
Differentiable probabilistic models of scientific imaging with the Fourier slice theorem

Karen Ullrich* **Rianne van den Berg** **Marcus Brubaker** **David Fleet†** **Max Welling‡**
University of Amsterdam University of Amsterdam York University University of Toronto University of Amsterdam

Abstract

Scientific imaging techniques, e.g., optical and electron microscopy or computed tomography, are used to study 3D structures through 2D observations. These observations are related to the 3D object through orthogonal integral projections. For computational efficiency, common 3D reconstruction algorithms model 3D structures in Fourier space, exploiting the Fourier slice theorem. At present it is somewhat unclear how to differentiate through the projection operator as required by learning algorithms with gradient-based optimization. This paper shows how back-propagation through the projection operator in Fourier space can be achieved. We demonstrate the approach on 3D protein reconstruction. We further extend the approach to learning probabilistic 3D object models. This allows us to predict regions of low sampling rates or to estimate noise. Higher sample efficiency can be reached by utilizing the learned uncertainties of the 3D structure as an unsupervised estimate of model fit. Finally, we demonstrate how the reconstruction algorithm can be extended with amortized inference on unknown attributes such as object pose. Empirical studies show that joint inference of the 3D structure and object pose becomes difficult when the underlying object contains symmetries, in which case pose estimation can easily get stuck in local optima, inhibiting a fine-grained high-quality estimate of the 3D structure.

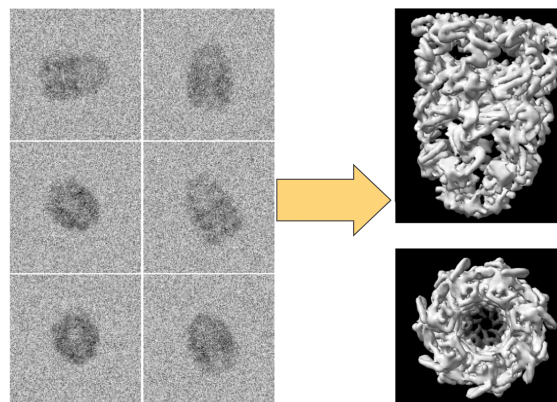


Figure 1: Example of electron cryo-microscopy with the GroEL-GroES protein (Xu et al., 1997). *Left*: 2D observations obtained by projections with an electron beam. *Right*: Two different views of the ground truth 3D protein structure represented by its electron density.

1 Introduction

The main goal of many scientific imaging methods is to reconstruct a $(d + 1)$ -dimensional structure $\mathbf{v} \in \mathcal{V} \subseteq \mathbb{R}^{D^{d+1}}$ from N (d) -dimensional observations $\mathbf{x}_n \in \mathcal{I} \subseteq \mathbb{R}^{D^d}$, where d is either one or two. For the sake of simplicity we will talk about the case $d = 2$ in the rest of this work. The contributions of this paper are:

1. We view the process of image formation through a graphical model in which latent variables correspond to physical quantities such as the hidden structure \mathbf{v} or the relative orientation/pose of a specimen. This enables one to predict errors in the reconstruction of 3D structures through uncertainty estimates. This is especially interesting when objects \mathbf{v} are only partially observable, as is the case in certain medical scans, such as breast cancer scans. Moreover, uncertainty prediction enables more data

* mail.karen.ullrich@gmail.com

† CIFAR AI Chair, Vector Institute, CIFAR LMB Program

‡ Qualcomm, CIFAR LMB Program

efficient model validation.

2. Based on the aforementioned innovations, we propose a new method for (unsupervised) reconstruction evaluation. Particularly, we demonstrate that learned uncertainties can replace currently used data inefficient methods of evaluation (see Section 6.2). We thus learn better model fits than traditional methods given the same amount of data.
3. We extend current approaches such as (Jaitly et al., 2010) to describe the generative process as a differentiable map by adopting recent techniques from the deep learning community (Jaderberg et al., 2015; Rezende et al., 2016). We demonstrate that this enables more advanced joint inference schemes over object pose and structure estimation.

Our experimental validation focuses on single particle electron cryo-microscopy (cryoEM). CryoEM is a challenging scientific imaging task, as it suffers from complex sources of observation noise, low signal to noise ratios, and interference corruption. Interference corruption attenuates certain Fourier frequencies in the observations. Radiation exposure is minimized because electron radiation severely damages biological specimens during data collection. Minimal radiation, however, leads to low signal-to-noise ratios, where sensor cells record relatively low electron counts. Since imaging techniques like CT suffer from a subset of these difficulties, we believe that evaluating and analyzing our method on cryoEM problems is appropriate.

2 Background and related work

Modelling nano-scale structures such as proteins or viruses is a central task in structural biology. By freezing such structures and subsequently projecting them via a parallel electron beam to a sensor grid (see figure 2), CryoEM enables reconstruction and visualization of such structures. The technique has been described as revolutionary because researchers are capable of observing structures that cannot be crystallized, as required for X-ray crystallography (Rupp, 2009).

The main task of reconstructing the structure from projections in cryoEM, and the wider field of medical imaging, is somewhat similar to multi-view scene reconstruction from natural images. There are, however, substantial differences. Most significantly, the projection operation in medical imaging is often an orthogonal integral projection, while in computer vision it is a non-linear perspective projection for which materials exhibit different degrees of opacity. Thus, the generative model in computer vision is more complex. Medical imaging domains,

on the other hand, face significant noise and measurement uncertainties, with signal-to-noise ratios as low as 0.05 (Baxter et al., 2009).

Most CryoEM techniques (De la Rosa-Trevín et al., 2013; Grigorieff, 2007; Scheres, 2012; Tang et al., 2007) iteratively refine an initial structure by matching a maximum a posteriori (MAP) estimate of the pose (orientation and position) under the proposal structure with the image observation. These approaches suffer from a range of problems such as high sensitivity to poor initialization (Henderson et al., 2012). In contrast to this approach, and closely related to our work, (Brubaker et al., 2015; Punjani et al., 2017) treat poses and structure as latent variables of a joint density model. MAP estimation enables efficient optimization in observation space. Previous work (Sigworth, 1998; Scheres et al., 2007a; Jaitly et al., 2010; Scheres, 2012) has suggested full marginalization, however due to its cost, it is usually intractable.

This paper extends the MAP approach by utilizing variational inference to approximate intractable integrals. Further, reparameterizing posterior distributions enables gradient based learning (Kingma and Welling, 2013; Rezende et al., 2014a). To our knowledge this is the first such approach that provides an efficient way to learn approximate posterior distributions in this domain.

3 Observation Formation Model

Given a structure \mathbf{v} , we consider a generic generative model of observations, one that is common to many imaging modalities. As a specific example, we take the structure \mathbf{v} to be a frozen (i.e. cryogenic) protein complex, although the procedure described below applies as well to CT scanning and optical microscopy. \mathbf{v} is in a specific pose \mathbf{p}_n relative to the direction of the electron radiation beam. This yields a pose-conditional projection, with observation \mathbf{x}_n . Specifically, the pose $\mathbf{p}_n = (\mathbf{r}_n, \mathbf{t}_n)$, consists of $\mathbf{r}_n \in SO(3)$, corresponding to the rotation of the object with respect to the microscope coordinate frame, and $\mathbf{t}_n \in E(3)$, the translation of the structure with respect to the origin.

The observations are then generated as follows: The specimen in pose \mathbf{p}_n is subjected to radiation (the electron beam), yielding an orthographic integral projection onto the plane perpendicular to the beam. The expected projection can be formulated as

$$\bar{\mathbf{x}}_n = P T_{\mathbf{t}_n} R_{\mathbf{r}_n} \mathbf{v}. \quad (1)$$

Here P is the projection operator, $T_{\mathbf{t}_n}$ is the linear translation operator for translation \mathbf{t}_n , and $R_{\mathbf{r}_n}$ is the linear operator corresponding to rotation \mathbf{r}_n . Without loss of generality we can choose the projection direction to be

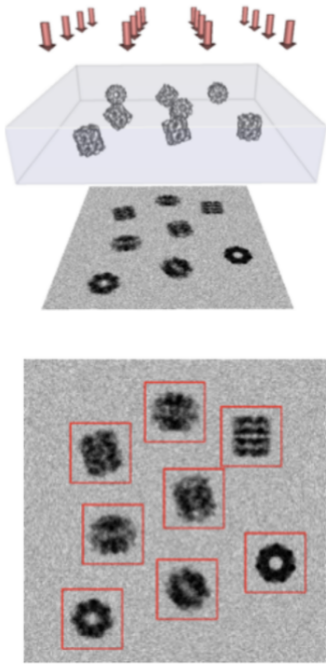


Figure 2: *Top*: Image formation on the example of cryo EM: The parallel Electron beam projects the electron densities on a surface where a grid of DDD sensors record the number of electrons that hit it. *Bottom*: To detect the projections (outlines in red) an algorithm seeks out areas of interest (Langlois et al., 2014; Zhao et al., 2013) (Figure from (Pintilie, 2010)).

along the z -direction \mathbf{e}_z . When the projection is recorded with a discrete sensor grid (i.e., sampled), information beyond the Nyquist frequency is aliased. Additionally, the recording is corrupted with noise stemming from the stochastic nature of electron detection events and sensor failures (Egelman, 2016). Low doses are necessary since electron exposure causes damage to sensitive biological molecules. Logically, the effect is more severe for smaller objects of study.

Many sophisticated noise models have been proposed for these phenomena (Faruqi et al., 2003; Vulović et al., 2013; Scheres et al., 2007b). In this work, for simplicity, we assume isotropic Gaussian noise; i.e.,

$$p(\mathbf{x}_n | \mathbf{p}_n, \mathbf{v}) = \mathcal{N}(\mathbf{x}_n | \bar{\mathbf{x}}_n, \mathbb{1}\sigma_\epsilon^2), \quad (2)$$

where σ_ϵ models the magnitude of the observation noise. The image formation process is depicted in Figure 2.

The final section below discusses how one can generalize to more sophisticated (learnable) noise models. Note that we do not model interference patterns caused by elec-

tron scattering, called defocus and modelled with a contrast transfer function (CTF). This will lead to less realistic generative models, however we see the problem of CTF estimation as somewhat independent of our problem. Ideally, we would like to model the CTF across multiple datasets, but we leave this to future work.

4 Back-propagating through the generative model

In this section we aim to bridge the gap between our knowledge of the generative process $p(\mathbf{x}_n | \mathbf{p}_n, \mathbf{v})$ and a differentiable mapping that facilitates direct optimization of hidden variables $(\mathbf{p}_n, \mathbf{v})$ with gradient-descent style schemes. We start with an explanation of a naive differentiable implementation in position space, followed by a computationally more efficient version by shifting the computations to the Fourier domain (momentum space).

4.1 Naive implementation: project in position space

Our goal is to optimize the conditional log-likelihood $\log p(\mathbf{x}_n | \mathbf{p}_n, \mathbf{v})$ with respect to the unobserved \mathbf{p}_n and \mathbf{v} , maximizing the likelihood of the 2D observations. This requires equation (2) to be a differentiable operator with respect to \mathbf{p}_n and \mathbf{v} . Note that the dependence on \mathbf{p}_n and \mathbf{v} is fully determined by equation (1). In order to achieve this, we first need to apply the group action $R_{\mathbf{r}_n}$ onto \mathbf{v} . Matrix representations of the group action such as the Euler angles matrix are defined on the sampling grid $G = \{(\nu_1^{(j)}, \nu_2^{(j)}, \nu_3^{(j)})\}_{j=1}^{D^3}$ of \mathbf{v} rather than the voxel values $\{v_j\}_{j=1}^{D^3}$. For example, the action induced by a rotation around the z -axis by an angle α on the position $(\nu_1^{(j)}, \nu_2^{(j)}, \nu_3^{(j)})$ of an arbitrary voxel j can be written as,

$$R_\alpha \cdot \nu^{(j)} = \begin{pmatrix} \cos(\alpha) & -\sin(\alpha) & 0 \\ \sin(\alpha) & \cos(\alpha) & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \nu_1^{(j)} \\ \nu_2^{(j)} \\ \nu_3^{(j)} \end{pmatrix}. \quad (3)$$

This entails two problems. First, the volume after transformation should be sampled at the same grid points as before. This requires interpolation. Second, to achieve a differentiable map we need to formulate the transformation of position values as a transformation of the voxel values. Jaderberg et al. (2015) offers a solution to both problems, known as differentiable sampling¹.

The j -th voxel $v'_j = (\mathbf{v}')_j$ of the transformed volume, $\mathbf{v}' = R_{\mathbf{r}_n} \mathbf{v}$, with index vector $\zeta^{(j)}$, can be expressed

¹Originally invented to learn affine transformations on images to ease the classification task for standard neural networks, the approach has since been extend to 3D reconstruction problems from images (Rezende et al., 2016).

as a weighted sum of all voxels before transformation $\{v_i, \nu^{(i)}\}_{i=1}^{D^3}$. The weights are determined by a sampling kernel $k(\cdot)$, the argument of which is the difference between the transformed voxel's position $\zeta^{(j)}$ and all transformed sampling grid vectors $R_\alpha \cdot \nu^{(i)}$:

$$v'_j = \sum_{i=1}^{D^3} v_i \cdot k(R_\alpha^{-1} \cdot \zeta^{(j)} - \nu^{(i)}) \quad (4)$$

A popular kernel in this context is the linear interpolation sampling kernel²

$$k(R_\alpha^{-1} \cdot \zeta^{(j)} - \nu^{(i)}) = \prod_{m=1}^3 \max(0, 1 - |(R_\alpha^{-1} \cdot \zeta^{(j)})_m - \nu_m^{(i)}|). \quad (5)$$

Computing one voxel v'_j only requires a sum over 8 voxels from the original structure. These are determined by flooring and ceiling the elements of $(R_\alpha^{-1} \zeta^{(j)})_m$. Furthermore, the partial derivatives are provided by,

$$\begin{aligned} \frac{\partial v'_j}{\partial v_i} &= k(R_\alpha^{-1} \cdot \zeta^{(j)} - \nu^{(i)}) \quad (6) \\ \frac{\partial v'_j}{\partial (R_\alpha^{-1} \zeta^{(j)})_m} &= \sum_{i=1}^{D^3} v_i \prod_{l \neq k} \max(0, 1 - |(R_\alpha^{-1} \zeta^{(j)})_l - \nu_l^{(i)}|) \\ &\quad \cdot \begin{cases} 0 & \text{if } |(R_\alpha^{-1} \zeta^{(j)})_m - \nu_m^{(i)}| \geq 1 \\ -1 & \text{elif } (R_\alpha^{-1} \zeta^{(j)})_m \geq \nu_m^{(i)} \\ 1 & \text{elif } (R_\alpha^{-1} \zeta^{(j)})_m < \nu_m^{(i)} \end{cases}. \quad (7) \end{aligned}$$

This framework was originally proposed for any differentiable kernel and any differentiable affine position transformation $\nu \rightarrow \zeta$. In our setting, we restrict ourselves to linear interpolation kernels. The group actions represented by R_r are affine. In this work we represent rotations as Euler angles by using the Z-Y-Z convention. One could also use quaternions or exponential maps. As with rotation, the translation operation is also a transformation of the voxel grid, rather than the voxel values. Thus, equation (4) can also be used to obtain a differentiable translation operation.

Finally, the orthogonal integral projection operator is applied by summing voxel values along one principal direction. Since the position of the hidden volume is arbitrary we can fix this direction to be the Z-axis as discussed in section 3. Denoting a volume, rotated and translated according to $\mathbf{p} = (\mathbf{r}, \mathbf{t})$, by $\mathbf{v}' = T_t R_r \mathbf{v}$, the (ζ_1, ζ_2) -th element of its expected projection is given by

$$\bar{\mathbf{x}}_n[\zeta_1, \zeta_2] = (P_{3 \rightarrow 2} \mathbf{v}')[\zeta_1, \zeta_2] = \sum_{\zeta_3} \mathbf{v}'[\zeta_1, \zeta_2, \zeta_3], \quad (8)$$

²Linear kernels are efficient and yield fairly good results. More complex ones such as the Lanczos re-sampling kernel may actually yield worse results due to smoothing.

where $\mathbf{v}'[\zeta_1, \zeta_2, \zeta_3]$ denotes the element $(\zeta_1, \zeta_2, \zeta_3)$ of \mathbf{v}' .

This concludes a naive approach to modelling a differential map of the expected observation in position space. This approach is not particularly efficient, as according to equation (8) we need to interpolate all D^3 voxels to compute one D^2 dimensional observation. Moreover, back-propagating through this mapping requires transporting gradients through all voxels. Next, we show how to reduce the cost of this naive approach without a loss of precision by shifting the problem to the Fourier domain.

4.2 Projection-slice theorem

The projection-slice theorem or Fourier-slice theorem states the equivalence between the Fourier transform F_d of the projection $P_{d' \rightarrow d}$ of a d' dimensional function $f(r)$ onto a d -dimensional submanifold $\mathcal{F}_d P_{d' \rightarrow d} f(r)$ and a d -dimensional slice of the d' -dimensional Fourier transform of that function. This slice is a d -dimensional linear submanifold through the origin in the Fourier domain that is parallel to the projection submanifold $\mathcal{S}_d \mathcal{F}_{d'} f(r)$.

In our setting, given an axis-aligned projection direction (see equation (1)) and the discrete grid G , the expected observation in 2D Fourier space is equal to a central slice through the Fourier transformed 3D structure parallel to the projection direction \mathbf{e}_z :

$$\begin{aligned} \mathcal{F}_2 \bar{\mathbf{x}} &= \mathcal{F}_2 P_{3 \rightarrow 2} \mathbf{v}' \\ &= \mathcal{S}_2 (\mathcal{F}_3 \mathbf{v}') = \mathcal{S}_2 \hat{\mathbf{v}}', \quad (9) \end{aligned}$$

where \mathcal{F}_2 and \mathcal{F}_3 denote discrete Fourier transforms in 2 and 3 dimensions. The slice operator \mathcal{S} is the Fourier equivariant of the projection operator. In our case, it is applied as follows:

$$(\mathcal{S}_2 \hat{\mathbf{v}}')[\omega_1, \omega_2] = \hat{\mathbf{v}}'[\omega_1, \omega_2, 0], \quad (10)$$

where ω_m are Fourier indexes.

This allows one to execute the generative model in position or momentum space. It has proven more efficient for most reconstruction algorithms to do computation in the Fourier domain (Sigworth, 2016). This also applies to our algorithm: (i) We reconstruct the structure in the Fourier domain. This means we only need to apply an inverse Fourier transformation at the end of optimization. (ii) We may save the Fourier transformed expected projections a-priori, further this is easily parallelized. Thus even though in its original formulation we can not expect a computational benefit, when sharing the computation across many data points to reconstruct one latent structure the gain is significant. We elaborate on this point with respect to differentiation below.

4.3 Differentiable orthographic integral projection

We incorporate the Fourier Slice Theorem into our approach to build an efficient differentiable generative model. For this we translate all elements of the model to their counterparts in the Fourier domain. The Gaussian noise model becomes (Havin and Jricke, 1994),

$$p(\mathcal{F}_2 \mathbf{x}_n | \mathbf{p}_n, \hat{\mathbf{v}}) = \mathcal{N}(\mathcal{F}_2 \mathbf{x}_n | \mathcal{F}_2 \bar{\mathbf{x}}_n, \mathbb{1} \frac{\sigma_z^2}{2}), \quad (11)$$

where $\hat{\mathbf{v}} = \mathcal{F}_3 \mathbf{v}$. In the following, we aim to determine an expression for the expected Fourier projection $\mathcal{F}_2 \bar{\mathbf{x}}_n = \mathcal{F}_2(P_{3 \rightarrow 2} T_{\mathbf{t}_n} R_{\mathbf{r}_n} \mathbf{v})$ by deriving the operators counterparts $\mathcal{F}_2(P_{3 \rightarrow 2} T_{\mathbf{t}_n} R_{\mathbf{r}_n} \mathbf{v}) = \hat{P}_{3 \rightarrow 2} \hat{T}_{\mathbf{t}_n} \hat{R}_{\mathbf{r}_n} \hat{\mathbf{v}}$.

We start by noting that it is useful to keep $\hat{\mathbf{v}}$ in memory $\hat{\mathbf{v}}$ to avoid computing the 3D discrete Fourier transform multiple times during optimization. The inverse Fourier transform $F_3^{-1} \hat{\mathbf{v}} = \mathbf{v}$ is then only applied once, after convergence of the algorithm. Next, we restate that the Fourier transformation is a rotation-equivariant mapping $\hat{R}_{\mathbf{r}_n} = R_{\mathbf{r}_n}$ (Chirikjian and Kyatkin, 2000). This means the derivations from Section 4.1 with respect to the rotation apply in this context as well. A translation in the Fourier domain $\hat{T}_{\mathbf{t}_n}$, however, induces a re-weighting of the original Fourier coefficients,

$$(\hat{T}_{\mathbf{t}_n} R_{\mathbf{r}_n} \hat{\mathbf{v}})[\omega_1, \omega_2, \omega_3] = e^{-i2\pi \mathbf{t}_n \cdot \boldsymbol{\omega}} (R_{\mathbf{r}_n} \hat{\mathbf{v}})[\omega_1, \omega_2, \omega_3]. \quad (12)$$

Finally, the last section established the equivalence of the slice and the projection operator $\hat{P}_{3 \rightarrow 2} = \mathcal{S}_2$ (see equation (9)) in momentum and position space. Specifically, for the linear interpolation kernel, we compute the set of interpolation points by flooring and ceiling the elements of the vector $R_{\mathbf{r}_n}^{-1} \boldsymbol{\omega}^{(j)}, \forall \boldsymbol{\omega}^{(j)} = (\omega_1^{(j)}, \omega_2^{(j)}, 0)$. This entails 6 interpolation candidates per voxel of the central slice, in total $6D^2$. Remember, this computation above involved D^3 voxels and $6D^3$ candidates.

We can further improve the efficiency of the algorithm by swapping the projection and translation operators. That is, due to parallel radiation, and hence orthographic projection,

$$PT_{\mathbf{t}_n} R_{\mathbf{r}_n} \mathbf{v} = T_{\boldsymbol{\tau}_n} P R_{\mathbf{r}_n} \mathbf{v}, \quad (13)$$

where $\boldsymbol{\tau}_n = (\mathbf{t}_n \cdot \mathbf{e}_x) \mathbf{e}_x + (\mathbf{t}_n \cdot \mathbf{e}_y) \mathbf{e}_y$. This is more efficient because we reduce the translation to a two dimensional translation. Thus this modifies equation (12) to its two dimensional equivalent.

For the naive implementation the cost of a real space forward projection is $O(D^3)$. In contrast, converting the volume to the Fourier space $O(D^3 \log D)$, projecting $O(D^2)$ and applying the inverse Fourier transform $O(D^2 \log D)$. At first glance this implies a higher computational cost. However, for large datasets the cost

of transforming the volume is amortized over all data points. For gradient descent schemes, we iterate over the dataset more than once, hence the cost of Fourier transforming the observations is further amortized. Furthermore, it is often reasonable to consider only a subset of Fourier frequencies, so back projection becomes $O(r^2)$ with $r < D$. The efficiency of this algorithm in the context of cryo-EM was first recognized by Grigorieff (1998). (We provide code for the differentiable observation model and in particular for the Fourier operators: <https://github.com/KarenUllrich/pytorch-backprojection>.)

5 Variational inference

Here we describe the variational inference procedure for 3D reconstruction in Fourier space, enabled by the Fourier slice theorem. We assume we have a dataset of observations from a single type of protein with ground truth structure \mathbf{v} , and its Fourier transform $\hat{\mathbf{v}} = \mathcal{F}_3 \mathbf{v}$. We consider two scenarios. In the first, both the pose of the protein and the projection are observed, and inference is only performed over the global protein structure. In the second, the pose of the protein for each observation is unknown. Therefore, inference is done over poses and the protein structure.

The first scenario is synonymous with the setting in tomography, where we observe a frozen cell or larger complex positioned in known poses. This case is often challenging because the sample cannot be observed from all viewing angles. For example, in cryo-EM tomography the specimen frozen in an ice slice can only be rotated till the verge of the slice comes into view. We find similar problems in CT scanning, for example in breast cancer scans. The second scenario is relevant for cryo-EM single particle analysis. In this case multiple identical particles are confined in a frozen sample and no information on their structure or position is available a priori.

In this work we lay the foundations for doing inference in either of the two scenarios. However, our experiments demonstrate that joint inference over the poses and the 3D structure is very sensitive to getting stuck in local minima that correspond to approximate symmetries of the 3D structure. Therefore, the main focus of this work is the setting where the poses are observed.

5.1 Inference over the 3D structure

Here the data comprise image projections and poses, $\{(\mathbf{x}_n, \mathbf{p}_n)\}_{n=1}^N$, with Fourier transformed projections denoted $\hat{\mathbf{x}}_n = \mathcal{F}_2 \mathbf{x}_n$. Our goal is to learn a model $q(\hat{\mathbf{v}})$ of the latent structure that as closely as possible resembles the true posterior $p(\hat{\mathbf{v}} | \{\hat{\mathbf{x}}_n\}, \{\mathbf{p}_n\})$. For this, we assume

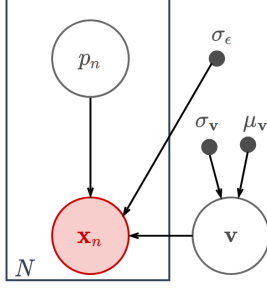


Figure 3: *Graphical model*: Latent structure \mathbf{v} , pose p_n and noise σ_ϵ can be learned from observations \mathbf{x}_n through back-propagation. The latent structure distribution is thereby characterized by a set of parameters, in the Gaussian example μ_v and σ_v .

a joint latent variable model $p(\{\hat{\mathbf{x}}_n\}_{n=1}^N, \hat{\mathbf{v}}|\{\mathbf{p}_n\}_{n=1}^N) = p(\{\hat{\mathbf{x}}_n\}_{n=1}^N|\{\mathbf{p}_n\}_{n=1}^N, \hat{\mathbf{v}})p(\hat{\mathbf{v}})$. To avoid clutter below, we use short-hand notations like $\{\hat{\mathbf{x}}_n\}$ for $\{\hat{\mathbf{x}}_n\}_{n=1}^N$.

Specifically, we minimize an upper bound to the Kullback-Leibler (KL) divergence:

$$\begin{aligned} & D_{\text{KL}} [q(\hat{\mathbf{v}})\|p(\hat{\mathbf{v}}|\{\hat{\mathbf{x}}_n\}, \{\mathbf{p}_n\})] \\ & \geq - \int d\hat{\mathbf{v}} q(\hat{\mathbf{v}}) \ln \left(\frac{p(\{\hat{\mathbf{x}}_n\}|\{\mathbf{p}_n\}, \hat{\mathbf{v}})p(\hat{\mathbf{v}})}{q(\hat{\mathbf{v}})} \right) \\ & = \sum_{n=1}^N -\mathbb{E}_{q(\hat{\mathbf{v}})} [\ln p(\hat{\mathbf{x}}_n|\mathbf{p}_n, \hat{\mathbf{v}})] + D_{\text{KL}} [q(\hat{\mathbf{v}})\|p(\hat{\mathbf{v}})]. \quad (14) \end{aligned}$$

Here, we have assumed that, given the volume, the data are IID: $p(\{\hat{\mathbf{x}}_n\}|\{\mathbf{p}_n\}, \hat{\mathbf{v}}) = \prod_{n=1}^N p(\hat{\mathbf{x}}_n|\mathbf{p}_n, \hat{\mathbf{v}})$. We have bounded the divergence by the data log-likelihood $\ln p(\{\hat{\mathbf{x}}_n\})$, a constant with respect to $\hat{\mathbf{v}}$ and $\{\mathbf{p}_n\}$. This is equivalent to lower bounding the model evidence by introducing a variational posterior (Jordan et al., 1998).

In this work we focus on modelling $q(\hat{\mathbf{v}})$ as isotropic Gaussian distribution. The prior is assumed to be a standard Gaussian. In practice, we use stochastic gradient descent-like optimization, with the data organized in mini-batches. That is, we learn the distribution parameters $\eta = \{\mu_v, \sigma_v\}$ for $q(\hat{\mathbf{v}}) = q_\eta(\hat{\mathbf{v}})$ through stochastic optimization, efficiently by using the reparameterization trick (Kingma and Welling, 2013; Rezende et al., 2014a).

In equation (14), the reconstruction term depends on the number of datapoints. The KL-divergence between the prior and approximate posterior does not. As the sum of the mini-batch objectives should be equal to the objective of the entire dataset, the mini-batch objective is

$$\sum_{n \in D_i} -\mathbb{E}_{q(\hat{\mathbf{v}})} [\ln p(\hat{\mathbf{x}}_n|\mathbf{p}_n, \hat{\mathbf{v}})] + \frac{|D_i|}{N} D_{\text{KL}} [q(\hat{\mathbf{v}})\|p(\hat{\mathbf{v}})]. \quad (15)$$

where D_i is the set of indices of the data in minibatch i , and $|D_i|$ denotes the size of the i -th minibatch.

5.2 Joint inference over the 3D structure and poses

In the second scenario the pose of the 3D structure for each observation is unknown. The data thus comprises the observed projections $\{\mathbf{x}_n\}_{n=1}^N$. Again, we perform inference in the Fourier domain, with transformed projections $\{\hat{\mathbf{x}}_n\}_{n=1}^N$ as data. We perform joint inference over the poses $\{\mathbf{p}_n\}_{n=1}^N$ and the volume $\hat{\mathbf{v}}$. We assume the latent variable model can be factored as follows, $p(\{\hat{\mathbf{x}}_n\}, \hat{\mathbf{v}}, \{\mathbf{p}_n\}) = p(\{\hat{\mathbf{x}}_n\}|\{\mathbf{p}_n\}, \hat{\mathbf{v}})p(\{\mathbf{p}_n\})p(\hat{\mathbf{v}})$.

Upper bounding the KL-divergence, as above, we obtain

$$\begin{aligned} & D_{\text{KL}} [q(\hat{\mathbf{v}})q(\{\hat{\mathbf{x}}_n\})\|p(\hat{\mathbf{v}}, \{\hat{\mathbf{x}}_n\}, \{\mathbf{p}_n\})] \\ & \geq - \iint d^N \mathbf{p}_n d\hat{\mathbf{v}} q(\{\mathbf{p}_n\})q(\hat{\mathbf{v}}) \\ & \quad \times \ln \left(\frac{p(\{\hat{\mathbf{x}}_n\}|\{\mathbf{p}_n\}, \hat{\mathbf{v}})p(\{\mathbf{p}_n\})p(\hat{\mathbf{v}})}{q(\{\mathbf{p}_n\})q(\hat{\mathbf{v}})} \right) \\ & = \sum_{n=1}^N -\mathbb{E}_{q(\hat{\mathbf{v}})} \mathbb{E}_{q(\mathbf{p}_n)} [\ln p(\hat{\mathbf{x}}_n|\mathbf{p}_n, \hat{\mathbf{v}})] \\ & \quad + \sum_{n=1}^N D_{\text{KL}} [q(\mathbf{p}_n)\|p(\mathbf{p}_n)] + D_{\text{KL}} [q(\hat{\mathbf{v}})\|p(\hat{\mathbf{v}})]. \quad (16) \end{aligned}$$

The prior, approximate posterior and conditional likelihood all factorize across datapoints: $p(\{\mathbf{p}_n\}) = \prod_{n=1}^N p(\mathbf{p}_n)$, $q(\{\mathbf{p}_n\}) = \prod_{n=1}^N q(\mathbf{p}_n)$, and $p(\{\hat{\mathbf{x}}_n\}|\{\mathbf{p}_n\}, \hat{\mathbf{v}}) = \prod_{n=1}^N p(\hat{\mathbf{x}}_n|\mathbf{p}_n, \hat{\mathbf{v}})$. Like equation (15), for mini-batch optimization algorithms we make use of the objective

$$\begin{aligned} & \sum_{n \in D_i} -\mathbb{E}_{q(\hat{\mathbf{v}})} \mathbb{E}_{q(\mathbf{p}_n)} [\ln p(\hat{\mathbf{x}}_n|\mathbf{p}_n, \hat{\mathbf{v}})] \\ & \quad + \sum_{n \in D_i} D_{\text{KL}} [q(\mathbf{p}_n)\|p(\mathbf{p}_n)] + \frac{|D_i|}{N} D_{\text{KL}} [q(\hat{\mathbf{v}})\|p(\hat{\mathbf{v}})]. \quad (17) \end{aligned}$$

In Section 5.1, we learn the structure parameters shared across all data points. Here the pose parameters are unique per observation and therefore require separate inference procedures per data point. As an alternative, we can also learn a function that approximates the inference procedure; This is called amortized inference (Kingma and Welling, 2013). In practice, a complex parameterized function f_ϕ such as a neural network predicts the parameters of the variational posterior $\eta = f_\phi(\cdot)$.

6 Experiments

We empirically test the formulation above with simulated data from the well-known GroEL-GroES protein (Xu et al., 1997). To this end we generate three datasets, each with 40K projections onto 128×128 images at a resolution of 2.8\AA per pixel. The three datasets have signal-to-noise ratios (SNR) of 1.0, 0.04 and 0.01, referred to

below as the *noise-free*, *medium-noise* and *high-noise* cases. Figure 7 shows one sample per noise level. As previously stated in Section 3, we do not model the microscope’s defocus or electron scattering effects, as captured by the CTF (Kohl and Reimer, 2008).

Using synthetic data allows us to evaluate the algorithm with the ground-truth structure, e.g., in terms of mean-squared error (MSE). With real-data, where ground truth is unknown, the resolution of a fitted 3D structure is often quantified using *Fourier Shell Correlation* (Rosenthal and Henderson, 2003): The N observations are partitioned randomly into two sets, A and B , each of which is then independently modeled with the same reconstruction algorithm. The normalized cross-correlation coefficient is then computed as a function of frequency $f = 1/\lambda$ to assess the agreement between the two reconstructions.

Given two 3D structures, F_A and F_B , in the Fourier domain, FSC at frequency f is given by

$$FSC(f|F_A, F_B) = \frac{\sum_{f_i \in S_f} F_A(f_i) \cdot F_B(f_i)^*}{\sqrt{\sum_{f_i \in S_f} |F_A(f_i)|^2 \cdot \sum_{f_i \in S_f} |F_B(f_i)|^2}} \quad (18)$$

where S_f denotes the set of frequencies in a shell at distance f from the origin of the Fourier domain (i.e. with wavelength λ). This yields FSC curves like those in Fig. 4. The quality (resolution) of the fit can be measured in terms of the frequency at which this curve crosses a threshold τ . When one of F_A or F_B is ground truth, then $\tau = 0.5$, and when both are noisy reconstructions it is common to use $\tau = 0.143$ (Rosenthal and Henderson, 2003)). The structure is then said to be resolved to wavelength $\lambda = 1/f$ for which $FSC(f) = \tau$.

6.1 Comparison to Baseline algorithms

When poses are known, the current state-of-the-art (SOTA) is a conventional tomographic reconstruction (a.k.a. back-projection). When poses are unknown, there are several well-known SOTA cryo-EM algorithms (De la Rosa-Trevín et al., 2013; Grigorieff, 2007; Scheres, 2012; Tang et al., 2007). All provide point estimates of the 3D structure. In terms of graphical models, point estimates correspond to the case in which the posterior $q(\hat{\mathbf{v}})$ is modeled as a delta function, $\delta(\hat{\mathbf{v}}|\mu_{\mathbf{v}})$, the parameter of which is the 3D voxel array, $\mu_{\mathbf{v}}$.

We compare this baseline to a model in which the posterior $q(\hat{\mathbf{v}})$ is a multivariate diagonal Gaussian, $\mathcal{N}(\hat{\mathbf{v}}|\mu_{\mathbf{v}}, \mathbb{1}\sigma_{\mathbf{v}}^2)$. While the latent structure is modeled in Fourier domain, the spatial domain signal is real-valued.

Posterior	$\delta(\mathbf{v} \mu_{\mathbf{v}})$	$\mathcal{N}(\mathbf{v} \mu_{\mathbf{v}}, \mathbb{1}\sigma_{\mathbf{v}}^2)$
Time until converged [s]	~ 390	~ 480
MSE [10 ⁻³ /voxel]	2.29	2.62
Resolution [Å]	5.82	5.82

Table 1: Results for modelling protein structure as latent variable. Fitting a Gaussian or Dirac posterior distribution with VI leads to similar model fits, as measured by MSE and FSC between fit and ground truth with $\tau = 0.5$.

We restrict the learnable parameters, $\mu_{\mathbf{v}}$ and $\sigma_{\mathbf{v}}$, accordingly³. We use the reparameterization trick thus correlate the samples accordingly. Finally, the prior in equation (14) in a multivariate standard Gaussian, $p(\mathbf{v}) = \mathcal{N}(\mathbf{v}|\mathbf{0}, \mathbb{1})$.

Table 1 shows results with these two models, with known poses (the tomographic setting), and with *noise-free* observations. Given the sensor resolution of $r = 2.8\text{Å}$, the highest possible resolution would be the Nyquist wavelength of 5.6Å . Our results show that both models approach this resolution, and in reasonable time.

6.2 Uncertainty estimation leads to data efficiency

In this section, we explore how modelling the latent structure with uncertainty can improve data efficiency. For this, recall that FSC is computed by comparing reconstructions based on dataset splits, A and B . As an alternative, we propose to utilize the learned model uncertainties $\sigma_{\mathbf{v}}$ to achieve a similar result. We thus only need one model fit that includes both dataset splits. Specifically, we propose to extend a measure first presented in Unser et al. (1987): the spectral SNR to the 3D case, and hence refer to it as spectral shell SNR (SS-SNR). When modelling the latent structure as diagonal Gaussian $\mathcal{N}(\mathbf{v}|\mu_{\mathbf{v}}, \mathbb{1}\sigma_{\mathbf{v}}^2)$, the SS-SNR can be computed to be

$$\alpha(f) = \frac{\sum_{f_i \in S_f} |\mu_{\mathbf{v}}(f_i)|^2}{\sum_{f_i \in S_f} \sigma_{\mathbf{v}}^2(f_i)}. \quad (19)$$

Following the formulation by Unser et al. (1987), we can then express the FSC in terms of the SS-SNR, i.e., $FSC \approx \alpha/(1 + \alpha)$.

Figure 4 shows FSC curves based on reconstructions from the *medium-noise* (top) and *high-noise* (bottom)

³That is $\forall \phi \in \{\mu_{\mathbf{v}}, \sigma_{\mathbf{v}}\}$: $\Re(\phi)[\zeta] = \Re(\phi)[- \zeta]$ and $\Im(\phi)[\zeta] = -\Im(\phi)[- \zeta]$ with $\zeta = (\zeta_1, \zeta_2, \zeta_3)$.

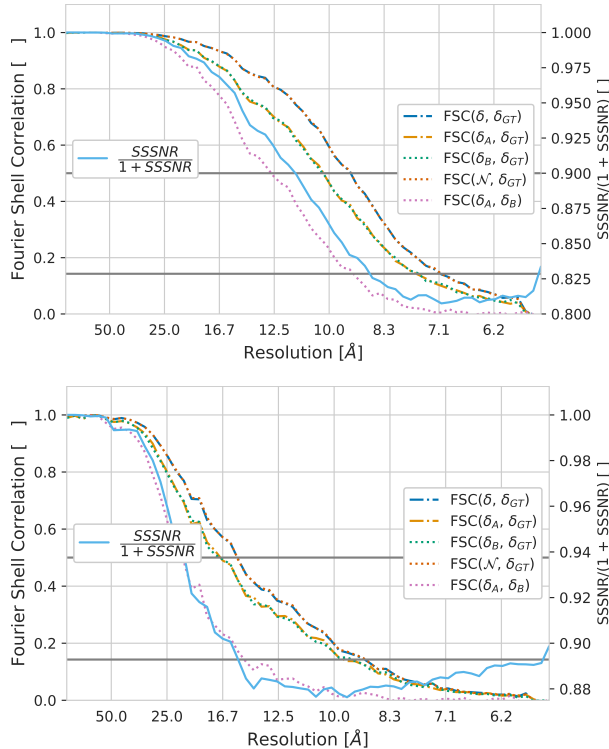


Figure 4: The FSC curves for various model fits. The grey lines indicate resolution thresholds. The higher the resolution of a structure the better the model fit. We contrast the FSC curves with the proposal we make to evaluate model fit.

datasets. First, we aim to demonstrate that the FSC curves between the Gaussian fit (with all data) vs ground truth, and the MAP estimate model with all data vs ground truth, i.e., $FSC(f|\delta, \delta_{GT})$ and $FSC(f|\mathcal{N}, \delta_{GT})$, yield the same fit quality. Note that we would not usually have access to the ground truth structure. Secondly, because in a realistic scenario we would not have access to the ground truth we would need to split the dataset in two. For this we evaluate the FSC between ground truth and two disjoint splits of the dataset $FSC(f|\delta_A, \delta_{GT})$ and $FSC(f|\delta_B, \delta_{GT})$. This curve not surprisingly lies under the previous curves. Also note, that the actual measure we would consider $FSC(f|\delta_A, \delta_B)$ is more conservative. Finally, we show that $\alpha/(1+\alpha)$ curve has the same inflection points as the FSC curve. As one would expect, it lies above the conservative $FSC(f|\delta_A, \delta_B)$.

Using α one can quantify the model fit with learned uncertainty rather than FSC curve. As a consequence there is no need to partition the data and perform two separate reconstructions, each with only half the data.

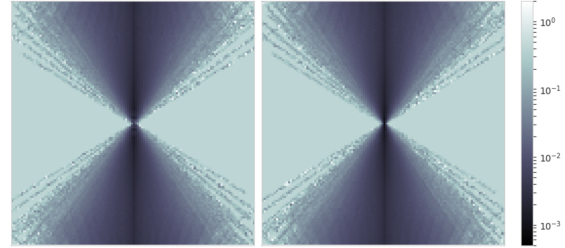


Figure 5: Center slice through the learned Fourier volume uncertainties σ_v . *Left*: real part, *Right*: imaginary part. We learn the model fit with observations coming only from a 30° cone, a scenario similar to breast cancer scans where observations are available only from some viewing directions. Uncertainty close to 1 means that the model has no information in these areas, close to zero represents areas of high sampling density. In contrast to other models, our model can identify precisely where information is missing (high variance).

6.3 Uncertainty identifies missing information

Above we discussed how global uncertainty estimation can help estimate the quality of fit. Here we demonstrate how local uncertainty can help evaluate the local quality of fit⁴. In many medical settings, such as breast cancer scans, limited viewing directions are available. This is an issue in tomography, and also occurs in single particle cryo-EM when the distribution of particle orientations around the viewing sphere are highly anisotropic. To recreate the same effect we construct a dataset of 40K observations, as before with no noise to separate the sources of information corruption⁵. We restrict viewing directions to Euler angles (α, β, γ) with $\beta \in (-15, +15)$; i.e., no observations outside a 30° cone. As above, we assume a Gaussian posterior over the latent protein structure. Figure 5 shows the result. We can show that the uncertainty the model has learned correlates with regions in which data has been observed (uncertainty close to 0) and has not been observed (uncertainty close to 1). Due to the pressure from the KL-divergence (see equation (14)) the latter areas of the model default to the prior $\mathcal{N}(\mathbf{v}|\mathbf{0}, \mathbb{1})$. This approach can be a helpful method to identify areas of low local quality of fit.

⁴Other studies recognize the importance of local uncertainty estimation, measuring FSC locally in a wavelet basis (Cardone et al., 2013; Kucukelbir et al., 2014; Vilas et al., 2018).

⁵For completeness we present the same experiment with noise in the appendix.

6.4 Limits: Treating poses as random variables

Extending our method from treating structures as latent variables to treating poses as latent variables is difficult. In the following we analyze why this is the case. Note though that, our method of estimating latent structure can be combined with common methods of pose estimation such as branch and bound (Punjani et al., 2017) without losing any of the benefits we offer. However, ideally it would be interesting to also learn latent pose posteriors. This would be useful to for example detect outliers in the dataset which is common in real world scenarios.

In an initial experiment, we fix the volume to the ground truth. We subsequently only estimate the poses with a simple Dirac model for each data point $\mathbf{p}_n \sim \delta(\mathbf{p}_n | \mu_{\mathbf{p}_n})$. In figure 6, we demonstrate the problem of SGD-based learning for this. For one example (samples on top), we show its true error surface, the global optimum (red star) and the changes over the course of optimization (red line). We observe that due to the high symmetries in the structure, the true pose posterior error surface has many symmetries as well. An estimate depending on its starting position, seems to converge to the closest local optimum only rather than the global one. We would hope to be able to fix this problem in the future by applying more advanced density estimation approaches.

7 Model Criticism and Future Work

This paper introduces practical probabilistic models into the scientific imaging pipeline, where practical refers to scalability through the use of the reparameterization trick. We show how to turn the operators in the pipeline into differentiable maps, as this is required to apply the trick. The main focus of the experiments is to show why this novelty is important, addressing issues such as data efficiency, local uncertainty, and cross validation. Specifically, we found that a parameterized distribution, i.e. the Gaussian, achieves the same quality of fit as a point estimate, i.e. the dirac, while relying on less data. We conclude that our latent variable model is a suitable building block. It can be plugged into many SOTA approaches seamlessly, such as (De la Rosa-Trevín et al., 2013; Grigorieff, 2007; Scheres, 2012; Tang et al., 2007). We also established that the learned uncertainty is predictive of locations with too few samples. Finally, we demonstrated the limits of our current methods in treating poses as latent variables. This problem, however, does not limit the applicability of our method to latent structures. We thus propose to combine common pose estimation with our latent variable structure estimation. This method benefit from the uncertainty measure but also find globally optimal poses.

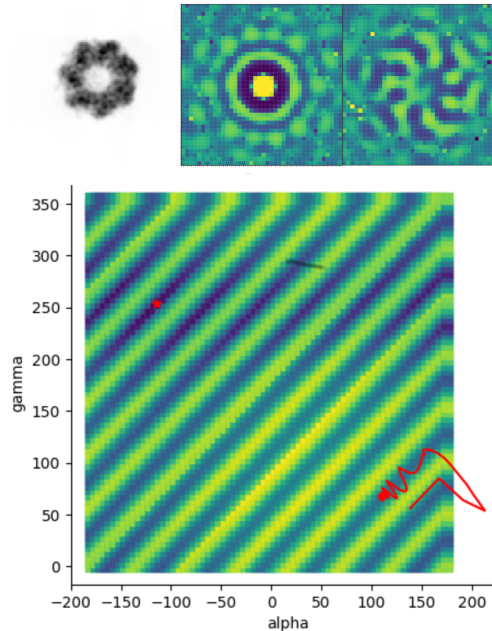


Figure 6: Example of gradient descent failing to estimate the latent pose of a protein. *Small images from left to right:* observation, real and imaginary part of the Fourier transpose. *Large figure:* Corresponding error surface of the poses for 2 of 3 Euler angles. The red curve shows the progression of the pose estimate over the course of optimization. It is clear that the optimization fails to recover the true global optimum (red star).

In future work we hope to find a way to efficiently learn pose posterior distributions as well. We hope that a reasonable approach would be to use multi-modal distributions and thus more advanced density estimation techniques. We will also try to incorporate amortized inference, mentioned in Section 5. Amortization would give the additional advantage of being able to transfer knowledge from protein to protein. Transfer could then lead to more advanced noise and CTF models. Bias in transfer will be a key focus of this effort; i.e., we only want to transfer features of the noise and not the latent structure. Another problem we see with the field of reconstruction algorithms is that the model evaluation can only help to detect variance but not bias in a model class. This is a problem with FSC comparison, but also with our proposal. We believe that an estimate of the data-log-likelihood of a hold out test dataset is generally much better suited. In a probabilistic view, this can be achieved by importance weighting the ELBO (Rezende et al., 2014b).

Acknowledgements

We thank the reviewers for their valuable feedback, in particular *AnonReviewer1*. This research was funded in part by Google and by the Canadian Institute for Advanced Research.

References

- Alemi, A. A., Poole, B., Fischer, I., Dillon, J. V., Saurous, R. A., and Murphy, K. (2017). Fixing a broken elbo. *arXiv preprint arXiv:1711.00464*.
- Baxter, W. T., Grassucci, R. A., Gao, H., and Frank, J. (2009). Determination of signal-to-noise ratios and spectral snrs in cryo-em low-dose imaging of molecules. *Journal of structural biology*, 166(2):126–132.
- Brubaker, M. A., Punjani, A., and Fleet, D. J. (2015). Building proteins in a day: Efficient 3d molecular reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3099–3108.
- Cardone, G., Heymann, J. B., and Steven, A. C. (2013). One number does not fit all: Mapping local variations in resolution in cryo-em reconstructions. *Journal of structural biology*, 184(2):226–236.
- Chirikjian, G. and Kyatkin, A. (2000). *Engineering Applications of Noncommutative Harmonic Analysis: With Emphasis on Rotation and Motion Groups*. CRC Press.
- De la Rosa-Trevín, J., Otón, J., Marabini, R., Zaldivar, A., Vargas, J., Carazo, J., and Sorzano, C. (2013). Xmipp 3.0: an improved software suite for image processing in electron microscopy. *Journal of structural biology*, 184(2):321–328.
- Egelman, E. H. (2016). The current revolution in cryo-em. *Biophysical journal*, 110(5):1008–1012.
- Faruqi, A., Cattermole, D., Henderson, R., Mikulec, B., and Raeburn, C. (2003). Evaluation of a hybrid pixel detector for electron microscopy. *Ultramicroscopy*, 94(3-4):263–276.
- Grigorieff, N. (1998). Three-dimensional structure of bovine nadh: ubiquinone oxidoreductase (complex i) at 22 Å in ice. *Journal of molecular biology*, 277(5):1033–1046.
- Grigorieff, N. (2007). Frealign: high-resolution refinement of single particle structures. *Journal of structural biology*, 157(1):117–125.
- Havin, V. and Jricke, B. (1994). *The Uncertainty Principle in Harmonic Analysis*. Springer-Verlag.
- Henderson, R., Sali, A., Baker, M. L., Carragher, B., Devkota, B., Downing, K. H., Egelman, E. H., Feng, Z., Frank, J., Grigorieff, N., et al. (2012). Outcome of the first electron microscopy validation task force meeting. *Structure*, 20(2):205–214.
- Jaderberg, M., Simonyan, K., Zisserman, A., et al. (2015). Spatial transformer networks. In *Advances in Neural Information Processing Systems*, pages 2017–2025.
- Jaitly, N., Brubaker, M. A., Rubinstein, J. L., and Lilien, R. H. (2010). A bayesian method for 3d macromolecular structure inference using class average images from single particle electron microscopy. *Bioinformatics*, 26(19):2406–2415.
- Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1998). An introduction to variational methods for graphical models. In *Learning in graphical models*, pages 105–161. Springer.
- Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.
- Kohl, H. and Reimer, L. (2008). *Transmission electron microscopy: physics of image formation*. Springer.
- Kucukelbir, A., Sigworth, F. J., and Tagare, H. D. (2014). Quantifying the local resolution of cryo-em density maps. *Nature methods*, 11(1):63.
- Langlois, R., Pallesen, J., Ash, J. T., Ho, D. N., Rubinstein, J. L., and Frank, J. (2014). Automated particle picking for low-contrast macromolecules in cryo-electron microscopy. *Journal of structural biology*, 186(1):1–7.
- Marino, J., Yue, Y., and Mandt, S. (2018). Iterative amortized inference. *arXiv preprint arXiv:1807.09356*.
- Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004). Ucsf chimera visualization system for exploratory research and analysis. *Journal of computational chemistry*, 25(13):1605–1612.
- Pintilie, G. (2010). Greg pintilie’s homepage: people.csail.mit.edu/gdp/cryoem.html.
- Punjani, A., Rubinstein, J. L., Fleet, D. J., and Brubaker, M. A. (2017). cryosparc: algorithms for rapid unsupervised cryo-em structure determination. *Nature Methods*, 14(3):290–296.
- Rezende, D. J., Eslami, S. A., Mohamed, S., Battaglia, P., Jaderberg, M., and Heess, N. (2016). Unsupervised learning of 3d structure from images. In *Advances in Neural Information Processing Systems*, pages 4996–5004.
- Rezende, D. J., Mohamed, S., and Wierstra, D. (2014a). Stochastic backpropagation and approximate infer-

- ence in deep generative models. *arXiv preprint arXiv:1401.4082*.
- Rezende, D. J., Mohamed, S., and Wierstra, D. (2014b). Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 1278–1286.
- Rosenthal, P. B. and Henderson, R. (2003). Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *Journal of Molecular Biology*, 333(4):721–745.
- Rupp, B. (2009). *Biomolecular crystallography: principles, practice, and application to structural biology*. Garland Science.
- Scheres, S. H. (2012). Relion: implementation of a bayesian approach to cryo-em structure determination. *Journal of structural biology*, 180(3):519–530.
- Scheres, S. H., Gao, H., Valle, M., Herman, G. T., Eggermont, P. P., Frank, J., and Carazo, J.-M. (2007a). Disentangling conformational states of macromolecules in 3d-em through likelihood optimization. *Nature methods*, 4(1):27.
- Scheres, S. H., Núñez-Ramírez, R., Gómez-Llorente, Y., San Martín, C., Eggermont, P. P., and Carazo, J. M. (2007b). Modeling experimental image formation for likelihood-based classification of electron microscopy data. *Structure*, 15(10):1167–1177.
- Sigworth, F. (1998). A maximum-likelihood approach to single-particle image refinement. *Journal of structural biology*, 122(3):328–339.
- Sigworth, F. J. (2016). Principles of cryo-em single-particle image processing. *Microscopy*, 65(1):57–67.
- Tang, G., Peng, L., Baldwin, P. R., Mann, D. S., Jiang, W., Rees, I., and Ludtke, S. J. (2007). Eman2: an extensible image processing suite for electron microscopy. *Journal of structural biology*, 157(1):38–46.
- Unser, M., Trus, B. L., and Steven, A. C. (1987). A new resolution criterion based on spectral signal-to-noise ratios. *Ultramicroscopy*, 23(1):39–51.
- Vilas, J. L., Gómez-Blanco, J., Conesa, P., Melero, R., de la Rosa-Trevín, J. M., Otón, J., Cuenca, J., Marabini, R., Carazo, J. M., Vargas, J., et al. (2018). Monores: automatic and accurate estimation of local resolution for electron microscopy maps. *Structure*, 26(2):337–344.
- Vulović, M., Ravelli, R. B., van Vliet, L. J., Koster, A. J., Lazić, I., Lücken, U., Rullgård, H., Öktem, O., and Rieger, B. (2013). Image formation modeling in cryo-electron microscopy. *Journal of structural biology*, 183(1):19–32.
- Xu, Z., Horwich, A. L., and Sigler, P. B. (1997). The crystal structure of the asymmetric groel-groes-(adp) 7 chaperonin complex. *Nature*, 388(6644):741.
- Zhao, J., Brubaker, M. A., and Rubinstein, J. L. (2013). Tmacs: A hybrid template matching and classification system for partially-automated particle selection. *Journal of structural biology*, 181(3):234–242.

Appendices

A Visual impressions

First, to give a visual impression of the observations, we learn from we present in Figure 7 samples from the 3 data sets we use. All of the datasets are based on the same protein estimate of the GroEL-GroES protein(Xu et al., 1997). They differ in the signal-to-noise ratio (SNR). The left row represents noise free data, the middle a moderate common noise level, and the right an extreme level of noise. For each observation example, we show also the corresponding Fourier transformations, their real part in the second column and their imaginary part in the third column. Further we present the qualitative results of fit-

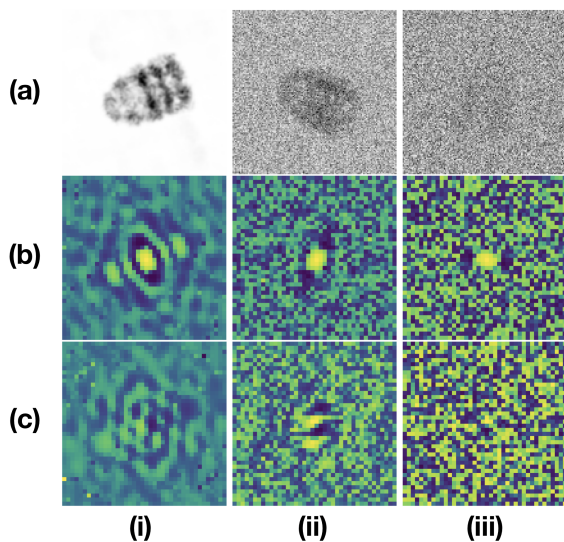


Figure 7: *Left to Right*: We show samples from the dataset we use: (i) no noise (such as in Experiment 6.1), (ii) moderate noise and (iii) high noise (such as in experiment 6.2). *Top to Bottom*: Observation in (a) real space, first 20 Fourier shells (b) real part and (c) imaginary part (for better visibility log-scaled). The latter two are being used for the optimization due to the application of the Fourier slice theorem explained in Section 4.

ting the mid and high level datasets with our method. We visualize the respective protein fit from experiment 6.2 in Figure 8 with the Chimera X software package (Pettersen et al., 2004). The two pictures on the top row represent the middle noise fit, respectively the bottom two the high noise fit.

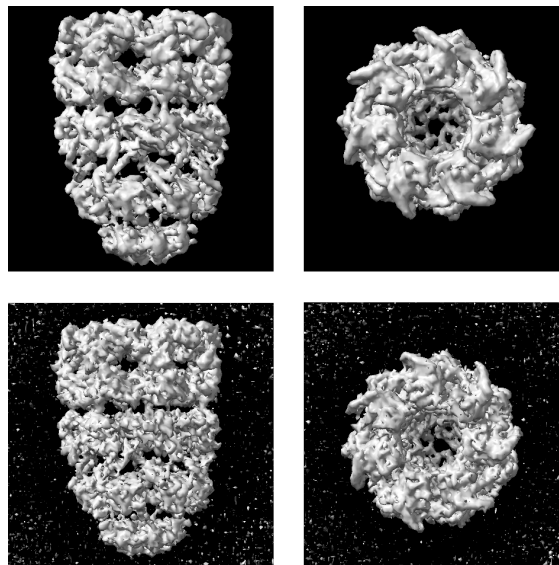


Figure 8: *Top*: Side and top view of the GroEL-GroES protein fit with moderate noise level data. *Bottom*: Side and top view of the respective high noise level dataset.

B Extension to experiment 6.3

We shall execute the same experiment as in section 6.3 given the dataset with intermediate noise. We display the experiments of this result in Figure 9. It is clear that while, missing information leads to large deviation in variance we also find that the noise leads to some variance in the observed area. Again we visualize the result of the fit in Figure 10.

C Amortized inference and variational EM for pose estimation

We have not used amortized inference in our experiments. In experiment 6.4 we have modelled poses as local variables and trained them by variational expectation maximization. Other work shows that the amortization gap can be significant (Scheres et al., 2007b; Alemi et al., 2017; Marino et al., 2018). Hence in order to exclude the gap as a reason for failure, we decided to model local variables. We did run experiments though with ResNets as encoders without success either. We believe the core problem in probabilistic pose estimation is the number of local optima. This makes simple SGD a somewhat poor choice, because we rely on finding the global minimum.

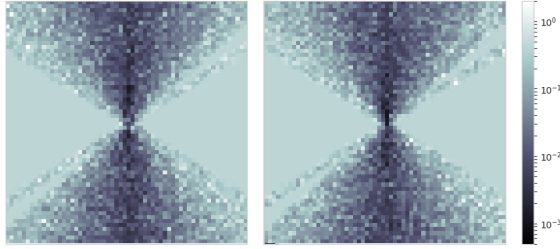


Figure 9: Center slice through the learned Fourier volume uncertainties σ_v . *Left*: real part, *Right*: imaginary part. We learn the model fit with observations coming only from a 30° cone, a scenario similar to breast cancer scans where observations are available only from some viewing directions. Uncertainty close to 1 means that the model has no information in these areas, close to zero represents areas of high sampling density. In contrast to other models, our model can identify precisely where information is missing (high variance).

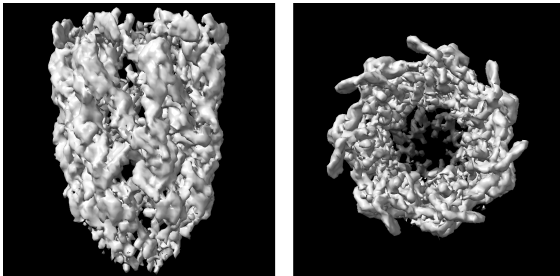


Figure 10: Side and top view of the GroEL-GroES protein fit with moderate noise level data and all observations stemming from a limited pose space.

D Remarks to the chosen observation model

The Gaussian noise model is a common but surely oversimplified model (Sigworth, 1998). A better model would be the Poisson distribution. The Gaussian is a good approximation to it given a high rate parameter meaning if there is a reasonable high count of radiation hitting the sensors. This is a good assumption for most methods of the field, but can actually be a poor model in some cases of cryo electron microscopy. An example of an elaborate model is presented in Vulović et al. (2013).