CryoDRGN2: *Ab initio* neural reconstruction of 3D protein structures from real cryo-EM images

Ellen D. Zhong
MIT
zhonge@mit.edu

Adam Lerer
Facebook AI
alerer@fb.com

Joseph H. Davis
MIT
jhdavis@mit.edu

Bonnie Berger
MIT
bab@mit.edu

Abstract

Protein structure determination from cryo-EM data requires reconstructing a 3D volume (or distribution of volumes) from many noisy and randomly oriented 2D projection images. While the standard homogeneous reconstruction task aims to recover a single static structure, recently-proposed neural and non-neural methods can reconstruct distributions of structures, thereby enabling the study of protein complexes that possess intrinsic structural or conformational heterogeneity. These heterogeneous reconstruction methods, however, require fixed image poses, which are typically estimated from an upstream homogeneous reconstruction and are not guaranteed to be accurate under highly heterogeneous conditions.

In this work we describe cryoDRGN2, an ab initio reconstruction algorithm, which can jointly estimate image poses and learn a neural model of a distribution of 3D structures on real heterogeneous cryo-EM data. To achieve this, we adapt search algorithms from the traditional cryo-EM literature, and describe the optimizations and design choices required to make such a search procedure computationally tractable in the neural model setting. We show that cryoDRGN2 is robust to the high noise levels of real cryo-EM images, trains faster than earlier neural methods, and achieves state-of-the-art performance on real cryo-EM datasets.

1. Introduction

The last decade has seen explosive growth in the development and application of single particle cryo-electron microscopy (cryo-EM) for 3D structure determination of proteins and other biomolecules. Driven by parallel developments in improved hardware and image processing algorithms, many challenging structures not amenable to crystallographic approaches have now been solved at atomic or near-atomic resolution with cryo-EM [20, 30, 37].

Central to structure determination with cryo-EM is the computational reconstruction of the target molecule's 3D electron scattering potential (i.e. volume) from an experimentally-derived dataset of microscopy images. In a cryo-EM experiment, a purified solution of the molecule of interest is frozen in a thin layer of vitreous ice and imaged at cryogenic temperatures using a transmission electron microscope. After initial pre-processing of the raw micrographs, the resulting imaging dataset contains thousands to millions of noisy and randomly oriented 2D projection images (Fig. 1a). The goal of the cryo-EM reconstruction task is to infer the underlying 3D structure or structures present in the recorded images.

The 3D reconstruction task is a challenging inverse problem primarily due to the unknown image poses and the high amount of noise in the images. It is further complicated by the potential for each molecule to adopt variable conformations. A major opportunity thus exists to use cryo-EM to visualize and study complex distributions of dynamic protein structures, and numerous algorithms have been proposed to extract multiple structures from the imaging dataset, termed *heterogeneous reconstruction* [44].

Recent neural methods have shown promise in instantiating expressive continuous latent variable models for structural variability in cryo-EM data. In particular, cryoDRGN performs heterogeneous reconstruction by learning a deep generative model for 3D cryo-EM volumes. The first instantiation of cryoDRGN performed joint optimization of pose and heterogeneity with a branch and bound (BNB) algorithm for pose search (referred to as cryoDRGN-BNB here) [55]; however, this version scaled poorly and could not produce high-quality reconstructions of real cryo-EM datasets. In follow-up work, high-quality reconstructions on real data were achieved by modifying cryoDRGN to take previously estimated poses from a homogeneous reconstruction as input, hence eliding the difficult pose search procedure [54]. By estimating the intrinsic structural heterogeneity separately from the extrinsic pose variables, these methods are limited to mildly heterogeneous conditions where pose inference remains accurate.

In this work, we revisit the problem of joint optimization of image pose and volumes in cryoDRGN. In particular, we

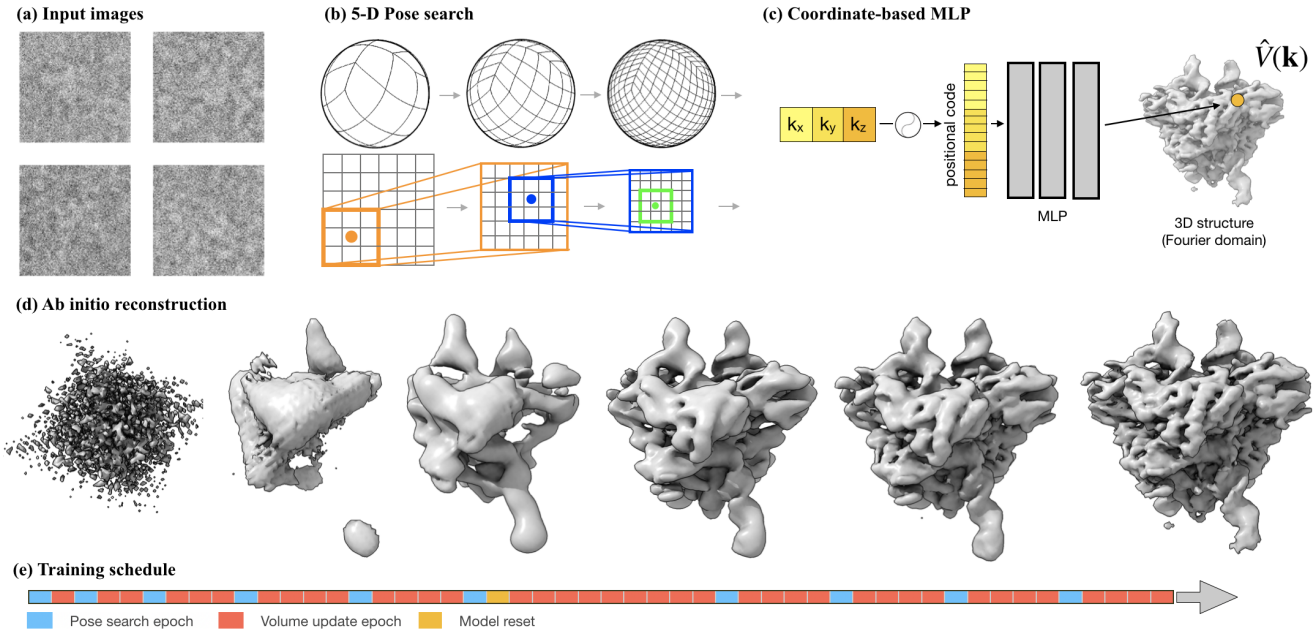

Figure 1: Overview of the cryoDRGN2 approach for *ab initio* reconstruction. (a) Example cryo-EM images of the RAG1-RAG2 complex [EMPIAR-10049]. (b) A multi-resolution 5-D pose search procedure doubles the resolution of the search grid at each iteration. (c) Coordinate-based MLP representation for volume (d) Volumes during *ab initio* training on the RAG dataset (e) A hypothetical training schedule interleaves pose search (expensive) and volume update (cheap) epochs. The model may also be reset after initial iterations to avoid vanishing gradients in training the neural volume.

consider 5-D camera pose optimization in the context of a feed-forward MLP representation of volume, and propose search techniques to address the high render time of MLPs relative to voxel-based representations. We further identify a pathological case of vanishing gradients during training that we hypothesize originates from distributional shifts of the objective function during joint optimization. With these techniques, we improve upon both the speed and accuracy of cryoDRGN-BNB and demonstrate for the first time that neural models can achieve state-of-the-art accuracy for fully unsupervised *ab initio* reconstruction on both homogeneous and heterogeneous real cryo-EM datasets.

2. Background and Related Work

The standard cryo-EM reconstruction task involves reconstructing a single volume $V : \mathbb{R}^3 \rightarrow \mathbb{R}$ from many noisy and randomly oriented 2D projection images of V. As cryo-EM images are orthographic integral projections of the volume, 2D images can be related to the 3D volume by the Fourier slice theorem [5], which states that the Fourier transform of a 2D projection is a central slice from the 3D Fourier transform of the volume. The generative process for image $\hat{X}$ in the Fourier domain is thus written:

$$\hat{X}(k_x, k_y) = \hat{g}S(t)\hat{V}(R^T(k_x, k_y, 0)^T) + \epsilon \qquad (1)$$

where $\hat{V} : \mathbb{R}^3 \rightarrow \mathbb{R}$ is the electron scattering potential (*volume*), $R \in SO(3)$, the 3D rotation group, is an unknown orientation of the volume, and $S(t)$ is a phase shift operator corresponding to in-plane translation in real space by $t \in \mathbb{R}^2$, which models imperfect centering of the volume within the image. The image signal is multiplied by $\hat{g}$, the contrast transfer function (CTF) for the microscope before being corrupted with frequency-dependent Gaussian noise and registered on a discrete grid of size $D \times D$, where D is the size of the image along one dimension.

Under this model, the probability of observing an image $\hat{X}$ with pose $\phi = (R, t)$ from volume $\hat{V}$ is:

$$p(\hat{X}|R, t, \hat{V}) = \frac{1}{Z} \exp\left(\sum_l \frac{-1}{2\sigma_l^2} \left|\hat{g}_l A_l(R)\hat{V} - S_l(t)\hat{X}_l\right|^2\right) \qquad (2)$$

where $A(R)\hat{V} = \hat{V}(R^T(\cdot, \cdot, 0^T)$ is a linear slice operator corresponding to rotation by R and linear projection along the z-axis in real space, l is a two-component index over Fourier coefficients for the image, σ_l is the width of the Gaussian noise expected at each frequency, and Z is a normalization constant. We refer the reader to [44] for a review of cryo-EM image formation and reconstruction methods.

Reconstruction algorithms are formulated as optimization of this statistical model, typically done in an iterative fashion

with expectation maximization (E-M) or gradient descent-based approaches [44]. In the E-M approach, starting from an initial model, images are aligned with the model (E-step). Then aligned images are "backprojected" to yield an updated estimate of V (M-step). Many software tools exist for 3D refinement[43, 47, 13, 35, 24]. Scheres [42] first proposed a Bayesian framework for *maximum a posteriori* estimation of $\hat{V}$, marginalizing over the posterior distribution of ϕ_i's.

While full marginalization can address uncertainty in pose variables, it is computationally demanding and many algorithms instead use a single maximum likelihood estimate for pose [13, 24, 47, 35]. In these iterative approaches, convergence of E-M to the correct structure depends strongly on the initialization, which is commonly obtained from other data sources, e.g. negative stain EM or approximations from previously-solved, related structures. In Brubaker et al. [6], stochastic gradient descent was proposed for data-driven, *ab initio* reconstruction of a low-resolution initial model, which was implemented in the cryoSPARC software package [35].

Heterogeneous reconstruction: Structural heterogeneity of the imaged protein complex is unknown *a priori* and can present in many forms (e.g. continuous motions vs. discrete compositional changes). Early approaches modeled structures as generated from a discrete mixture model of a small number of volumes [41, 40, 24], and the modelling of continuous motions was seen as a major challenge for the field. Advanced methods for heterogeneity analysis have since been proposed that learn continuous models of molecular variation [28, 33, 11, 27, 22, 21, 54, 7, 34, 56]. These methods all model structural heterogeneity with previously-posed images (e.g. from a homogeneous reconstruction), which limits the scope of heterogeneity analysis to structures where the consensus reconstruction is accurate.

Neural cryo-EM reconstruction: Until recently, all existing cryo-EM reconstruction methods used 3D voxel arrays to parameterize volume(s). Zhong *et al.* proposed cryoDRGN [55], a coordinate-based neural architecture to directly approximate the continuous 3D density function. Preliminary work describing cryoDRGN proposed the joint optimization of pose and heterogeneity with a branch and bound (BNB) algorithm for pose search (referred to as cryoDRGN-BNB in this work) [55]. Later work extended the cryoDRGN approach to real datasets by using image poses from a homogeneous consensus reconstruction [54], and has been successfully applied to identify novel structures [54, 14]. CryoGAN presented an alternate paradigm at the proof-of-concept stage for homogeneous reconstruction, which obviates the need for inference of image pose via distribution matching [15]. Recently, learning-based approaches for reconstruction attempt to infer pose by optimizing a parametric function to approximate the posterior over pose variables [38, 29]. These approaches have only been shown on synthetic datasets, and it remains to be seen how robust the optimization of this function is for real cryo-EM data.

Related work in computer vision: CryoDRGN models a continuous volume representation for protein structure that is related to the 3D representation used in other domains of computer vision [23, 31, 46, 45, 26]. Most similar is the neural radiance fields (NeRF) model used for novel view synthesis (NVS) of 3D natural scenes [26]. Unlike the natural image data used to train NeRF and related models however, the cryo-electron microscope produces noisy integral projections and thus the cryoDRGN coordinate-based neural model is specified in the Fourier domain with image generation modeled as central slices instead rendering with ray tracing.

In the standard setup of optimizing NeRF models for NVS [26], camera poses are treated as known. iNeRF inverts this process, and estimates camera poses via gradient descent to backpropagate the loss on a pretrained NeRF model directly into pose parameters[51]. NeRF– performs joint optimization of camera pose and the 3D scene/shape using gradient descent, updating poses from their initial, random values [49]. Here, we show that gradient descent on the cryo-EM reconstruction objective for image pose optimization fails. Instead, we propose an exhaustive pose search procedure performed concurrently with the optimization of the volume representation to achieve state-of-the-art performance on real cryo-EM datasets.

3. Method

In this section, we briefly overview the cryoDRGN architecture. Then, we describe the pose search algorithm in cryoDRGN2 and a series of strategies we use to speed up pose search. We then describe our overall training schedule, which alleviates a potential pathology when optimizing neural models under a nonstationary objective.

3.1. Overview of cryoDRGN

CryoDRGN parameterizes cryo-EM volumes using a coordinate-based MLP with parameters θ to directly approximate the continuous density function, $\hat{V}_\theta : \mathbb{R}^3 \rightarrow \mathbb{R}$ (Figure 1c). The model is specified in Hartley space [16] (which is closely related to Fourier space as the real minus imaginary Fourier components for real-valued signal). Thus, input Cartesian coordinates represent Hartley transform coefficients, and cryo-EM images (i.e. integral projections) are 2D central slices of the model whose orientation is determined from the image pose (Section 2). In heterogeneous reconstruction, the volume representation is augmented with a latent variable that is learned using amortized variational inference in the framework of variational autoencoders (VAEs) (Figure S1). Image poses, $\phi \in SO(3) \times \mathbb{R}^2$, are treated explicitly as geometric operations on a Cartesian coordinate lattice spanning $[-0.5, 0.5]^2$ that are input to the model. Training a cryoDRGN network involves optimizing neural

network weights θ, and image poses ϕ_i to maximize the likelihood of the experimental data under the image formation model (Equation 1). For more details, see [55].

3.2. 5-D pose search

In neural reconstruction, each model evaluation $\hat{V}_\theta(k)$ of coordinate k (which corresponds to a Fourier coefficient of an image's pixel) requires an expensive MLP evaluation. This is in contrast to voxel-based reconstruction, where image pixel values are computed by linear interpolation. In this work, we rethink the search procedure to minimize the number of neural network evaluations.

In cryoDRGN2, the pose ϕ_i for a given image $\hat{X}_i$ is estimated using a hierarchical search procedure over multi-resolution grids on the space of rotations and in-plane translations. We begin with an exhaustive search in the 5-D space of rotations and in-plane translations at some base resolution, γ_0, followed by an iterative refinement of the K most likely candidate poses by binary search at successively higher resolution grids, $\gamma_1 = \gamma_0/2, ..., \gamma_M$, (Fig. 1b). We also employ *frequency marching* [4], where we band-limit the signal to low frequency components of the image, and successively increase the frequency band-limit from k_{min} to k_{max} through the M iterations of pose refinement; this both decreases computational cost and prevents over-fitting on high-frequency noise while the grid is too coarse to align the high-frequency features. Finally, we note that the choice of the base grid resolution γ_0 has a significant impact on the accuracy of pose search, and that the base grid resolution used in state of the art tools was not computationally tractable in cryoDRGN-BNB. In the next sections, we discuss various speedups of the pose search procedure in cryoDRGN2 to enable fast and accurate pose search comparable with traditional state of the art tools.

3.2.1 Speeding up exhaustive search by interpolation

Consider the cost of the exhaustive search procedure. Using a base resolution of 15° and a translation base grid of 14×14 (our defaults) leads to 903,168 pose evaluations for a single image. Each pose evaluation consists of a squared error evaluation between the model $\hat{V}$ and the D^2 pixels of a central slice.

To minimize neural network evaluations, we combine the interpolation ideas from voxel-based reconstruction with our neural model. Instead of evaluating the MLP for each pixel in each pose, for the exhaustive search one can compute a 3D lattice within the frequency cutoff and compute pixel estimates by interpolation. In practice, we use interpolation only for the in-plane rotations, which reduces model evaluations by a factor of 24 (for the 15° resolution grid), taking the exhaustive search step off the critical path.

Interpolation is only accurate if the underlying function is smooth. Smoothness in Fourier space corresponds to the function being flat at large $\mathbf{r}$ in real space, which is satisfied as long as the model output is centered and smaller than the box size. Importantly, it's *not* satisfied for the images, which have a high degree of noise throughout the image; therefore, we found it crucial to interpolate the model output rather than the image.

3.2.2 Leveraging a cheap translation operator

Search over in-plane translations does not require extra model evaluations because translations in real space map to multiplication by an exponential function in Fourier space, which can be computed exactly without additional model evaluations.

For efficiency, we apply translations to the (single) image rather than the 4,608 model estimates at different poses. Computing the optimal pose now consists of finding the minimum mean squared error (MSE between approximately 10^5 model estimates with about 10^2 translated images. Taking advantage of the identity $(A-B)^2 = |A|^2 + |B|^2 - 2A \cdot B$, all the MSEs can be computed as a single matrix multiply between the rotated estimates and the translated images (plus some norms), which is both memory-efficient and very fast on modern CPU and GPU architectures.

The fact that evaluating translations is essentially free leads us to a new approach for pose refinement, which effectively factorizes the search over $SO(3) \times \mathbb{R}^2$ to independent searches over $SO(3)$ and $\mathbb{R}^2$ under some (standard) assumptions. In earlier work, the top $K \leq 24$ most likely candidate poses were selected for refinement; a grid of 2^{3+2} new poses was evaluated for each candidate [55]. In practice, these candidates often corresponded to multiple translations of the *same rotation*, while other promising rotations were discarded. In this work, we instead pick the $K \approx 8$ most likely candidate *rotations* and the single most-likely translation for each of these rotations, t^*; at the next resolution of pose refinement, we search a grid of 2^3 new rotations at the higher resolution but check a large grid of candidate translation grid points centered around t^* at 2x the resolution with .5x the translational grid extent (see Fig. 1(b))[1]. This allows us to pursue a larger number of candidate poses, and makes the algorithm less sensitive to the choice of translation resolution.

3.3. Training Schedule and Model Re-initialization

Traditional cryo-EM reconstruction consists of an expectation-maximization procedure of alternating pose in-

[1] Since we only check a local grid of translations around the minimum for each rotation candidate, a key assumption for this approach is that the loss surface with respect to translation is unimodal given the rotation. This is satisfied for biological datasets with respect to translation. We note that it's not satisfied for rotations, e.g. molecular complexes with symmetries will have a local minimum at each symmetry operator offset from the global minimum, so it is crucial that multiple candidate rotations be refined.

ference (E-step) and volume estimation (M-step). In neural reconstruction, the volume estimation consists of gradient descent on a reconstruction loss for the maximum-likelihood poses [55]. However, we observe that the representation quality of the coordinate-based MLP is limited by the number of gradient updates, and the computational cost of each update is dominated by the pose search procedure (about 10x slower with pose search than without). Thus, in cryoDRGN2 we increase the number of gradient updates by reusing each computed pose for N gradient updates. Specifically, we alternate training epochs that perform pose search with epochs that reuse the latest computed poses (Fig. 1e). For simplicity we set a constant pose search frequency (e.g. N=5), however, further speedups can likely be achieved with different (e.g. exponential) training schedules.

3.3.1 Vanishing gradients

We observed a pathology in neural network training when performing *ab initio* reconstruction on a particularly challenging dataset. Due to the alternating updates of pose and volume, the neural network training objective changes during the course of training: in early epochs, the pose estimates are less accurate leading to an inability to resolve features for large $|k|$; as a result, SGD minimizes the L2 loss by predicting a constant function at these high frequencies (Fig. 4a). Later in training when high-frequency features can be resolved, the gradient of these high-frequency predictions is 0 with respect to the input k and model parameters, leading to an inability to update the volume approximation given the new poses (Fig. 4c). We validated that this is a vanishing gradients issues rather than e.g. a local minimum of the loss, by explicitly computing the gradient $d\hat{V}_k/dw_j$ and $d\hat{V}_k/dk$ for the coordinate k highlighted in Figure 4c) and observing that they are zero.

Training pathologies due to sparse or vanishing gradients to the parameters is well documented and a variety of solutions have been proposed [8, 9]. However, these analyses typically focus on supervised learning, whereas we conjecture that it is precisely the non-stationarity of the objective that leads to this pathology. We found that proposed solutions like a Leaky ReLU activation [25] or residual corrections [18] did not fully solve the problem.

We found that resetting the coordinate MLP model and optimizer state intermittently during training (while retaining the image poses inferred from the old model) resolved the vanishing gradient problem, as shown in Fig. 4. The training schedule including model reset is illustrated in Fig. 1(e). We leave a further analysis of this vanishing gradient problem as well as alternative methods for warm-starting training from an old model [3] to future work.

Method	**Grid setting**	**Time**	**Accuracy**
cDRGN-BNB	$30°, 2.8$ pix	0:01:32	0.691
cryoDRGN2	$30°, 2.8$ pix	0:00:23	0.643
cryoDRGN2	$15°, 1.4$ **pix**	**0:00:52**	**0.004**

Table 1: Comparison of pose search algorithm and hyperparameter choices. Timing and accuracy (mean rotation error) are measured for the alignment of 1000 images from the 80S dataset on a pre-trained cryoDRGN model.

3.4. Hyperparameters

A listing of the cryoDRGN2 pose search algorithm and its hyperparameters is given in Appendix A. To choose reasonable defaults for the base resolution γ_0, number of grid subdivision M, kept poses per subdivision K, and frequency marching bounds k_{min} and k_{max}, we perform a hyperparameter sweep of possible values, evaluated by aligning a subset of images to a pre-trained cryoDRGN model (Appendix A).

As training speed depends on many external factors, we do not perform an ablation of each of these techniques to assess computational speedups. Instead, we verify that our overall pose search algorithm is accurate and compare the overall training time of our reconstruction method relative to prior work (Table 1).

We note that existing traditional reconstruction methods achieved high pose accuracy with a $\gamma_0 = 15°$ or $7.5°$ on $SO(3)$ (which corresponds to 4,608 or 36,864 rotations, respectively)[57], but cryoDRGN-BNB [55] was restricted to $\gamma_0 = 30°$ (576 rotations) due to computational limitations. Depending on the smoothness of the underlying objective, using too coarse of a search resolution can lead to missing the global minimum. With the techniques described above, we are able to use a base resolution of $15°$ or even $7.5°$, leading to much higher pose accuracy (Table 1).

4. Results

We qualitatively and quantitatively evaluate cryoDRGN2 for *ab initio* reconstruction in both homogeneous and heterogeneous settings. We first validate our pose search algorithm on synthetic homogeneous datasets (*hand, spike*) and compare to baseline methods. Next, we perform homogeneous reconstruction on three real cryo-EM datasets of variable difficulty (*80S, RAG12, spliceosome*). We highlight a particularly challenging test case of the RAG1-RAG2 complex. Lastly, we show heterogeneous reconstruction on synthetic and real heterogeneous cryo-EM data (*Linear1d, spliceosome*).

4.1. Homogeneous reconstruction of synthetic datasets

Data and setup: We create two synthetic *homogeneous* datasets from a ground truth volume of a hand and of the

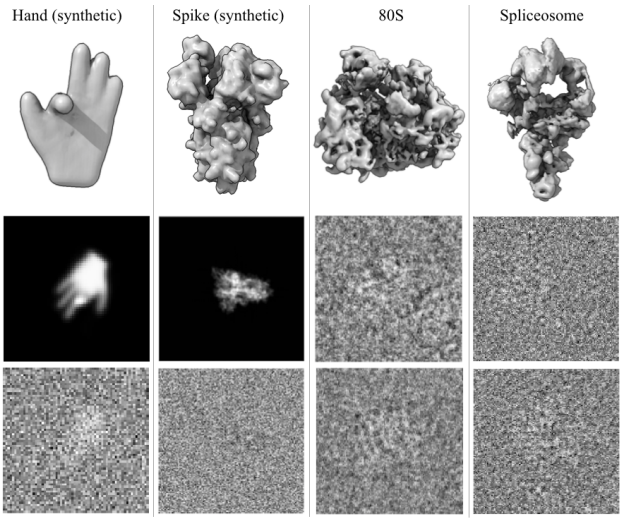

Figure 2: Ground truth (synthetic) or reference (real) volumes with corresponding example cryo-EM images below. Synthetic datasets show a noiseless and a corresponding noisy image (SNR=0.1).

SARS-CoV-2 spike protein (PDB: 6VYB) [2] by following the standard image formation model (50k images, D=64/128 for hand/Spike, Appendix B for more details). We test on both the noiseless version and a noisy (SNR=0.1), realistic version of the dataset. We compare cryoDRGN2 against two methods that use a branch and bound algorithm for pose search: prior work cryoDRGN-BNB [55] and cryoSPARC [35], a state-of-the-art, traditional (voxel-array-based) software for cryo-EM reconstruction. Results with cryoSPARC are obtained from *ab initio* reconstruction followed by homogeneous refinement in cryoSPARC v2.15. We additionally compare the performance of cryoDRGN2 to two other paradigms for pose estimation: a learning-based method for pose inference (pose-VAE) where we use a variational encoder to predict 3D pose variables and the direct gradient-based optimization of pose variables (pose-GD). For pose-GD, we randomly initialize 3D pose variables, and initialize the volume from a pre-trained model. Additional experimental details are in Appendix C.

We find that cryoDRGN2 obtains high accuracy on our synthetic datasets similar to other pose search algorithms (cryoDRGN-BNB and cryoSPARC). Similar to [48] we find that the gradient-based approaches perform poorly, likely due to the non-convexity of the objective with respect to pose. Pose errors to ground truth poses are given in Table 2. Visualizations of the reconstructed volumes of the Hand are given in Figure S2.

4.2. Homogeneous reconstruction of real datasets

Data and setup: We use three experimental cryo-EM datasets publicly available on the EMPIAR database: the 80S ribosome (EMPIAR-10028) [50], the RAG1-RAG2 complex

Method	**Hand**		**Spike**	
	Noiseless	*SNR=0.1*	*Noiseless*	*SNR=0.1*
Pose VAE	6.66	6.64	6.67	6.65
Pose GD	6.61	6.65	6.63	6.66
cryoSPARC	0.0015	0.071	0.0003	**0.002**
cDRGN-BNB	0.007	0.25	0.0006	0.012
cryoDRGN2	**0.0003**	**0.027**	**0.0001**	0.011

Table 2: Homogeneous reconstruction pose accuracy on synthetic datasets quantified by median rotation error to the ground truth image poses. Mean error statistics are provided in the Appendix. Rotation error is defined as $||R - \hat{R}||_F^2$ between the predicted and ground truth rotation.

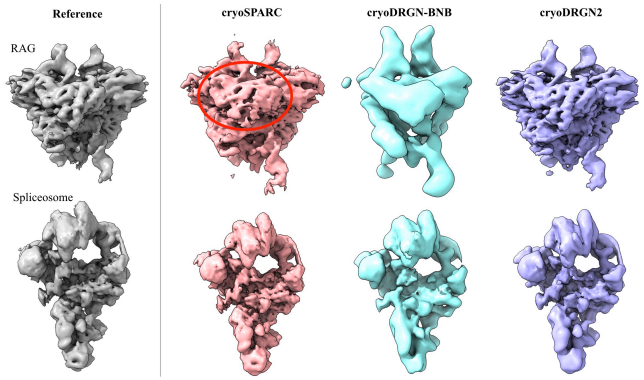

Figure 3: Reconstructed volumes from different homogeneous *ab initio* reconstruction algorithms and the reference volume.

(EMPIAR-10049) [1], and the pre-catalytic spliceosome (EMPIAR-10180) [32, 28]. Images were downsampled to D=128 for all experiments. Real datasets have varying degree of difficulty due to differences in contrast (*i.e.* signal) for different molecules, non-uniformity of pose distributions, and varying degrees of underlying structural heterogeneity and symmetry. As real datasets lack ground truth, to produce a reference model for comparison, we train a cryoDRGN coordinate-based MLP using published poses [54]. We note that the published structures were originally obtained using prior knowledge from other related complexes (as their initial models for refinement), and in the case of the spliceosome, also involved many rounds of hierarchical processing due to the heterogeneity of the complex. We baseline against cryoDRGN-BNB and cryoSPARC *ab initio* reconstruction followed by homogeneous refinement (Additional details in Appendix B).

On real cryo-EM datasets, cryoDRGN2 is able to obtain high quality structures *ab initio*, matching that of the reference refinement and competitive with existing *ab initio* methods. We report the difference in the estimated poses to the reference poses in Table 3. Visualizations of the reconstructed volumes of the RAG and spliceosome datasets are

Method	80S		RAG12		Spliceosome	
	Mean	*Median*	*Mean*	*Median*	*Mean*	*Median*
cryoSPARC	**0.0186**	**0.0001**	3.7806	0.3084	**0.0853**	**0.0015**
cryoDRGN-BNB	0.6151	0.0020	4.1621	4.6371	2.2187	0.1854
cryoDRGN2	0.0578	0.0008	3.4254	0.0386	0.1958	0.0046
cryoDRGN2+r	0.0590	0.0008	**3.3730**	**0.0226**	0.1947	0.0044

Table 3: Homogeneous reconstruction pose accuracy on real cryo-EM datasets quantified by mean/median rotation error to the reference. Rotation error between the predicted and reference pose is defined as $||R - \hat{R}||_F^2$ after global alignment of the set of images. Translation error statistics are provided in the Appendix.

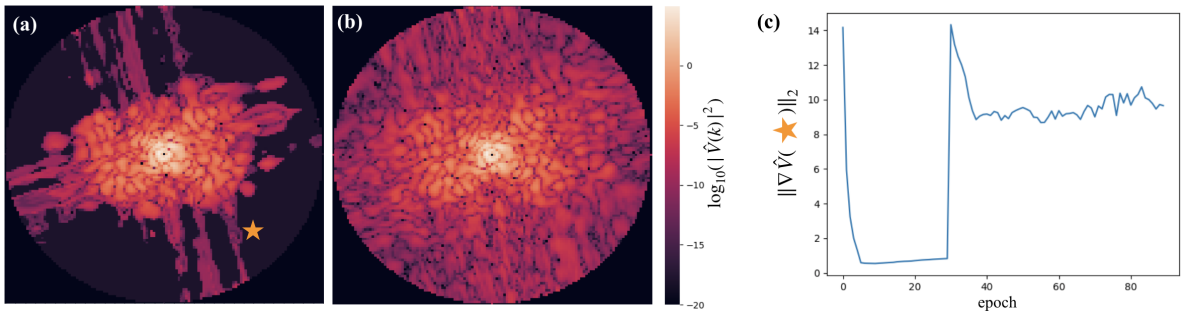

Figure 4: Power spectral density $|\hat{V}(k)|^2$ of a slice from the neural volume at different stages of training. **(a)** Model slice after 30 epochs of joint optimization of pose and $\hat{V}$. **(b)** Model slice after the model was reinitialized and trained for 30 epochs using fixed poses from **(a)**. **(c)** L2 norm of the gradient of a dummy loss computed at the starred coordinate in **(a)** with respect to last layer of weights of $\hat{V}$.

given in Figure 3. We additionally quantify that the reconstructed volumes match the ground truth using Fourier Shell Correlation (FSC) curves (Figures S4,S3 and Tables S5,S6).

80S: The 80S ribosome dataset is a common cryo-EM benchmark dataset with high contrast images and static structure, and all methods perform well with low pose error (Table 3) and good FSC metrics (Tables S5,S6, Figures S4,S3).

RAG: The RAG complex is a much more challenging dataset, e.g. a replicate of cryoSPARC refinement using the same initial model of the published structure produces a 0.91/0.03 mean/median rotation error. The discrepancy between the mean and median statistic in Table 3 is likely from the approximate 2-fold symmetry of the core of the complex. In the qualitative comparison of the reconstructed volumes, we observe that high resolution features are more resolved in the cryoDRGN2 volume than in cryoSPARC (Figure 3). CryoDRGN-BNB only produces an approximately correct low-resolution shape (Fig. 3). We also observe improvements of the cryoDRGN2 volume relative to the *reference volume* in the (non-symmetric) DNA extensions of the complex, likely due to alignment of images to the correct symmetry copy by cryoDRGN2 (Figure S5).

Spliceosome: CryoDRGN2 produces a volume closely matching the reference with low pose error (Fig. 3). Relative to cryoDRGN2, cryoDRGN-BNB has higher pose error and captures an approximately correct, though much lower resolution shape. Initial results with cryoSPARC were poor (e.g. the median image alignment error was 5.8 for rotations), however once images were recentered based on published poses, cryoSPARC was able to produce a high quality consensus reconstruction (Figures 3,S4,S3).

4.2.1 Importance of model reset

We identified a pathological case of vanishing gradients which we hypothesize results from distributional shifts in pose variables when training on the RAG complex dataset (discussed in Section 3.3). We note that the RAG dataset pose distribution is highly skewed towards a preferred orientation. We observe that while the initial round of cryoDRGN2 training produced a low resolution structure matching the reference, the model output at high resolution (large $|k|$) was essentially zero (visualization in Fig. 4). We compute the norm of the gradient of the last layer during the stages of training (Fig. 4), showing disappearance of the gradient in the initial stage and the recovery of gradient information after model reset. Further refinement of the model with pose search (cryoDRGN2+r), is able to learn high resolution features with concomitant improvement in pose accuracy (Fig. 4 right, Fig. 1d). This observation motivates our multi-stage training procedure (cryoDRGN2+r), and may be relevant in other application domains of neural volume rendering.

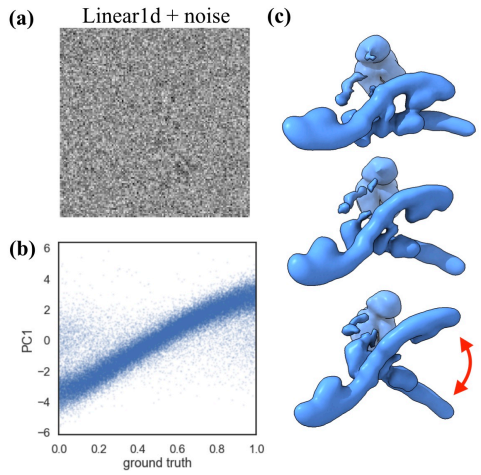

Figure 5: (a) Example images of the Linear1d dataset. (b) CryoDRGN2 latent embeddings of particle images. (c) CryoDRGN2 reconstructed structures along the PC1 axis of the latent embeddings.

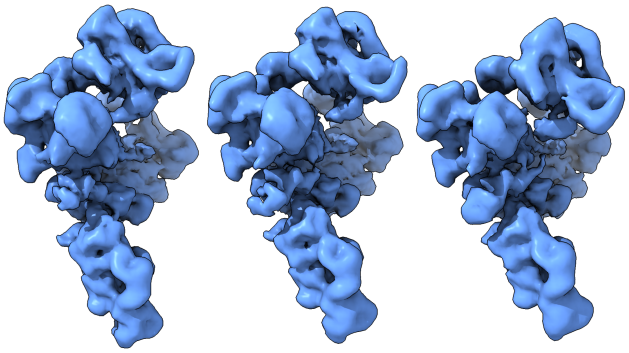

Figure 6: CryoDRGN2 reconstructed volumes of the spliceosome generated along the PC1 axis of the latent embeddings.

4.3. Heterogeneous reconstruction

Data and setup: We perform *ab initio* heterogeneous reconstruction (i.e. joint inference of $\hat{V}_z$ and ϕ_is) on two datasets that contain large structural variations. The Linear1d dataset is a synthetic dataset containing a large continuous 1D motion [55]. We generate a dataset containing 50k images with CTF and noise (SNR=0.1, D=128) from 50 ground truth models simulating a continuous motion (Fig. 5a). We also test cryoDRGN2 on the pre-catalytic spliceosome dataset (EMPIAR-10180), which contains large continuous motions [28, 54, 33]. We compare against cryoDRGN-BNB as all other methods for reconstructing continuous heterogeneity require previously assigned poses.

We find that CryoDRGN2 is able to reconstruct the underlying continuous 1D motion of the synthetic dataset (Fig. 5, Fig. S6). Trained on the spliceosome dataset, cryoDRGN2 volumes sampled along the PC1 axis of the latent embeddings show large-scale flexing of the molecular complex (Fig. 6), similar to previous analyses with pose supervision [28, 54, 33]. CryoDRGN-BNB captures the same qualitative motions in these datasets however volumes are lower resolution (Figure S7). Visualizing the inferred pose distribution highlights local minima in the cryoDRGN-BNB pose search (Fig. 7).

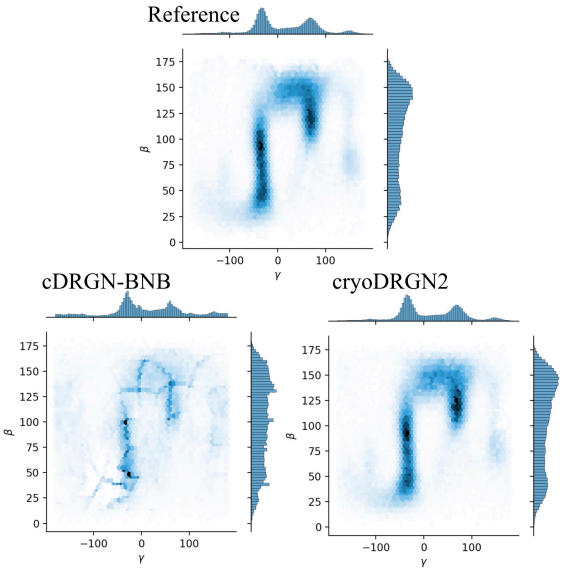

Figure 7: Pose distribution of the spliceosome dataset inferred from *ab initio* heterogeneous reconstruction with cryoDRGN2 and cryoDRGN-BNB.

5. Conclusion

We present cryoDRGN2, a method for reconstructing single or heterogeneous distributions of protein structure from unlabelled 2D cryo-EM images. By addressing inaccuracies and computational bottlenecks in earlier unsupervised optimization of cryoDRGN models, we demonstrate that neural models can achieve state-of-the-art accuracy for *ab initio* reconstruction of challenging, real cryo-EM datasets. Although we reanalyze publicly available datasets here, we are optimistic that this and future improvements will be fruitful for structure determination of novel datasets, especially for structurally heterogeneous complexes, for which no other reconstruction algorithms exist. The techniques shown here may be useful in other domains in computer vision, including graphics, inverse rendering, and robotics.

6. Acknowledgements

We thank the MIT-IBM Satori team for computing resources and support. This work was funded by the NSF GRFP to E.D.Z., NIH grant R01-GM081871 to B.B., NIH grant R00-AG050749 to J.H.D., and a grant from the MIT J-Clinic for Machine Learning and Health to J.H.D. and B.B.

References

[1] Molecular Mechanism of V(D)J Recombination from Synaptic RAG1-RAG2 Complex Structures. *Cell*, 163(5):1138–1152, nov 2015.

[2] Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell*, 181(2):281–292.e6, 2020.

[3] Jordan T Ash and Ryan P Adams. On warm-starting neural network training. *arXiv preprint arXiv:1910.08475*, 2019.

[4] Alex Barnett, Leslie Greengard, Andras Pataki, and Marina Spivak. Rapid solution of the cryo-EM reconstruction problem by frequency marching. *arXiv.org*, Oct. 2016.

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

[6] Marcus A. Brubaker, Ali Punjani, and David J. Fleet. Building Proteins in a Day: Efficient 3D Molecular Reconstruction. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June-2015:3099–3108, apr 2015.

[7] Muyuan Chen, Steven Ludtke, and Verna Marrs. Deep learning based mixed-dimensional GMM for characterizing variability in CryoEM. *arXiv*, 2021.

[8] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.

[9] Yann Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. *arXiv preprint arXiv:1406.2572*, 2014.

[10] Luca Falorsi, Pim de Haan, Tim R Davidson, Nicola De Cao, Maurice Weiler, Patrick Forré, and Taco S Cohen. Explorations in Homeomorphic Variational Auto-Encoding. *arXiv.org*, July 2018.

[11] Joachim Frank and Abbas Ourmazd. Continuous changes in structure mapped by manifold embedding of single-particle data in cryo-EM. *Methods (San Diego, Calif.)*, 100:61–67, May 2016.

[12] 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.

[13] Timothy Grant, Alexis Rohou, and Nikolaus Grigorieff. cisTEM, user-friendly software for single-particle image processing. *eLife*, 7:e14874, mar 2018.

[14] Miao Gui, Meisheng Ma, Erica Sze-Tu, Xiangli Wang, Fujiet Koh, Ellen D Zhong, Bonnie Berger, Joseph H Davis, Susan K Dutcher, Rui Zhang, et al. Structures of radial spokes and associated complexes important for ciliary motility. *Nature structural & molecular biology*, 2020.

[15] Harshit Gupta, Michael T. McCann, Laurène Donati, and Michael Unser. CryoGAN: A new reconstruction paradigm for single-particle cryo-EM via deep adversarial learning. *bioRxiv*, 2020.

[16] Ralph VL Hartley. A more symmetrical fourier analysis applied to transmission problems. *Proceedings of the IRE*, 30(3):144–150, 1942.

[17] David Haselbach, Ilya Komarov, Dmitry E Agafonov, Klaus Hartmuth, Benjamin Graf, Olexandr Dybkov, Henning Urlaub, Berthold Kastner, Reinhard Lührmann, and Holger Stark. Structure and Conformational Dynamics of the Human Spliceosomal Bact Complex. *Cell*, 172(3):454–464.e11, Jan. 2018.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

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

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

[21] 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.org*, July 2019.

[22] Roy R Lederman and Amit Singer. Continuously heterogeneous hyper-objects in cryo-EM and 3-D movies of many temporal dimensions. *arXiv.org*, Apr. 2017.

[23] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural Volumes: Learning Dynamic Renderable Volumes from Images. *ACM Transactions on Graphics*, 38(4):14, jun 2019.

[24] Lyumkis, Dmitry, Brilot, Axel F, Theobald, Douglas L, and Grigorieff, Nikolaus. Likelihood-based classification of cryo-EM images using FREALIGN. *Journal of structural biology*, 183(3):377–388, Sept. 2013.

[25] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3. Citeseer, 2013.

[26] 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. *ECCV*, 2020.

[27] Amit Moscovich, Amit Halevi, Joakim Andén, and Amit Singer. Cryo-EM reconstruction of continuous heterogeneity by Laplacian spectral volumes. *arXiv.org*, July 2019.

[28] Takanori Nakane, Dari Kimanius, Erik Lindahl, and Sjors Hw Scheres. Characterisation of molecular motions in cryo-EM single-particle data by multi-body refinement in RELION. *eLife*, 7:e36861, June 2018.

[29] Youssef SG Nashed, Frederic Poitevin, Harshit Gupta, Geoffrey Woollard, Michael Kagan, Chuck Yoon, and Daniel Ratner. End-to-end simultaneous learning of single-particle orientation and 3d map reconstruction from cryo-electron microscopy data. *arXiv preprint arXiv:2107.02958*, 2021.

[30] Eva Nogales. The development of cryo-EM into a mainstream structural biology technique. *Nature methods*, 13(1):24–27, Jan. 2016.

[31] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. *CVPR*, 2019.

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

[33] Ali Punjani and David J Fleet. 3D Variability Analysis: Directly resolving continuous flexibility and discrete heterogeneity from single particle cryo-EM images. *bioRxiv*, 11(2):2020.04.08.032466, apr 2020.

[34] Ali Punjani and David J Fleet. 3d flexible refinement: Structure and motion of flexible proteins from cryo-em. *bioRxiv*, 2021.

[35] 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, Mar. 2017.

[36] Ali Punjani, Haowei Zhang, and David J Fleet. Non-uniform refinement: adaptive regularization improves single-particle cryo-em reconstruction. *Nature methods*, 17(12):1214–1221, 2020.

[37] Jean-Paul Renaud, Ashwin Chari, Claudio Ciferri, Wen-Ti Liu, Hervé-William Rémigy, Holger Stark, and Christian Wiesmann. Cryo-EM in drug discovery: achievements, limitations and prospects. *Nature reviews. Drug discovery*, 17(7):471–492, July 2018.

[38] 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.

[39] Peter B Rosenthal and Richard Henderson. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *Journal of molecular biology*, 333(4):721–745, 2003.

[40] Sjors HW Scheres. Maximum-likelihood methods in cryo-em. part ii: Application to experimental data. *Methods in enzymology*, 482:295, 2010.

[41] Sjors HW Scheres, Mikel Valle, Rafael Nuñez, Carlos OS Sorzano, Roberto Marabini, Gabor T Herman, and Jose-Maria Carazo. Maximum-likelihood multi-reference refinement for electron microscopy images. *Journal of molecular biology*, 348(1):139–149, 2005.

[42] Sjors H W Scheres. A Bayesian view on cryo-EM structure determination. *Journal of molecular biology*, 415(2):406–418, Jan. 2012.

[43] Sjors H W Scheres. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *Journal of structural biology*, 180(3):519–530, Dec. 2012.

[44] Amit Singer and Fred J. Sigworth. Computational Methods for Single-Particle Electron Cryomicroscopy. *Annual Review of Biomedical Data Science*, 3(1):163–190, jul 2020.

[45] Vincent Sitzmann, Julien N.P. Martel, Alexander W. Bergman, David B. Lindell, Gordon Wetzstein, and Stanford University. Implicit Neural Representations with Periodic Activation Functions. *NeurIPS*, 2020.

[46] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhöfer. DeepVoxels: Learning Persistent 3D Feature Embeddings. *arXiv.org*, Dec. 2018.

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

[48] Karen Ullrich, Rianne van den Berg, Marcus Brubaker, David Fleet, and Max Welling. Differentiable probabilistic models of scientific imaging with the Fourier slice theorem. *arXiv.org*, June 2019.

[49] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. NeRF- - Neural Radiance Fields Without Known Camera Parameters. *CoRR*, 2021.

[50] Wilson Wong, Xiao Chen Bai, Alan Brown, Israel S. Fernandez, Eric Hanssen, Melanie Condron, Yan Hong Tan, Jake Baum, and Sjors H.W. Scheres. Cryo-EM structure of the Plasmodium falciparum 80S ribosome bound to the anti-protozoan drug emetine. *eLife*, 2014.

[51] Lin Yen-Chen, Pete Florence, Jonathan T. Barron, Alberto Rodriguez, Phillip Isola, and Tsung-Yi Lin. iNeRF: Inverting Neural Radiance Fields for Pose Estimation. *arXiv*, 2020.

[52] Anna Yershova, Swati Jain, Steven M LaValle, and Julie C Mitchell. Generating Uniform Incremental Grids on SO(3) Using the Hopf Fibration. *The International Journal of Robotics Research*, 29(7):801–812, May 2010.

[53] Ellen D. Zhong. zhonge/cryodrgn_empiar: Initial release, Jan. 2021.

[54] 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.

[55] Ellen D Zhong, Tristan Bepler, Joseph H Davis, and Bonnie Berger. Reconstructing continuous distributions of 3D protein structure from cryo-EM images. *ICLR*, 2020.

[56] Ellen D Zhong, Adam Lerer, Joseph H Davis, and Bonnie Berger. Exploring generative atomic models in cryo-em reconstruction. *arXiv preprint arXiv:2107.01331*, 2021.

[57] Jasenko Zivanov, Takanori Nakane, Björn O. Forsberg, Dari Kimanius, Wim J.H. Hagen, Erik Lindahl, and Sjors H.W. Scheres. New tools for automated high-resolution cryo-EM structure determination in RELION-3. *eLife*, 7, nov 2018.