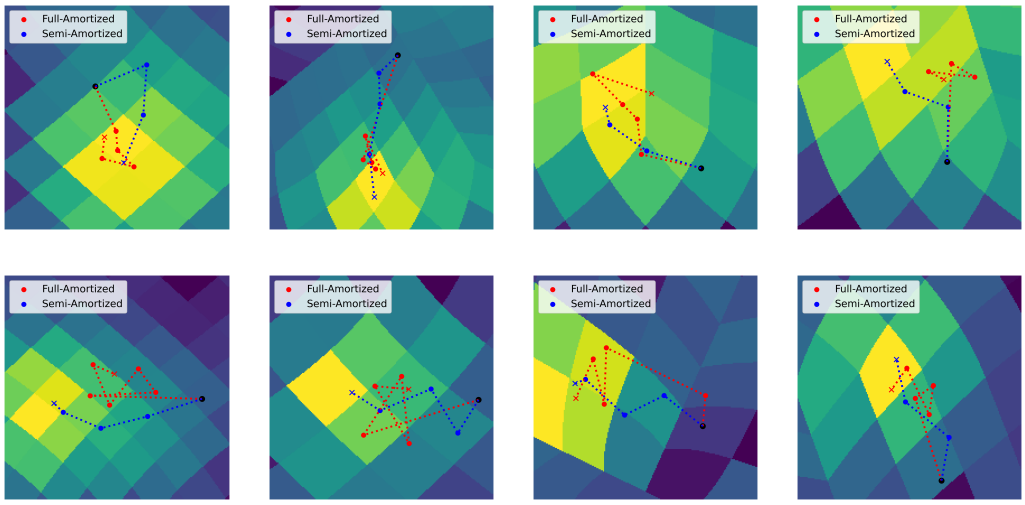


Figure 7: We compare the behavior of fully-amortized and semi-amortized pose inference on four examples per dataset. Two rows correspond to Spike and Spliceosome datasets, respectively. Each plot shows the approximate log pose posterior, marginalized over in-plane rotations represented as a heat map 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** cross is the starting point while **blue** dots and **red** dots show poses estimated by fully- and semi-amortized methods, respectively.

## E More on Multi-Modal Pose Posterior

Through more examples (Fig. 8), 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.


https://doi.org/10.52202/079017-1773

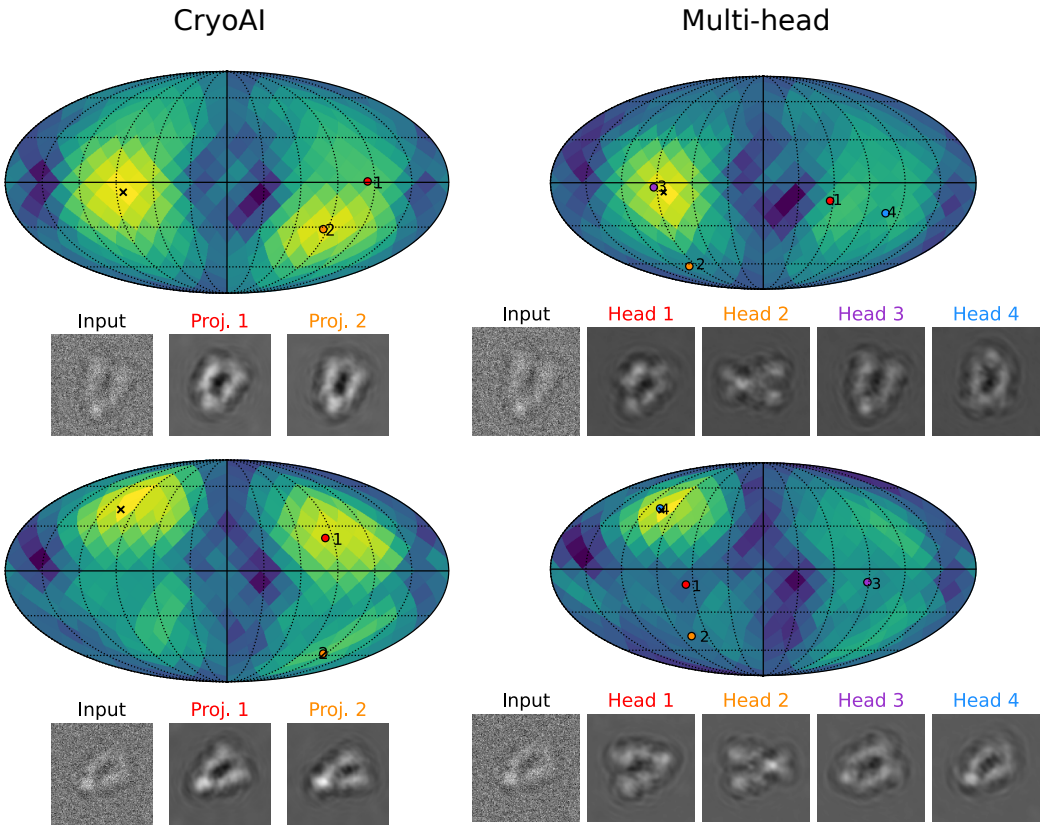


Figure 8: The approximate log posterior of view-direction visualized on the unit sphere with highlighted areas showing modes of the distribution. CryoAI [10] and our multi-head encoders provide two and four pose estimates, respectively, which are marked with colored dots on the sphere (the order of poses is arbitrary). The corresponding projections are also illustrated. CryoAI cannot identify the correct mode of pose distribution.

## NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

**The checklist answers are an integral part of your paper submission.** They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We clearly state the claims in the abstract, discuss them through the introduction section, and summarize them in a list at the end of introduction section.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]


https://doi.org/10.52202/079017-1773

Justification: We discuss limitations in the conclusion section. We acknowledge that ours assumptions about homogeneity and centered images could be limiting in some experimental setups. However, our framework could be extended to account for heterogeneity and translation parameters.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: We provide no theoretical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

Justification: In a subsection called implementation details, we included all details needed to reproduce reconstruction results of our method, cryoAI and cryoSPARC. Steps to generate synthetic datasets and preprocessing steps to prepare experimental data are also provided.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification:

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.


https://doi.org/10.52202/079017-1773

- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We have discussed all the necessary information about experimental settings in implementation details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: When reporting the resolution, we follow the standard to compute FSC by splitting data into two halves and make independent reconstructions and compare resulting half-maps.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We have mentioned that we are using A40 Nvidia GPU to prodcue the results. Also, in Fig. 3 we compare the execution time of our method vs. cryoAI.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: We conform with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Societal impacts of the work is dicussed in the conclusion section.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**


https://doi.org/10.52202/079017-1773

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: For cryoAI, we are using the available code provided in github. For cryoSPARC, we have mentioned which version of the software is used. Synthetic datasets are generated based on atomic models on PDB which are free of all copyright restrictions and freely available for both non-commercial and commercial use. For experimental data, we use EMPIAR database which is freely and publicly available to the global community under the CC0 licence.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: We don't release any new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.

- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No crowdsourcing or research with human subject.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No crowdsourcing or research with human subject.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.


https://doi.org/10.52202/079017-1773