Building Proteins in a Day: Efficient 3D Molecular Structure Estimation with Electron Cryomicroscopy

Ali Punjani, *Member, IEEE,* Marcus A. Brubaker, *Member, IEEE,* and David J. Fleet *Member, IEEE,*

Abstract—Discovering the 3D atomic-resolution structure of molecules such as proteins and viruses is one of the foremost research problems in biology and medicine. Electron Cryomicroscopy (cryo-EM) is a promising vision-based technique for structure estimation which attempts to reconstruct 3D atomic structures from a large set of 2D transmission electron microscope images. This paper presents a new Bayesian framework for cryo-EM structure estimation that builds on modern stochastic optimization techniques to allow one to scale to very large datasets. We also introduce a novel Monte-Carlo technique that reduces the cost of evaluating the objective function during optimization by over five orders of magnitude. The net result is an approach capable of estimating 3D molecular structure from large-scale datasets in about a day on a single CPU workstation.

1 Introduction

Discovering the 3D atomic-resolution structure of molecules such as proteins and viruses is a fundamental open problem in biology and medicine. Without exaggeration, the ability to routinely determine the 3D structure of such molecules would likely revolutionize the process of drug development, and accelerate research into key biological processes. Electron Cryomicroscopy (cryo-EM) is an emerging, vision-based approach to 3D macromolecular structure determination that is applicable to medium to large-sized molecules in their native state [10]. This is in contrast to X-ray crystallography which requires a crystal of the target molecule that are often impossible to grow [39], or nuclear magnetic resonance (NMR) spectroscopy, which is limited to relatively small molecules [20].

The cryo-EM reconstruction task is to estimate the 3D density of a target molecule from a large set of images of the molecule (called particle images), obtained with a transmission electron microscope. The problem is similar in spirit to multi-view scene carving [8, 22] and to large-scale, uncalibrated, multi-view reconstruction [2]. Like multi-view scene carving, the goal is to estimate a dense 3D occupancy representation of shape from a set of different views. But unlike many approaches to scene carving, we do not assume calibrated cameras, since the 3D poses of the molecule in different images are unknown. Like uncalibrated, multi-view reconstruction, we aim to use large numbers of uncalibrated views to obtain high fidelity 3D reconstructions. But the low signal-to-noise levels in cryo-EM particle images (often as low as 0.05 [6]; see Fig. 1) and the different image formation model prevent the use of common feature matching techniques to establish correspondences. Computed Tomography (CT) [15, 18] uses a similar imaging model (orthographic integral projection), however in CT the projection direction of each image is known, whereas with cryo-EM the 3D pose for each particle image is unknown.

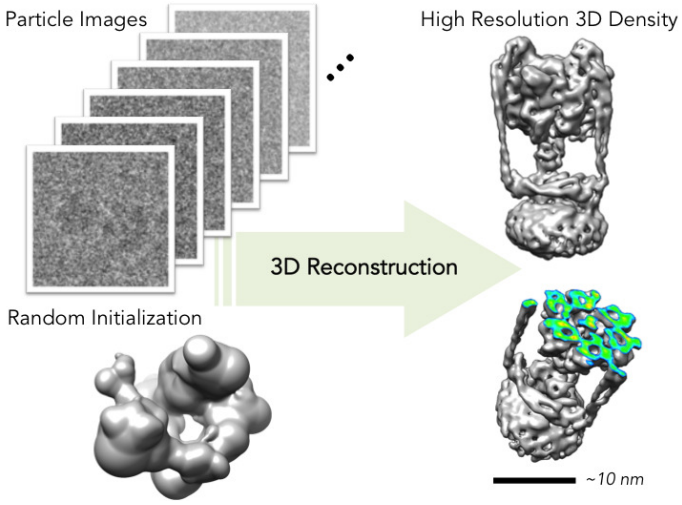


Fig. 1: The goal is to reconstruct the 3D structure of a molecule (right), at sub-nanometer scales, from a large number of noisy, uncalibrated 2D projections obtained from cryogenically frozen samples in an electron microscope (left).

Existing cryo-EM techniques (*e.g.*, [11, 16, 42, 45]) suffer from two key problems. First, without good initialization, they converge to poor or incorrect solutions [17], often with little indication that something went wrong. Second, they are extremely slow, which limits the number of particles images one can use to mitigate the effects of noise; *e.g.*, the website of the RELION package [1, 42] reports requiring two weeks on 200 cores to process a dataset with 100,000 images. This

• *Financial support was provided to DJF from NSERC Canada and the CIFAR Learning in Machines and Brains Program. MAB was funded in part by an NSERC Postdoctoral Fellowship. The authors would like thank John L. Rubinstein for providing data and invaluable feedback.*

problem is likely to worsen as high-throughput data collection becomes more common, providing larger datasets, and better hardware enables higher resolution images.

We introduce a framework for cryo-EM density estimation, formulating the problem as one of stochastic optimization to perform maximum-a-posteriori (MAP) estimation in a probabilistic model. The approach is efficient, providing useful low resolution density estimates in just a few hours. We also show that our stochastic optimization technique is insensitive to initialization, allowing the use of random initializations. We further introduce a novel importance sampling scheme that dramatically reduces the computational costs associated with high resolution reconstruction. This leads to speedups of 100,000-fold or more, allowing structures to be determined in a day on a modern CPU workstation. In addition, the proposed framework is flexible, allowing parts of the model to be changed and improved without impacting the overall framework; *e.g.*, we compare the use of three different priors. To demonstrate the effectiveness of the method, we perform reconstructions on two real datasets and one synthetic dataset. A preliminary version of this work appeared in [9].

2 Background and Related Work

Biological processes occur as the result of binding and chemical interactions between molecules inside cells. The majority of these molecules are protein structures, constructed from 20 different amino acid monomer building blocks. Each different type of protein, coded in DNA, is a unique sequence of these monomers joined into a chain. These chains fold into 3D shapes during construction and it is this final 3D structure that determines the function of the protein. Because function depends on structure, discovering the structures of proteins and other macromolecules is fundamental to studying and understanding biological processes. It is also an important part of discovering drugs that can inhibit or accelerate the action of specific proteins involved in disease pathways.

Electron Cryomicroscopy (cryo-EM) provides a way to determine this critical 3D atomic-resolution structural information for many proteins and other macromolecules. In cryo-EM, a purified solution of the target molecule is cryogenically frozen into a thin (single molecule thick) film, and then imaged with a transmission electron microscope. A large number of such samples are obtained, each of which provides a micrograph containing hundreds of visible, individual molecules. In a process known as *particle picking*, individual molecules are detected, resulting in a stack of cropped *particle images*. Particle picking is often done manually, however there have been recent attempts to partially or fully automate the process [23, 48]. Each particle image provides a noisy view of the molecule, but with unknown pose relative to the molecule, see Fig. 2 (right). The reconstruction task is to estimate the 3D electron density of the target molecule from the potentially large set of particle images.

Traditional approaches to cryo-EM density estimation use a form of iterative refinement (*e.g.*, [11, 16, 45]). Based on an initial estimate of the 3D density, they determine the best matching alignment (*i.e.*, 3D pose and image position) for each particle image. A new density estimate is then constructed using the Fourier Slice Theorem, much like Computed Tomography [18]. In effect, using the 3D pose and position of the particle in each particle image, the new density is found through interpolation and averaging of the observed particle images, often performed in the Fourier domain.

This approach is fundamentally limited in several ways. Even if one knew the correct 3D molecular density, the very low SNR makes it difficult to accurately identify the correct pose and position for each particle image. This problem is exacerbated when the initial density is inaccurate. As a consequence, poor initializations result in estimated structures that are either clearly wrong (see Fig. 11) or, worse, appear plausible but are misleading in reality, yielding incorrect 3D structures [17]. Finally, and crucially for the case of density estimation with many particle images, all data are used at each refinement iteration, causing these methods to be extremely slow. Mallick et al. [30] proposed an approach which attempted to establish weak constraints on the relative 3D poses between different particle images. This was used to initialize an iterative refinement algorithm. In contrast to such approaches, ours does not require accurate initialization.

To avoid the need to estimate a single 3D pose and position for each particle image, Bayesian approaches have been proposed in which pose and position for each particle image are treated as latent variables, and then marginalized numerically. This approach was originally proposed by Sigworth [43] for 2D image alignment and later by Scheres et al. [41] for alignment comprising 3D orientation and 2D image position. It has since been used by Jaitly et al. [19], where batch, gradient-based optimization was performed. Nevertheless, due to the computational cost of marginalization, the method was only applied to small numbers of class-average images obtained by clustering, aligning and averaging individual particle images according to their 2D appearance, to reduce noise and the number of particle images used during the optimization.

The state-of-the-art RELION package [42] uses pose marginalization and a batch Expectation-Maximization algorithm for density estimation. This approach is, however, computationally expensive. The website for the RELION software reports that reconstruction from a dataset of 100,000 particle images typically take two weeks on 200 cores [1]. We advocate the use of a similar marginalized likelihood, but with stochastic rather than batch optimization. This allows for more efficient optimization, and for robustness to initialization. We further introduce a novel importance sampling technique that dramatically reduces the computational cost of the marginalization when working at higher resolutions.

3 A Framework for 3D Density Estimation

Our formulation of the cryo-EM reconstruction problem comprises a probabilistic generative model of image formation, given the 3D electron density of the molecule, where the 3D pose and 2D position of the particle in each image are treated as unknown, latent variables. Stochastic optimization is employed to cope with large-scale datasets, and importance

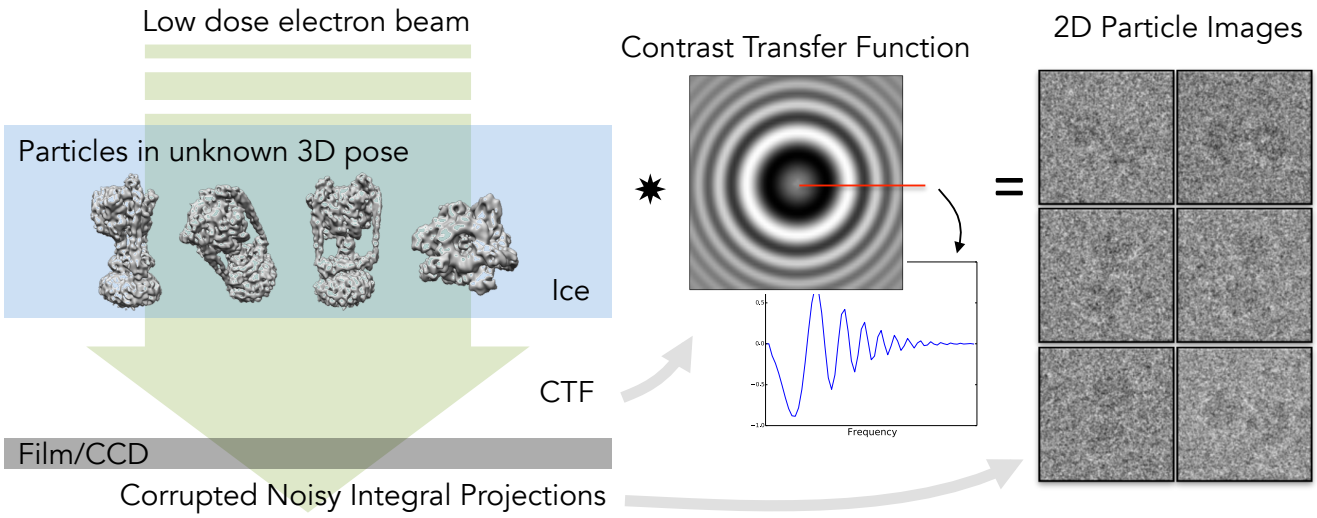


Fig. 2: A generative image formation model in cryo-EM. The electron beam results in an orthographic integral projection of the electron density of the specimen. This projection is modulated by the Contrast Transfer Function (CTF) and corrupted with noise. The images pictured here showcase the low SNR typical in cryo-EM. The zeros in the CTF (which completely destroy some spatial information) make estimation particularly challenging, however their locations vary as a function of microscope parameters. These are set differently across particle images in order to mitigate this problem. Particle images and density from [24].

sampling is used to efficiently marginalize over the unknown pose and position for each particle image.

3.1 Generative Model

This section introduces the key elements of a generative model of image formation for cryo-EM. Details of the formulation are provided in the appendix.

A cryo-EM particle image is assumed to be an orthogonal, integral projection of the 3D target density, V. The 3D pose of the particle in the image, $\mathbf{R} \in \mathcal{SO}(3)$, is unknown a priori, as is the 2D position of the particle within the image, $\mathbf{t} \in \mathbb{R}^2$. The projection is corrupted by blur and other microscope aberrations, analogous to the effects of defocus in a conventional light camera. Such distortions are characterized by modulation with a contrast transfer function (CTF) in the Fourier domain, denoted $\hat{c}(\boldsymbol{\omega}; \theta)$ where $\boldsymbol{\omega} \equiv (\omega_1, \omega_2)^{\mathsf{T}}$ represents 2D Fourier coordinates, and θ denotes the CTF parameters. Figure 2 depicts a typical CTF, with periodic phase reversals that imply, among other things, that particle images are not strictly positive. Finally the image is corrupted with additive noise, clearly visible in Fig. 2. Such large amounts of noise are primarily due to very low exposures, necessitated by the sensitive nature of biological specimens. The noise is modelled as IID and Gaussian.

In cryo-EM, largely for computational convenience, it is common to express the generative model in the Fourier domain where the key elements have a particularly simple form. The effect of the CTF reduces to modulation in the Fourier domain. Image translation corresponds to a simple phase shift in the Fourier domain. And the Fourier Slice Theorem specifies that the Fourier transform of the integral projection of a rotated object yields a slice through the 3D spectrum of the object, in a plane normal to the projection direction. As formulated in the appendix, the generative model in the Fourier domain is given by:

$$\hat{I}(\boldsymbol{\omega}) \;=\; \hat{c}(\boldsymbol{\omega}; \theta)\, e^{-\iota 2\pi\, \boldsymbol{\omega}^{\mathsf{T}} \mathbf{t}}\, \hat{V}(\omega_1 \mathbf{n}_1 + \omega_2 \mathbf{n}_2) \;+\; \hat{\nu}(\boldsymbol{\omega}) \quad (1)$$

where $\hat{I}(\boldsymbol{\omega})$ denotes 2D Fourier spectrum of the image, $e^{-\iota 2\pi\, \boldsymbol{\omega}^{\mathsf{T}} \mathbf{t}}$ is the phase shift induced by the image translation (with $\iota^2 = -1$), $\hat{V}$ is the 3D Fourier spectrum of the target density V, and where $\mathbf{n}_1$ and $\mathbf{n}_2$ are orthogonal unit vectors that span the plane normal to the viewing direction $\mathbf{d}$. These vectors are given by the rows of the rotation matrix for the 3D pose, *i.e.*, $\mathbf{R} = [\mathbf{n}_1, \mathbf{n}_2, \mathbf{d}]^{\mathsf{T}}$. Finally, because the Fourier transform is a linear, unitary transform, the additive, Gaussian image noise remains Gaussian and white in the Fourier domain up to Hermitian symmetry. Given IID mean-zero Gaussian noise with variance σ^2, the noise in the Fourier coefficients is mean-zero, independent, and complex normal with variance $2\sigma^2$, written $\hat{\nu}(\boldsymbol{\omega}) \sim \mathcal{CN}(0, 2\sigma^2)$.

One key benefit of expressing the generative model and likelihood in the Fourier domain is the ease with which we can formulate the estimation of low-resolution structures. This is accomplished straightforwardly by only considering the low-order Fourier terms of the observed images. Computational savings stem from the fact that the number of Fourier coefficients increases quadratically with frequency. The low frequency coefficients also have higher SNR, due in part to CTF attenuation of higher frequencies.

3.2 Marginalized Likelihood

In practice, we discretize the target density V into 3D grid, $\mathcal{V}$, with density represented at each of D^3 voxels, the discrete Fourier transform (DFT) of which is denoted $\hat{\mathcal{V}}$. Similarly, the

observed particle images have D^2 pixels, with 2D DFT, $\hat{\mathcal{I}}$, defined on a discrete set of frequencies, $\boldsymbol{\omega} \in \Omega$. Given the white, complex normal noise model, the joint likelihood over the image Fourier coefficients up to a maximum radius in the frequency domain, ω_*, is given by

$$p(\hat{\mathcal{I}} \,|\, \theta, \mathbf{R}, \mathbf{t}, \mathcal{V}, \sigma^2) \propto \prod_{\substack{\boldsymbol{\omega}\in\Omega \\ ||\boldsymbol{\omega}|| < \omega_*}} \mathcal{CN}(\hat{\mathcal{I}}[\boldsymbol{\omega}]; \mu(\boldsymbol{\omega}), 2\sigma^2) \qquad (2)$$

where $\hat{\mathcal{I}}[\boldsymbol{\omega}]$ denotes the image Fourier coefficient at frequency $\boldsymbol{\omega}$, and $\mu(\boldsymbol{\omega}) = \hat{c}(\boldsymbol{\omega};\, \theta)\, e^{-\iota 2\pi \boldsymbol{\omega}^{\mathrm{T}}\mathbf{t}}\, \hat{\mathcal{V}}(\omega_1\mathbf{n}_1 + \omega_2\mathbf{n}_2)$. Computing the mean requires interpolation of the discrete 3D Fourier coefficients and we use trilinear interpolation with premultiplication. See the appendix and [31] for more details.

The 3D pose, $\mathbf{R} \in \mathcal{SO}(3)$, and shift, $\mathbf{t} \in \mathbb{R}^2$, of each particle image are unknown and treated as latent variables that are marginalized [41, 43]. Assuming $\mathbf{R}$ and $\mathbf{t}$ are independent of each other and the density $\mathcal{V}$, one obtains

$$p(\hat{\mathcal{I}} \,|\, \theta, \mathcal{V}, \sigma^2) = \int_{\mathbb{R}^2} \int_{\mathcal{SO}(3)} p(\hat{\mathcal{I}} \,|\, \theta, \mathbf{R}, \mathbf{t}, \mathcal{V}, \sigma^2)\, p(\mathbf{R})\, p(\mathbf{t})\, d\mathbf{R}\, d\mathbf{t} \qquad (3)$$

where $p(\mathbf{R})$ is a prior over 3D poses, and $p(\mathbf{t})$ is a prior over translations. In general, nothing is known about the projection direction so $p(\mathbf{R})$ is assumed to be uniform. Particles are picked to be close to the center of each image, so $p(\mathbf{t})$ is chosen to be a broad Gaussian distribution centered in the image.

The double integral in Eq. (3) is not analytically tractable, so numerical quadrature is used. To perform quadrature over 3D pose, it is convenient to use the subgroup structure of $\mathcal{SO}(3)$ and parameterize the rotation matrix $\mathbf{R}$ in terms of the viewing direction $\mathbf{d} \in \mathcal{S}$ and an in-plane rotation angle $\psi \in [0, 2\pi)$. To generate a valid quadrature scheme over $\mathcal{SO}(3)$, it suffices to combine quadrature schemes over $\mathcal{S}$ and $[0, 2\pi)$ [14, 28]. The conditional probability of an image (*i.e.*, the likelihood) $p(\hat{\mathcal{I}} \,|\, \theta, \mathcal{V}, \sigma^2)$ is then approximated as a weighted sum

$$\sum_{j=1}^{M_{\mathbf{d}}} w_j^{\mathbf{d}} \sum_{k=1}^{M_\psi} w_k^{\psi} \sum_{\ell=1}^{M_{\mathbf{t}}} w_\ell^{\mathbf{t}}\, p(\hat{\mathcal{I}}|\theta, \mathbf{R}_{j,k}, \mathbf{t}_\ell, \mathcal{V}, \sigma^2)\, p(\mathbf{R}_{j,k}) p(\mathbf{t}_\ell) \qquad (4)$$

where $\mathbf{R}_{j,k} \equiv \mathbf{R}(\mathbf{d}_j, \psi_k)$ is the rotation matrix with viewing direction $\mathbf{d}_j$ and inplane rotation ψ_k. Further, $\{(\mathbf{d}_j, w_j^{\mathbf{d}})\}_{j=1}^{M_{\mathbf{d}}}$ are weighted quadrature points over $\mathcal{S}$, $\{(\psi_k, w_k^{\psi})\}_{k=1}^{M_\psi}$ are weighted quadrature points over $[0, 2\pi)$ and $\{(\mathbf{t}_\ell, w_\ell^{\mathbf{t}})\}_{\ell=1}^{M_{\mathbf{t}}}$ are weighted quadrature points over $\mathbb{R}^2$.

To generate quadrature points over $\mathcal{S}$ we use the approach described in [40], which generates a requested number of points $M_{\mathbf{d}}$ which (approximately) uniformly cover $\mathcal{S}$ and then use $w^{\mathbf{d}} = \frac{4\pi}{M_{\mathbf{d}}}$. Quadrature points over $[0, 2\pi)$ are generated uniformly, so $w^{\psi} = \frac{2\pi}{M_\psi}$. Finally, due to the Gaussian form of $p(\mathbf{t})$, Gauss-Hermite quadrature over $\mathbb{R}^2$ is used where the weights $w^{\mathbf{t}}$ are determined by the quadrature scheme.

The accuracy of the quadrature scheme, and consequently the values of $M_{\mathbf{d}}$, M_ψ and $M_{\mathbf{t}}$, are set based on ω_*, the specified maximum frequency considered. The use of higher frequencies requires finer quadrature sampling. Specifically, if $\alpha = \frac{1}{2} \arccos \frac{(D\omega_*)^2 - 1}{(D\omega_*)^2 + 1}$ is the angle between discrete wave numbers in Fourier space at a radius of ω_*, then we seek a quadrature scheme which has a maximum angular spacing of approximately α. To achieve this we use $M_{\mathbf{d}} = (\frac{3.6}{\alpha})^2$ [40] and $M_\psi = \frac{2\pi}{\alpha}$. For quadrature in the plane, we use $M_{\mathbf{t}} = (D\omega_*)^2$ which corresponds to spacing quadrature points at one full period at frequency ω_*.

3.3 Stochastic Optimization

Given a set of K particle images, each with CTF parameters and noise levels, $\mathfrak{D} = \{(\hat{\mathcal{I}}_i, \theta_i, \sigma_i^2)\}_{i=1}^K$, and assuming conditional independence of the images, the posterior probability of a density $\mathcal{V}$ is

$$p(\mathcal{V}|\mathfrak{D}) \propto p(\mathcal{V}) \prod_{i=1}^{K} p(\hat{\mathcal{I}}_i \,|\, \theta_i, \mathcal{V}, \sigma_i^2)\ , \qquad (5)$$

where $p(\mathcal{V})$ is a prior over 3D molecular densities. Several choices of prior are explored below, but we found that a simple independent exponential prior worked well. Specifically, $p(\mathcal{V}) = \prod_{i=1}^{D^3} \lambda e^{-\lambda \mathcal{V}_i}$ where $\mathcal{V}_i$ is the density of the ith voxel and λ is the inverse scale parameter. Other choices of prior are possible and we explore this later.

Estimating the density corresponds to finding $\mathcal{V}$ that maximizes the posterior in Eq. (5). In doing so, we also constrain the density at each voxel to be positive, as negative density is physically unrealistic. Them taking the negative log and dropping constant factors, the optimization problem becomes $\arg\min_{\mathcal{V}\in\mathbb{R}_+^{D^3}} f(\mathcal{V})$,

$$f(\mathcal{V}) = -\log p(\mathcal{V}) - \sum_{i=1}^{K} \log p(\hat{\mathcal{I}}_i \,|\, \theta_i, \mathcal{V}, \sigma_i^2)\ . \qquad (6)$$

Optimizing Eq. (6) directly is costly due to the marginalization in Eq. (4) as well as the large number (K) of particle images in a typical dataset. To address these challenges the following sections describe our use of stochastic optimization and importance sampling.

3.3.1 Stochastic Average Gradients

To efficiently cope with the large number of particle images in a typical dataset, we advocate the use of stochastic optimization. Stochastic optimization methods exploit the large amount of redundancy in most datasets by only considering subsets of data (*i.e.*, images) at each iteration. There are many such methods, ranging from simple stochastic gradient descent with momentum [34, 35, 44] to more complex methods such as Natural Gradient methods [4, 5, 25, 26] and Hessian-free optimization [32].

We have explored many of these methods for 3D reconstruction [36]. However, here we use Stochastic Average Gradient Descent (SAGD) [27] which has several important properties. First, it is effectively self-tuning, using a line-search to determine and adapt the learning rate. This is particularly important, as many methods require significant manual tuning for new objective functions, or datasets. Further, it is specifically designed for the finite dataset case allowing for faster convergence. Finally, SAGD explicitly produces an

estimate of the full gradient over the entire dataset, providing a natural way to assess convergence.

Our goal, as stated above is to minimize the negative log posterior in Eq. (6), *i.e.*,

$$\begin{aligned} f(\mathcal{V}) &= -\log p(\mathcal{V}) - \sum_{i=1}^{K} \log p(\hat{\mathcal{I}}_i|\theta_i, \mathcal{V}, \sigma_i^2) \\ &= \sum_{i=1}^{K} \left[-\frac{1}{K} \log p(\mathcal{V}) - \log p(\hat{\mathcal{I}}_i|\theta_i, \mathcal{V}, \sigma_i^2) \right] \\ &= \sum_{i=1}^{K} f_i(\mathcal{V}) \ . \end{aligned} \tag{7}$$

At each iteration τ, SAGD [27] selects a random particle image, indexed by i_τ, the corresponding objective for which is the log likelihood, denoted, $f_{i_\tau}(\mathcal{V})$. Also, let the gradient of the objective with respect to the 3D density be denoted $\mathbf{g}_{i_\tau}(\mathcal{V}) \equiv \nabla_{\mathcal{V}} f_{i_\tau}(\mathcal{V})$. As explained in the appendix, the objective is continuous and straightforwardly differentiable.

SAGD then computes an update given by

$$\mathcal{V}_{\tau+1} = \mathcal{V}_\tau - \frac{\epsilon}{L} \sum_{i=1}^{K} \left[d\mathcal{V}_i^\tau - \frac{1}{K} \frac{\partial}{\partial \mathcal{V}} \log p(\mathcal{V}) \right] \ , \tag{8}$$

where ϵ is a base learning rate, L is a Lipschitz constant of the gradient $\mathbf{g}_{k_\tau}(\mathcal{V})$ and

$$d\mathcal{V}_i^\tau = \begin{cases} \mathbf{g}_{i_\tau}(\mathcal{V}_\tau) & i = i_\tau \\ d\mathcal{V}_i^{\tau-1} & \text{otherwise} \end{cases} \tag{9}$$

That is, at iteration τ we only compute the gradient for data point i_τ, but the gradient update at iteration τ also uses the most recently computed gradients for all other data points.

In practice, the sum in Eq. (8) is not computed at each iteration. Rather, a running total is maintained and updated as follows:

$$\begin{aligned} \bar{\mathbf{g}}_\tau &= \sum_{i=1}^{K} d\mathcal{V}_i^\tau \\ \bar{\mathbf{g}}_{\tau+1} &= \bar{\mathbf{g}}_\tau - d\mathcal{V}_{i_\tau}^\tau + \mathbf{g}_{i_\tau}(\mathcal{V}_\tau) \end{aligned}$$

Altogether, this allows for SAGD to take many steps and make fast progress before a batch optimization algorithm would be able to take even a single step. Further, rather than selecting a single data point at each iteration, we select a subset of data points (minibatches) and compute the gradient for the sum of the objective f_i over the entire minibatch. This allows for computational parallelization and helps to reduce the memory required by SAGD. For the experiments below, we used a minibatch size of 200 particle images.

3.3.2 SAGD Parameters

The SAGD algorithm requires a Lipschitz constant L which is not generally known. Instead it is estimated using a line search algorithm where an initial value of L is increased until the instantiated Lipschitz condition $f(\mathcal{V}) - f(\mathcal{V} - L^{-1}d\mathcal{V}) < \frac{\|d\mathcal{V}\|^2}{2L}$ is met. The line search for the Lipschitz constant L is only performed once every 20 iterations. More sophisticated line search could be performed if desired. A good initial value of L is found using a bisection search where the upper bound is the smallest L found so far to satisfy the condition and the lower bound is the largest L found so far which fails the condition. In between line searches, L is gradually decreased to try to take larger steps as described in [27].

Algorithm 1 SAGD

```
Initialize V and L
Initialize ḡ ← 0
Initialize dV_k ← 0 for all k = 1..K
for τ = 1..τ_max do
    Select data subset k_τ
    Compute objective gradient g_{k_τ}(V)
    ḡ ← ḡ − dV_{k_τ} + g_{k_τ}(V)
    dV_{k_τ} ← g_{k_τ}(V)
    V ← V − (ε/L) [ḡ − ∂/∂V log p(V)]
    if mod(τ,20) == 0 then
        Perform line search
        while f_{k_τ}(V) − f_{k_τ}(V − L^{-1} dV_{k_τ}) < ||dV_{k_τ}||^2 / 2L do
            L ← 2L
    else
        L ← 2^{−1/150} L
```

Like other stochastic optimization algorithms, convergence of SAGD is only guaranteed for convex functions. In practice, while the objective function defined here is non-convex, we find good performance regardless, consistent with myriad other applications of stochastic optimization. In the convex case, convergence is only assured for values of $\epsilon \leq \frac{1}{16}$ [27]. However we found larger values at early iterations to be beneficial, and consequently use $\epsilon = \max(\frac{1}{16}, 2^{1-\lfloor \tau/150 \rfloor})$. Finally, to enforce the positivity of density, negative values of $\mathcal{V}$ are truncated to zero after each iteration. A summary of the entire SAGD algorithm is provided in Algorithm (1).

3.4 Importance Sampling

While stochastic optimization allows us to scale to large numbers of images, the cost of computing the required gradient for each image remains high due to marginalization over 3D orientations and 2D shifts in Eq. (4). Intuitively, one could consider randomly selecting a subset of the terms in Eq. (4) and using this as an approximation. This idea is formalized by importance sampling (IS) which allows for an efficient and accurate approximation of the discrete sums in Eq. (4).[1] (For a review of importance sampling, see [46].)

To formulate importance sampling, we first re-express the inner sum from Eq. (4) as follows

$$\phi_{j,k}^{\mathbf{d},\psi} = \sum_{\ell=1}^{M_{\mathbf{t}}} w_\ell^{\mathbf{t}} \, p_{j,k,\ell} = \sum_{\ell=1}^{M_{\mathbf{t}}} q_\ell^{\mathbf{t}} \left(\frac{w_\ell^{\mathbf{t}} p_{j,k,\ell}}{q_\ell^{\mathbf{t}}} \right) \tag{10}$$

where $p_{j,k,\ell} = p(\hat{\mathcal{I}}|\theta, \mathbf{R}_{j,k}, \mathbf{t}_\ell, \hat{\mathcal{V}}) \, p(\mathbf{R}_{j,k}) \, p(\mathbf{t}_\ell)$ and $\mathbf{q}^{\mathbf{t}} \equiv (q_1^{\mathbf{t}}, \ldots, q_{M_{\mathbf{t}}}^{\mathbf{t}})^T$ is the parameter vector of a multinomial importance distribution such that $\sum_{\ell=1}^{M_{\mathbf{t}}} q_\ell^{\mathbf{t}} = 1$ and $q_\ell^{\mathbf{t}} > 0$. The domain of $\mathbf{q}^{\mathbf{t}}$ corresponds to the set of quadrature points in Eq. (4). Then, $\phi_{j,k}^{\mathbf{d},\psi}$ can be thought of as the expected

1. One can also apply importance sampling directly to the continuous integrals in Eq. (3) but it can be computationally advantageous to have a fixed set of projections and shifts which can be reused across particle images.

value $E_\ell[\frac{w_\ell^{\mathbf{t}} p_{j,k,\ell}}{q_\ell^{\mathbf{t}}}]$, where ℓ is a random variable distributed according to $\mathbf{q}^{\mathbf{t}}$. If a set of $N_{\mathbf{t}} \ll M_{\mathbf{t}}$ random indexes, $\mathfrak{I}^{\mathbf{t}}$, are drawn according to $\mathbf{q}^{\mathbf{t}}$, then

$$\phi_{j,k}^{\mathbf{d},\psi} \approx \frac{1}{N_{\mathbf{t}}} \sum_{\ell \in \mathfrak{I}^{\mathbf{t}}} \frac{w_\ell^{\mathbf{t}} p_{j,\ell}}{q_\ell^{\mathbf{t}}} \ . \tag{11}$$

Thus, we can approximate $\phi_{j,k}^{\mathbf{d},\psi}$ by drawing samples according to the importance distribution $\mathbf{q}^{\mathbf{t}}$ and computing the average.

Using this approximation, Eq. (4) becomes

$$p(\hat{\mathcal{I}}|\theta,\hat{\mathcal{V}}) \approx \sum_{j=1}^{M_{\mathbf{d}}} w_j^{\mathbf{d}} \sum_{k=1}^{M_\psi} w_k^\psi \frac{1}{N_{\mathbf{t}}} \left(\sum_{\ell \in \mathfrak{I}^{\mathbf{t}}} \frac{w_\ell^{\mathbf{t}} p_{j,k,\ell}}{q_\ell^{\mathbf{t}}} \right) \ . \tag{12}$$

Importance sampling can be similarly used for the other summations:

$$p(\hat{\mathcal{I}}|\theta,\hat{\mathcal{V}}) \approx \sum_{j \in \mathfrak{I}^{\mathbf{d}}} \sum_{k \in \mathfrak{I}^\psi} \sum_{\ell \in \mathfrak{I}^{\mathbf{t}}} \frac{w_j^{\mathbf{d}} w_k^\psi w_\ell^{\mathbf{t}}}{N_{\mathbf{d}} N_\psi N_{\mathbf{t}} q_j^{\mathbf{d}} q_k^\psi q_\ell^{\mathbf{t}}} p_{j,k,\ell} \tag{13}$$

where $\mathfrak{I}^{\mathbf{d}}$ and $\mathfrak{I}^\psi$ are samples drawn from the importance distributions $\mathbf{q}^{\mathbf{d}} = (q_1^{\mathbf{d}}, \ldots, q_{M_{\mathbf{d}}}^{\mathbf{d}})^T$ and $\mathbf{q}^\psi = (q_1^\psi, \ldots, q_{M_\psi}^\psi)^T$.

The accuracy of the approximation in Eqs. (13) to (4) is determined in part by the number of samples used, with the error going to zero as the number of samples increases. For $N_{\mathbf{d}}$ we use $s_0 s(\mathbf{q}_{\mathbf{d}})$ samples where $s(\mathbf{q}) = \left(\sum_\ell q_\ell^2\right)^{-1}$ is the effective sample size [12] and s_0 is a scaling factor. This choice ensures that when the importance distribution is diffuse, more samples are used.

3.4.1 Importance Distribution

As long as the importance distributions (*i.e.*, $\mathbf{q}^{\mathbf{t}}$, $\mathbf{q}^{\mathbf{d}}$ and $\mathbf{q}^\psi$) are non-zero over their respective domains, the resulting weighted samples are properly weighted and the estimates provided by IS are unbiased. Nevertheless, their estimator variance can arbitrarily poor if the importance distributions are not well chosen. In what follows we explain our choice of importance distributions, exploiting the iterative nature of the of the algorithm, while ensuring properly weighted samples.

To begin, note that one natural choice for effective importance distributions would be based on the marginal sums

$$\phi_j^{\mathbf{d}} = \sum_{k=1}^{M_\psi} \sum_{\ell=1}^{M_{\mathbf{t}}} w_k^\psi w_\ell^{\mathbf{t}} p_{j,k,\ell} \tag{14}$$

$$\phi_k^\psi = \sum_{j=1}^{M_{\mathbf{d}}} \sum_{\ell=1}^{M_{\mathbf{t}}} w_j^{\mathbf{d}} w_\ell^{\mathbf{t}} p_{j,k,\ell} \tag{15}$$

$$\phi_\ell^{\mathbf{t}} = \sum_{j=1}^{M_{\mathbf{d}}} \sum_{k=1}^{M_\psi} w_j^{\mathbf{d}} w_k^\psi p_{j,k,\ell} \ . \tag{16}$$

These sums essentially tell us how significant a particular viewing direction, inplane rotation, or inplane shift is for the quadrature, and thus how strongly we should sample it. Unfortunately computing these sums requires as much work as computing Eq. (4) directly.

To avoid such expense we make two observations. First, once the rough shape of the structure has been determined, the marginal sums do not change dramatically from iteration to iteration of stochastic optimization. Intuitively, once a structure has been coarsely determined, the correct pose and shift for a particular image will not change much as the 3D structure is further updated. This is evident in Fig. 3, where, by the third epoch the KL divergence of $\phi^{\mathbf{d}}$ from one epoch to the next is extremely small. This suggests that ϕ from the previous epoch may be useful in constructing the importance distribution at the current epoch. Second, these distributions can also be approximated by importance sampling; *i.e.*,

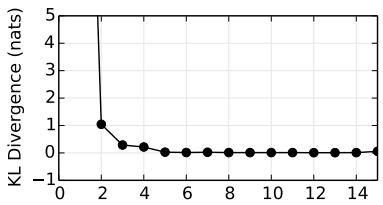


Fig. 3: The KL divergence between the values of $\phi^{\mathbf{d}}$ at the current and previous epochs on the thermus dataset. As optimization progresses, the coarse structure of the molecule is quickly determined, and the KL divergence becomes very small by the third epoch. This indicates that the significantly likely poses for each image have stabilized, and so importance sampling can focus quadrature on these regions preferentially, providing significant speedups.

$$\tilde{\phi}_j^{\mathbf{d}} = \sum_{k \in \mathfrak{I}^\psi} \sum_{\ell \in \mathfrak{I}^{\mathbf{t}}} \frac{w_k^\psi w_\ell^{\mathbf{t}} p_{j,k,\ell}}{N_\psi N_{\mathbf{t}} q_k^\psi q_\ell^{\mathbf{t}}} \tag{17}$$

$$\tilde{\phi}_k^\psi = \sum_{j \in \mathfrak{I}^{\mathbf{d}}} \sum_{\ell \in \mathfrak{I}^{\mathbf{t}}} \frac{w_j^{\mathbf{d}} w_\ell^{\mathbf{t}} p_{j,k,\ell}}{N_{\mathbf{d}} N_{\mathbf{t}} q_j^{\mathbf{d}} q_\ell^{\mathbf{t}}} \tag{18}$$

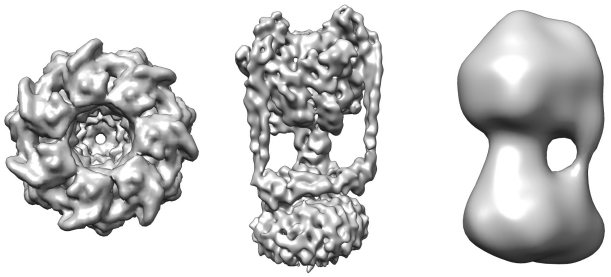
$$\tilde{\phi}_\ell^{\mathbf{t}} = \sum_{j \in \mathfrak{I}^{\mathbf{d}}} \sum_{k \in \mathfrak{I}^\psi} \frac{w_j^{\mathbf{d}} w_k^\psi p_{j,k,\ell}}{N_{\mathbf{d}} N_\psi q_j^{\mathbf{d}} q_k^\psi} \ . \tag{19}$$

We use these quantities, computed from the previous iterations, to construct the importance distributions at the current iteration. We further anneal these distributions to smooth out sharp peaks.

The actual importance distributions we use are two-component mixture models, comprising a convex combination of a uniform distribution and the approximate marginals from the last epoch. The uniform distribution helps to ensure that the importance distribution is non-zero everywhere. In effect, it encourages search over the entire space, while the other component of the mixture focuses search in regions that previously had high probability. Dropping the superscripts for clarity, let $\mathfrak{I}$ be the set of samples evaluated at the previous iteration and ϕ_i be the computed values for $i \in \mathfrak{I}$. Then the importance distribution used at the current iteration is

$$q_j = (1-\alpha) Z^{-1} \sum_{i \in \mathfrak{I}} \tilde{\phi}_i^{1/T} \mathbf{K}_{i,j} + \alpha\psi \tag{20}$$

where $\psi = M^{-1}$ is the uniform distribution, α is the mixing proportion with the uniform distribution, T is an annealing parameter, $\mathbf{K}_{i,j}$ is a kernel evaluated between quadrature points i and j, and $Z = \sum_j \sum_{i \in \mathfrak{I}} \tilde{\phi}_i^{1/T} \mathbf{K}_{i,j}$ is a normalization



GroEL-GroES [47] Thermus ATPase [24] Bovine ATPase [38]

Fig. 4: Previously published structures for the datasets used in this paper.

constant. The values for $\alpha = \max(0.05, 2^{-0.25\lfloor \tau_{prev}/50 \rfloor})$ and $T = \max(1.25, 2^{10.0/\lfloor \tau_{prev}/50 \rfloor})$ are set so that at early iterations, when the underlying density is changing, we rely more heavily on the uniform ψ.

The kernel $\mathbf{K}$ is used to diffuse probability around quadrature points as neighbouring points are more likely to be useful. It also enables smooth transitions between resolutions of quadrature points. Specifically, we use a squared exponential kernel for the shifts and an exponential kernel for the projection directions and in-plane rotations:

$$\mathbf{K}_{\mathbf{d}}(\mathbf{d}_i, \mathbf{d}_j) = \exp(\kappa_{\mathbf{d}} \mathbf{d}_i^T \mathbf{d}_j) \quad (21)$$

$$\mathbf{K}_{\psi}(\psi_i, \psi_j) = \exp(\kappa_{\psi} \cos \angle(\psi_i, \psi_j)) \quad (22)$$

$$\mathbf{K}_{\mathbf{t}}(\mathbf{t}_i, \mathbf{t}_j) = \exp(-\kappa_{\mathbf{t}} \|\mathbf{t}_i - \mathbf{t}_j\|^2) \quad (23)$$

where $\kappa_{\mathbf{d}} = \frac{\log 4}{1-\cos r_{\mathbf{d}}}$, $\kappa_{\psi} = \frac{\log 2}{1-\cos r_{\psi}}$, and $\kappa_{\mathbf{t}} = \frac{1}{2r_{\mathbf{t}}^2}$ are kernel bandwidth parameters and $\angle(\psi_i, \psi_j)$ is the angular difference between ψ_i and ψ_j. The bandwidth parameters are set based on the resolution of the quadrature schemes where $r_{\mathbf{d}}$ and r_{ψ} are the angular distances between projection directions and in-plane rotations respectively and $r_{\mathbf{t}}$ is the distance between translation quadrature points.

4 EXPERIMENTS

The proposed method was applied to two experimental and one synthetic dataset. All experiments were initialized using randomly generated densities. The maximum frequency considered was gradually increased from a relatively low frequency to near the Nyquist rate. Optimizations were run until the maximum resolution was reached and the average error on a held-out set of 200 particle images stopped improving, around 5000 iterations. In principle, higher resolution could be achieved but for this work the focus was on speed to achieve comparable resolution to existing published structures.

4.1 Datasets

The first dataset was ATP synthase from the *thermus thermophilus* bacteria, a large transmembrane protein complex. Transmembrane proteins are an important class of targets for cryo-EM. They constitute nearly 27% of human proteins [3] and play a crucial role in many diseases and cellular processes, yet they constitute less than 1% of the solved structures to date due to the fundamental limitations of other structure determination methods, notable, x-ray crystalography and NMR spectroscopy [13].

The *thermus* dataset contained $46,105$ particle images, provided by Lau and Rubinstein [24]. The high resolution structure from [24] and six sample images are shown in Fig. 2. The second dataset was *bovine* mitochondrial ATP synthase, containing $5,984$ particle images, and provide by Rubinstein et al. [38], For both datasets the particle images comprised 128×128 square pixels, the sides of which were 2.8Å in length (*i.e.*, $0.28nm$). The CTF information for each particle image was provided, having been estimated previously using CTFFIND3 [33]. The noise level, σ, was estimated by computing the standard deviation of pixels near the boundary of the particle images.

To demonstrate the ability of our method to handle a different type of structure, a third dataset was synthesized by taking an existing structure from the Protein Data Bank[2], GroEL-GroES-(ADP)7 [47], and generating $50,000$ random projections according to the generative model (see Fig. 5(top left)). The CTF, signal-to-noise level, image and pixel sizes, along with other parameters, were set to plausible values based on the *thermus* dataset values. This structure, as well as previously solved structures of the bovine and thermus ATP synthase molecules are depicted in Fig. 4. GroEL-GroES was selected because it is structurally unlike either of the bovine or thermus ATP synthase molecules.

4.2 Estimated Structures

For these datasets, structure estimation begins with a maximum frequency ω_* that corresponds to a wavelength of 40Å. The maximum frequency was then increased gradually, yielding finer resolution structures. At iteration τ, the maximum frequency, in cycles per Angstrom, was $\omega_* = \min(0.11, 0.02 + 0.005\lfloor \tau/150 \rfloor)$. The maximum frequency at the final iteration is $1/9$Å, close to the Nyquist frequency of $1/5.6$Å for these datasets. Importantly, a frequency of $1/9$Å corresponds to the resolution of the best published results for the thermus dataset used here [24].

Results on these datasets are shown in Fig. 5. Sample particle images are shown, along with an iso-surface and slices of the final estimated density. Computing these reconstructions took less than 24 hours in all cases. Further, even at early iterations, reasonable structures are available. Fig. 6 shows the estimated structure for the *thermus* dataset at different stages during optimization. Notably, after just one hour (during which only a fraction of the full dataset is seen), the low-resolution shape of the structure has already been determined.

4.3 Quantitative Evaluation

Traditionally, the cryo-EM field has used the *Fourier Shell Correlation* (FSC) to measure the similarity between two 3D densities. FSC is the normalized cross-correlation, computed using the 3D Fourier coefficients of the density as a function of frequency, *i.e.*, measured in shells of fixed frequency magnitude. This produces a function of frequency that is

2. Structure 1AON from http://pdb.org

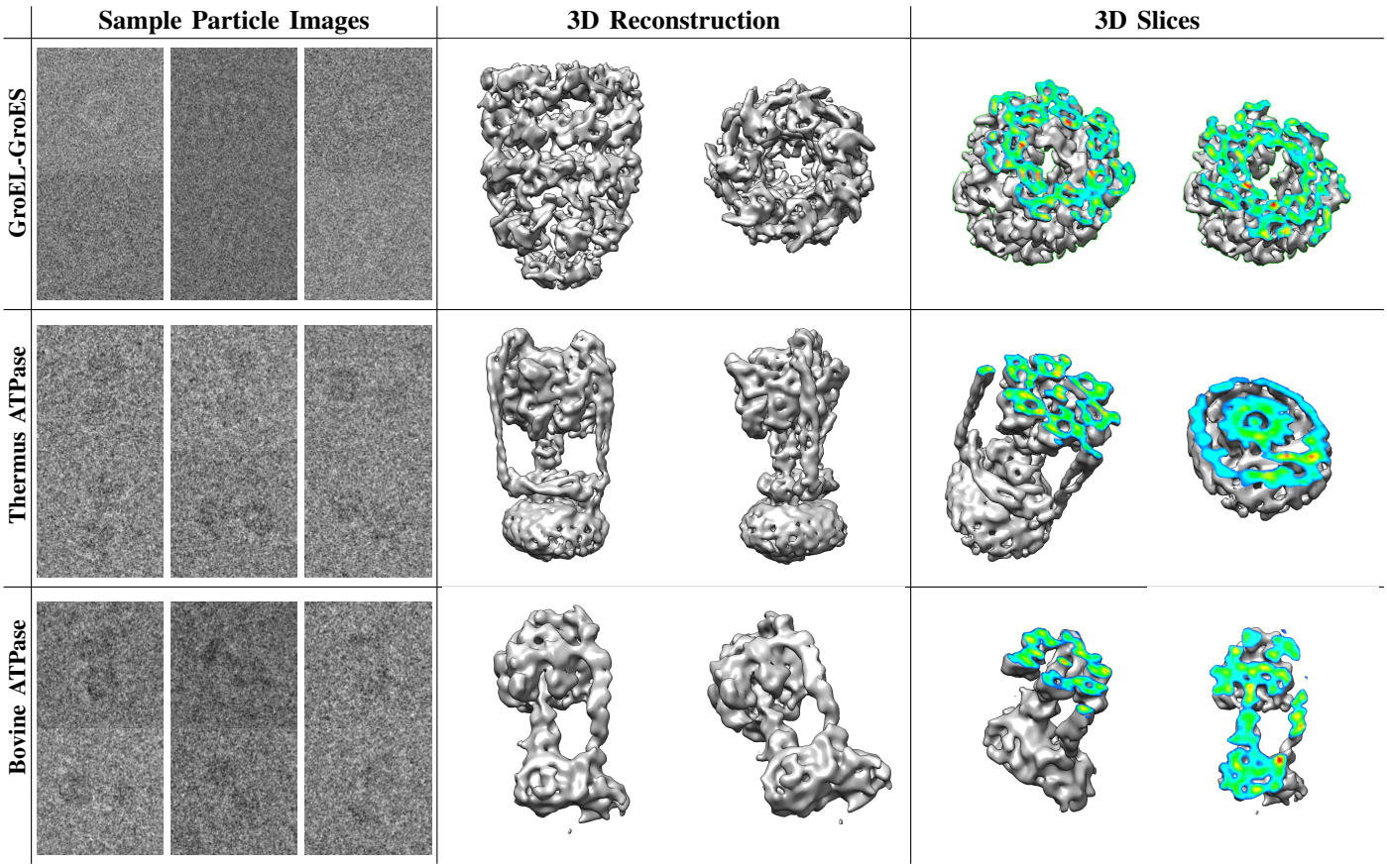


Fig. 5: Sample particle images (left), an isosurface of the reconstructed 3D density (middle) and slices through the 3D density with colour indicating relative density (right) for GroEL-GroES (top), thermus thermophilus ATPase (middle) and bovine mitochondrial ATPase (bottom). Reconstructions took a day or less on a single 16 core workstation.

typically near one for low frequency shells and decreases for higher frequency shells, as similarity between the densities decreases. The point at which the correlation drops below half is nominally defined as the resolution to which to two densities match. FSC requires that the two densities being compared are aligned. If this is not the case, we can compute an optimal alignment by maximizing the (overall) normalized cross-correlation using a simplex based search.

Ground-truth is rarely available for cryo-EM which makes it difficult to assess the accuracy of estimated densities. However, because the GroEL-GroES dataset is synthetic, we can compare against the ground truth structure to determine whether the estimated structure is accurate. The FSC curve for the GroEL-GroES dataset is show in Fig. 7. Here it can be seen that the estimated resolution of 9.1Å is consistent with the highest frequency of coefficients that were considered, which was 9Å. Power in the solved structures above this frequency is due to the influence of non-negativity and the prior.

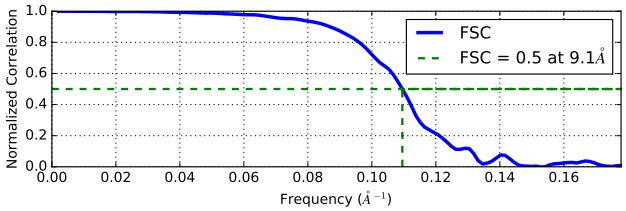


Fig. 7: Fourier Shell Correlation between the estimated structure and ground truth for GroEL/GroES. The estimated resolution of 9.1Å is consistent with the highest frequencies considered, *i.e.*, the largest value of ω_* (in cycles per Anstrom).

4.4 Importance Sampling

To validate our importance sampling approach we evaluated the error made in computing $\log p(\hat{\mathcal{I}}|\theta, \hat{\mathcal{V}})$ using IS against computing the exact sum in Eq. (4) without IS. This error is plotted in Fig. 8, along with the fraction of quadrature points used at various values of s_0. Based on these plots we selected a factor of $s_0 = 10$ for all experiments as a trade-off between accuracy and speed achieving a relative error of less then 0.1% while still providing significant speedups.

To see just how much of a speedup importance sampling provides in practice, Fig. 9 shows the fraction of quadrature points evaluated during optimization. Initially all quadrature points are evaluated, but as optimization progresses, and the density becomes better determined, importance sampling yields larger and larger speedups. At the full resolution, importance sampling provided more than a 100,000-fold speedup.

No prior knowledge of pose was assumed. Nevertheless, for many particles, certain views are more likely than others.

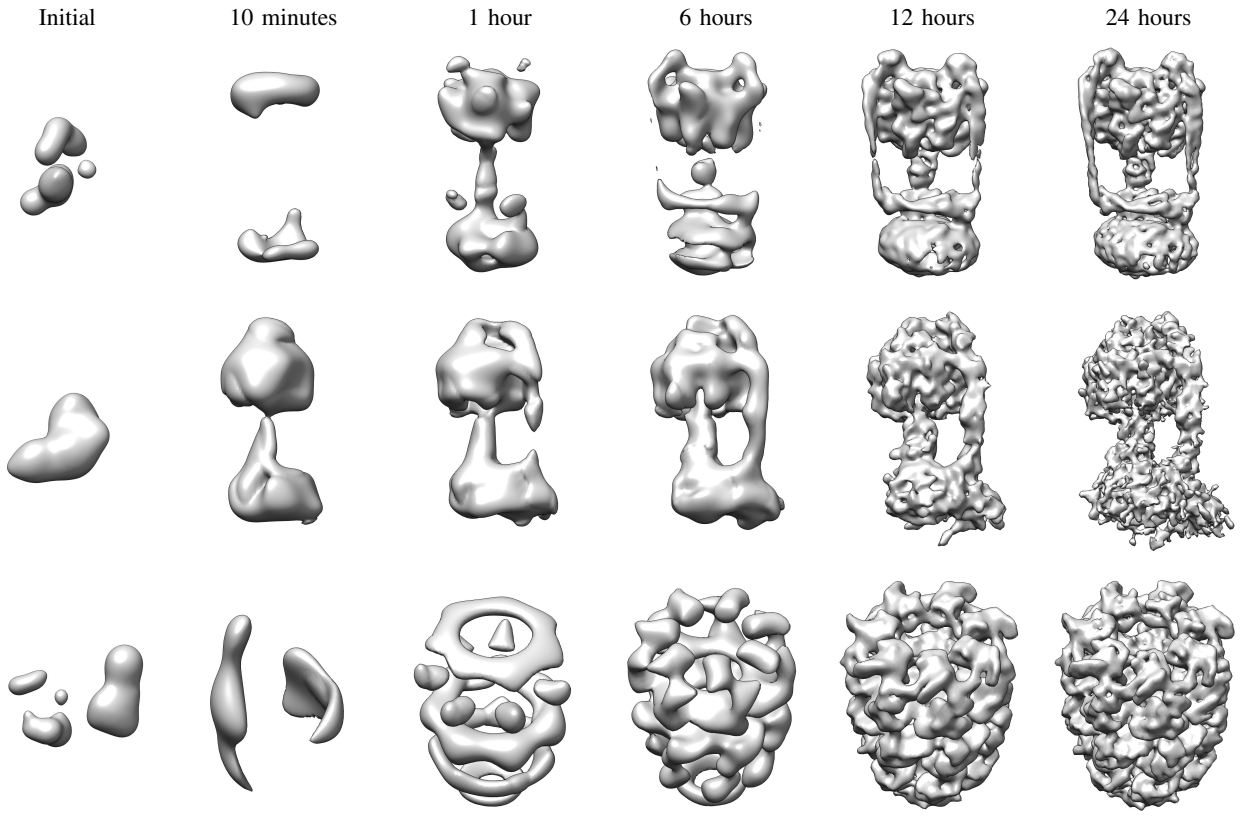


Fig. 6: Reconstruction progress at several times during a run of our method. Top row *thermus*, middle row *bovine*, and bottom row *GroEL-GroES* datasets. Initializations are generated randomly as a sum of spheres. Note that within an hour of computation, the gross structure is already well determined, after which fine details emerge gradually. Video sequences of reconstruction progress can be found at http://cs.toronto.edu/~alipunjani/pami16cryoem

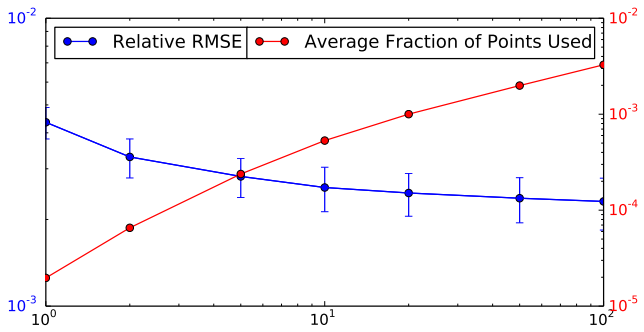


Fig. 8: Relative error (blue, left axis) and fraction of total quadrature points (red, right axis) used in computing $\log p(\hat{\mathcal{I}}|\theta, \hat{\mathcal{V}})$ as a function of the ESS scaling factor, s_0 (horizontal axis), on log-log axes.. Error bars represent the variance over a population of 100 individual images.

This fact can be seen by examining the average importance distribution for the thermus dataset, shown in Fig. 10 for a typical iteration. Here, the distribution of views forms an equatorial belt around the particle, while top or bottom views are rarely if ever seen. This phenomenon is well known for particles like these (*e.g.*, see [38] where this knowledge was

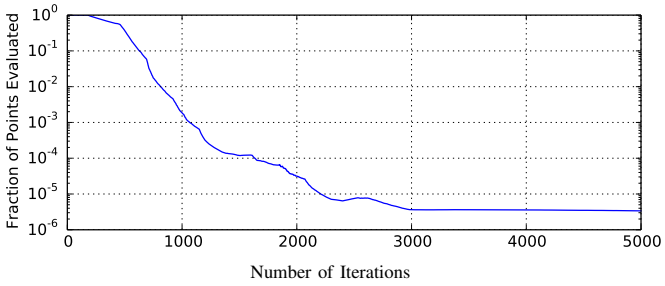


Fig. 9: The fraction of naive quadrature points evaluated on average during optimization, when using importance sampling. As resolution increases, the speedup obtained increases significantly yielding more than a 100,000 fold speedup.

used directly in estimation), validating our sampling approach and suggesting a use of this average importance distribution to supplement the uniform component of the mixture model importance distribution in Eq. (20).

4.5 Sensitivity to Initialization

To demonstrate the robustness of our method to initialization 15 different random initial densities were generated and used with the thermus dataset. Qualitatively, we find that the resulting structures from all runs are remarkably similar to one

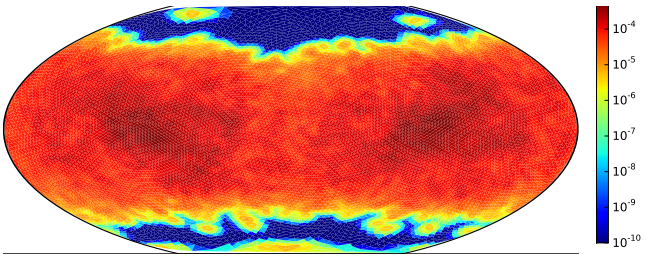


Fig. 10: A Winkel-Tripel projection of the importance distribution of view directions, $\mathbf{q}^{\mathbf{d}}$, averaged over the *thermus* dataset at a typical iteration. Clearly visible is the equatorial belt of likely views, while axis aligned views (those on the top or bottom of the plot) are rarely seen.

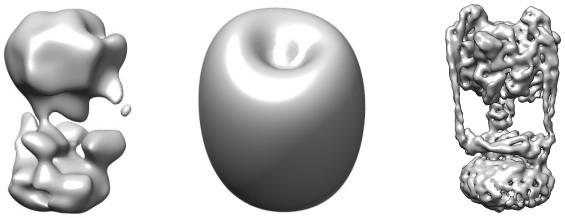


Fig. 11: Baseline comparisons to two existing standard methods. Iterative projection matching and reconstruction (left) and RELION [42] (middle). The proposed method (right) is able to determine the correct structure while projection matching and RELION both become trapped in poor local optima. See Fig. 11(middle) for comparison. All methods were given the same random initialization.

another. To assess quantitative similarity, the aligned FSC was computed between all pairs of the 15 final density maps. As mentioned above, FSC curves measure consistency of density maps as a function of resolution. Figure 12 shows FSC mean and standard deviation for the 105 pairs of density maps. The curve indicates that the densities are extremely close in structure all the way down to a resolution of 12Å on average, which is approximately half the Nyquist frequency, meaning that the results are consistent to approximately 2 pixels. This indicates that the different random initializations are effectively converging to very similar solutions.

To compare this method to others, especially its sensitivity to initialization, we selected two well-known, state-of-the-art approaches for structure determination. The first is a standard iterative projection matching scheme where images are matched to an initial density through a global cross-correlation search (*e.g.*, as in [16]). The density is then reconstructed based on these fixed orientations and this process is iterated. The second benchmark is the RELION package described in [42]. It uses a similar marginalized model as our method but with a batch EM algorithm to perform optimization.

We used publicly available code[3] for both of these approaches on the *thermus* dataset and initialized using the same randomly generated density. We ran each method for a number

3. For projection matching we used the code from https://sites.google.com/site/rubinsteingroup/home

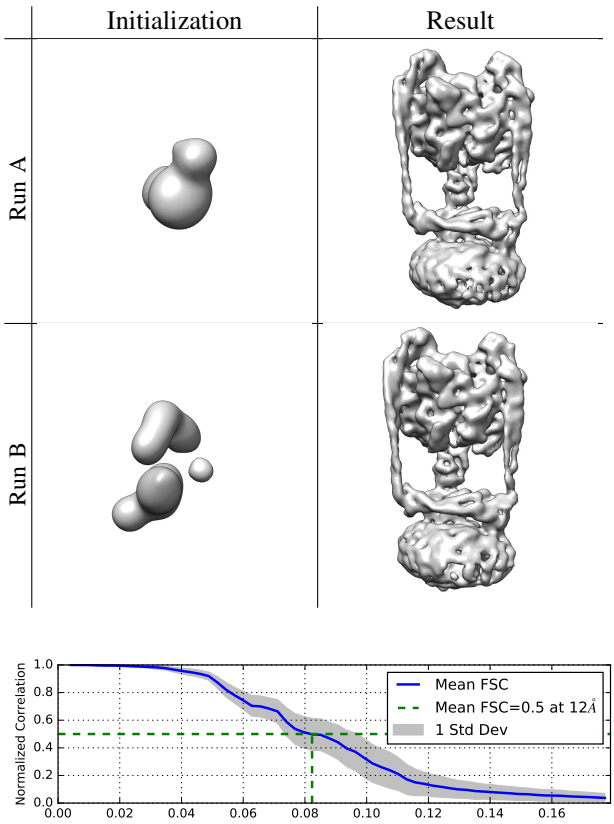


Fig. 12: Comparison of multiple restarts with the same Thermus data using different randomly generated initializations. Top: Example initializations and results from two individual runs. Bottom: Average Fourier Shell Correlation between the 15 independent runs. On average, the results were consistent to 12Å which is approximately twice wavelength of the Nyquist frequency, *i.e.*, the structures are consistent to approximately 2 pixels. This demonstrates the robustness of the method to random initializations.

of iterations roughly equivalent in terms of CPU time to the 5000 iterations used by our method. Fig. 11 shows the results. Both approaches have clearly determined an incorrect structure, and appear to have converged to a local minimum as no further progress was made beyond this point.

We note that both projection matching and RELION have been used successfully for reconstruction by others, and are not recommended to be used without a good initialization. Our results support this recommendation as neither approach converges from random initializations. In practice, it is difficult to construct good initializations for molecules of unknown shape [17], giving the proposed method a significant advantage. However, while it has not been seen in practice with our method, it is still possible that poor local optima could be found as the optimization problem is non-convex and potentially has many local optima.

4.6 Comparing Priors

The above results used an exponential prior for the density at the voxels of $\mathcal{V}_i$, however the presented framework allows for

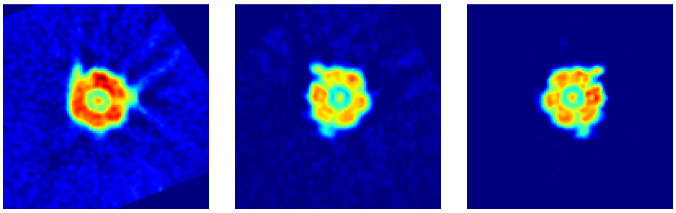


Fig. 13: Slices through the reconstructions with (from left to right) uniform, CAR and exponential priors. The exponential prior does the best job of suppressing noise in the background without oversmoothing fine details within the structure. Blue corresponds to small or zero density and red corresponds to high density.

any continuous and differentiable prior. To demonstrate this, we explored two other priors, namely, an improper (uniform) prior, $p(\mathcal{V}_i) \propto 1$, and a conditionally autoregressive (CAR) prior [7] $p(\mathcal{V}_i|\mathcal{V}_{-i}) = \mathcal{N}(\mathcal{V}_i|\frac{1}{26}\sum_{j\in\text{Nbhd}(i)}\mathcal{V}_j, \sigma^2_{CAR})$ which is a smoothness prior biasing each voxel towards the mean of its 26 immediate neighbours $\text{Nbhd}(i)$. Slices through the resulting densities on *thermus* under these priors are shown in Fig. 13. With an improper uniform prior (Fig. 13(left), there is significant noise visible in the background. This noise is somewhat suppressed with the CAR prior (Fig. 13(middle) however the best results are clearly obtained using the exponential prior which suppresses the background noise without smoothing out internal details.

4.7 Limitations

Beyond the work described in this paper, there remain a number of unresolved questions for future research. While an exponential prior was found to be effective, more sophisticated priors could be explored or learned, potentially enabling higher resolution estimation without the need to collect more data and providing a kind of of atomic-scale super-resolution.

As noted in the appendix, the Gaussian approximation to the Poisson noise model is well motivated, however experimental data suggests that there are correlations in the noise. This suggests that a coloured noise model may be yield even better results. Further, some datasets exhibit crowding, where there are structured outliers in particle images that are not well captured in the existing noise model.

The optimization problem is challenging, and, while SAGD was successful here, it is likely that more efficient stochastic optimization methods are possible which exploit the problem structure to a greater degree. This will likely be critical in recovering even higher resolution structures as the Hessian matrix can be shown to become poorly conditioned, meaning that first-order optimization methods will likely struggle to reach the optimal solution.

Finally, validation and resolution assessment remains a significant open problem [17]. Unfortunately, the nature of the problem is such that there is no ground truth for novel structures. Existing validation techniques have limitations (*e.g.*, gold-standard FSC) or are experimentally cumbersome (*e.g.*, tilt-pair tests) and new tools for validation are necessary.

5 Conclusions

This paper introduces a framework for efficient 3D molecular reconstruction from cryo-EM images. It comprises MAP estimation of 3D structure with a generative model, marginalization over 3D particle poses, and optimization using SAGD. A novel importance sampling scheme was used to reduce the computational cost of marginalization. The resulting approach can be applied to large stacks of cryo-EM images, providing high resolution reconstructions in a day on a 16-core workstation, starting from a random initialization.

The problem of density estimation for cryo-EM is a fascinating vision problem. The low SNR in particle images makes it remarkable that any molecular structure can be estimated, let alone the high resolution densities which are now common. Recent research [29] suggests that the combination of new techniques and new sensors may facilitate atomic resolution reconstructions for arbitrary molecules. This development will be ground-breaking in both biological and medical research.

In order to encourage others to work on this problem, source code is available from the authors' website.

References

[1] "RELION," http://www2.mrc-lmb.cam.ac.uk/relion, Accessed: 2015-10-30.

[2] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski, "Building rome in a day," in *ICCV*, 2009.

[3] M. S. Almén, K. J. Nordström, R. Fredriksson, and H. B. Schiöth, "Mapping the human membrane proteome: a majority of the human membrane proteins can be classified according to function and evolutionary origin," *BMC biology*, vol. 7, no. 1, 2009.

[4] S.-I. Amari, H. Park, and K. Fukumizu, "Adaptive method of realizing natural gradient learning for multilayer perceptrons," *Neural Computation*, vol. 12, no. 6, pp. 1399–1409, 2000.

[5] S. Amari, "Natural gradient works efficiently in learning," *Neural Computation*, vol. 10, no. 2, pp. 251–276, 1998.

[6] W. T. Baxter, R. A. Grassucci, H. Gao, and J. Frank, "Determination of signal-to-noise ratios and spectral snrs in cryo-em low-dose imaging of molecules," *J Struct Biol*, vol. 166, no. 2, pp. 126–32, May 2009.

[7] J. Besag, "Statistical analysis of non-lattice data," *Journal of the Royal Statistical Society. Series D (The Statistician)*, vol. 24, no. 3, pp. 179–195, 1975.

[8] R. Bhotika, D. Fleet, and K. Kutulakos, "A probabilistic theory of occupancy and emptiness," in *ECCV*, 2002.

[9] M. A. Brubaker, A. Punjani, and D. J. Fleet, "Building Proteins in a Day: Efficient 3D Molecular Reconstruction," in *Proc. CVPR*, 2015.

[10] E. Callaway, "The revolution will not be crystallized: A new method sweeps through structural biology," *Nature*, vol. 525, no. 7568, pp. 172–174, 2015.

[11] J. de la Rosa-Trevín, J. Otòn, R. Marabini, A. Zaldívar, J. Vargas, J. Carazo, and C. Sorzano, "Xmipp 3.0: An improved software suite for image processing in electron microscopy," *Journal of Structural Biology*, vol. 184, no. 2, pp. 321–328, 2013.

[12] A. Doucet, S. Godsill, and C. Andrieu, "On sequential Monte Carlo sampling methods for Bayesian filtering," *Statistics and Computing*, vol. 10, no. 3, pp. 197–208, 2000.

[13] S. J. Fleishman, V. M. Unger, and N. Ben-Tal, "Transmembrane protein structures without x-rays," *Trends in Biochemical Sciences*, vol. 31, no. 2, pp. 106 – 113, 2006.

[14] M. Gräf and D. Potts, "Sampling sets and quadrature formulae on the rotation group," *Numerical Functional Analysis and Optimization*, vol. 30, no. 7-8, pp. 665–688, 2009.

[15] J. Gregson, M. Krimerman, M. B. Hullin, and W. Heidrich, "Stochastic tomography and its applications in 3d imaging of mixing fluids," *ACM Trans. Graph. (Proc. SIGGRAPH 2012)*, vol. 31, no. 4, pp. 52:1–52:10, 2012.
[16] N. Grigorieff, "Frealign: high-resolution refinement of single particle structures," *J Struct Biol*, vol. 157, no. 1, p. 117–125, Jan 2007.
[17] R. Henderson, A. Sali, M. L. Baker, B. Carragher, B. Devkota, K. H. Downing, E. H. Egelman, Z. Feng, J. Frank, N. Grigorieff, W. Jiang, S. J. Ludtke, O. Medalia, P. A. Penczek, P. B. Rosenthal, M. G. Rossmann, M. F. Schmid, G. F. Schröder, A. C. Steven, D. L. Stokes, J. D. Westbrook, W. Wriggers, H. Yang, J. Young, H. M. Berman, W. Chiu, G. J. Kleywegt, and C. L. Lawson, "Outcome of the first electron microscopy validation task force meeting," *Structure*, vol. 20, no. 2, pp. 205 – 214, 2012.
[18] J. Hsieh, *Computed Tomography: Principles, Design, Artifacts, and Recent Advances*. SPIE, 2003.
[19] N. Jaitly, M. A. Brubaker, J. Rubinstein, and R. H. Lilien, "A Bayesian Method for 3-D Macromolecular Structure Inference using Class Average Images from Single Particle Electron Microscopy," *Bioinformatics*, vol. 26, pp. 2406–2415, 2010.
[20] J. Keeler, *Understanding NMR Spectroscopy*. Wiley, 2010.
[21] E. J. Kirkland, *Advanced Computing in Electron Microscopy*, 2nd ed. Springer, 2010.
[22] K. N. Kutulakos and S. M. Seitz, "A theory of shape by space carving," *IJCV*, vol. 38, no. 3, pp. 199–218, 2000.
[23] R. Langlois, J. Pallesen, J. T. Ash, D. N. Ho, J. L. Rubinstein, and J. Frank, "Automated particle picking for low-contrast macromolecules in cryo-electron microscopy," *Journal of Structural Biology*, vol. 186, no. 1, pp. 1 – 7, 2014.
[24] W. C. Y. Lau and J. L. Rubinstein, "Subnanometre-resolution structure of the intact Thermus thermophilus H+-driven ATP synthase," *Nature*, vol. 481, pp. 214–218, 2012.
[25] N. Le Roux and A. Fitzgibbon, "A fast natural Newton method," in *ICML*, 2010.
[26] N. Le Roux, P.-A. Manzagol, and Y. Bengio, "Topmoumoute online natural gradient algorithm," in *NIPS*, 2008, pp. 849–856.
[27] N. Le Roux, M. Schmidt, and F. Bach, "A stochastic gradient method with an exponential convergence rate for strongly convex optimization with finite training sets," in *NIPS*, 2012.
[28] V. J. Lebedev and D. N. Laikov, "A quadrature formula for the sphere of the 131st algebraic order of accuracy," *Doklady Mathematics*, vol. 59, no. 3, pp. 477 – 481, 1999.
[29] X. Li, P. Mooney, S. Zheng, C. R. Booth, M. B. Braunfeld, S. Gubbens, D. A. Agard, and Y. Cheng, "Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-em," *Nature Methods*, vol. 10, no. 6, pp. 584–590, 2013.
[30] S. P. Mallick, S. Agarwal, D. J. Kriegman, S. J. Belongie, B. Carragher, and C. S. Potter, "Structure and view estimation for tomographic reconstruction: A Bayesian approach," in *CVPR*, 2006.
[31] T. Malzbender, "Fourier volume rendering," *ACM Trans. Graph.*, vol. 12, no. 3, pp. 233–250, Jul. 1993.
[32] J. Martens, "Deep learning via hessian-free optimization," in *ICML*, 2010.
[33] J. A. Mindell and N. Grigorieff, "Accurate determination of local defocus and specimen tilt in electron microscopy," *Journal of Structural Biology*, vol. 142, no. 3, pp. 334–47, 2003.
[34] Y. Nesterov, "A method of solving a convex programming problem with convergence rate o (1/k2)," *Soviet Mathematics Doklady*, vol. 27, no. 2, pp. 372–376, 1983.
[35] B. Polyak, "Some methods of speeding up the convergence of iteration methods," *USSR Computational Mathematics and Mathematical Physics*, vol. 4, no. 5, pp. 1–17, 1964.
[36] A. Punjani and M. A. Brubaker, "Microscopic Advances with Large-Scale Learning: Stochastic Optimization for Cryo-EM," in *Neural Information Processing Systems Workshop: Machine Learning in Computational Biology (MLCB)*, December 2014.
[37] L. Reimer and H. Kohl, *Transmission Electron Microscopy: Physics of Image Formation*. Springer, 2008.
[38] J. L. Rubinstein, J. E. Walker, and R. Henderson, "Structure of the mitochondrial atp synthase by electron cryomicroscopy," *The EMBO Journal*, vol. 22, no. 23, pp. 6182–6192, 2003.
[39] B. Rupp, *(2009). Biomolecular Crystallography: Principles, Practice and Application to Structural Biology*. Garland Science, 2009.
[40] E. B. Saff and A. B. J. Kuijlaars, "Distributing many points on a sphere," *The Mathematical Intelligencer*, vol. 19, no. 1, pp. 5–11, 1997.
[41] S. H. W. Scheres, H. Gao, M. Valle, G. T. Herman, P. P. B. Eggermont, J. Frank, and J.-M. Carazo, "Disentangling conformational states of macromolecules in 3d-em through likelihood optimization," *Nature Methods*, vol. 4, pp. 27–29, 2007.
[42] S. H. Scheres, "RELION: Implementation of a Bayesian approach to cryo-EM structure determination ," *Journal of Structural Biology*, vol. 180, no. 3, pp. 519 – 530, 2012.
[43] F. Sigworth, "A maximum-likelihood approach to single-particle image refinement," *Journal of Structural Biology*, vol. 122, no. 3, pp. 328 – 339, 1998.
[44] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *ICML*, 2013.
[45] G. Tang, L. Peng, P. Baldwin, D. Mann, W. Jiang, I. Rees, and S. Ludtke, "EMAN2: an extensible image processing suite for electron microscopy," *Journal of Structural Biology*, vol. 157, no. 1, pp. 38–46, 2007.
[46] S. T. Tokdar and R. E. Kass, "Importance sampling: a review," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 1, pp. 54–60, 2010.
[47] Z. Xu, A. L. Horwich, and P. B. Sigler, "The crystal structure of the asymmetric GroEL-GroES-(ADP)7 chaperonin complex," *Nature*, vol. 388, pp. 741–750, 1997.
[48] J. Zhao, M. A. Brubaker, and J. L. Rubinstein, "TMaCS: A hybrid template matching and classification system for partially-automated particle selection," *Journal of Structural Biology*, vol. 181, no. 3, pp. 234 – 242, 2013.

APPENDIX A
FORMULATION OF GENERATIVE MODEL

The physics of image formation in a transmission electron microscope are described by the Schrödinger equation of quantum mechanics. This partial differential equation would be prohibitively expensive and difficult to work with directly. However, due to the thin nature of samples used in cryo-EM, the weak phase object approximation is widely used which results in the linear image formation model we describe below. For more details on the physics of the microscope and this approximation we refer the reader to [21].

In the weak phase object approximation the electron microscope takes an integral projection of a rotated, translated 3D molecular density. Let $V(\mathbf{y})$ be the 3D density at $\mathbf{y} \in \mathbb{R}^3$, and let the transformation from the particle coordinate frame to the microscope coordinate frame be

$$\mathbf{x}_3 = \mathbf{R}\mathbf{y} + \mathbf{t}_3 \tag{24}$$

where $\mathbf{R}$ and $\mathbf{t}_3 \equiv (t_1, t_2, t_3)^\mathsf{T}$ are the 3D rotation and translation between coordinate frames, and $\mathbf{x}_3 \equiv (x_1, x_2, x_3)^\mathsf{T}$. Without loss of generality we assume the microscope projection is parallel to the x_3-axis, and we express the rotation as $\mathbf{R}^\mathsf{T} = [\mathbf{n}_1, \mathbf{n}_2, \mathbf{d}]$, so the projection direction in the particle

frame is $\mathbf{d}$. The integral projection, onto $\mathbf{x} \equiv (x_1, x_2)^\mathsf{T}$, is then given by

$$P(\mathbf{x};\, \mathbf{R}, \mathbf{t}) \;=\; \int V(\mathbf{R}^\mathsf{T}(\mathbf{x}_3 - \mathbf{t}_3))\; dx_3 \;. \qquad (25)$$

In addition to the rotation, the orthgonal projection only depends critically on the two components of translation in the plane normal to the projection direction $\mathbf{d}$, *i.e.*, $\mathbf{t} \equiv (t_1, t_2)$.

According to the well-known Fourier Slice Theorem, the Fourier transform of an orthogonal projection of V is equivalent to a planar slice through the 3D Fourier spectrum of V. That is, the 2D transform of $P(\mathbf{x};\, \mathbf{R}, \mathbf{t})$, denoted $\hat{P}(\boldsymbol{\omega};\, \mathbf{R}, \mathbf{t})$, with $\boldsymbol{\omega} \equiv (\omega_1, \omega_2)^\mathsf{T}$, can be written as

$$\hat{P}(\boldsymbol{\omega}; \mathbf{R}, \mathbf{t}) \equiv \int P(\mathbf{x};\, \mathbf{R}, \mathbf{t})\; e^{-\iota 2\pi \mathbf{x}^\mathsf{T}\boldsymbol{\omega}} d\mathbf{x} \qquad (26)$$

$$= e^{-\iota 2\pi \boldsymbol{\omega}^\mathsf{T}\mathbf{t}}\; \hat{V}(\omega_1 \mathbf{n}_1 + \omega_2 \mathbf{n}_2) \qquad (27)$$

In words, the 2D Fourier spectrum comprises a slice through the 3D spectrum of V and a phase shift. The slice through $\hat{V}$ is on a plane through the origin and normal to $\mathbf{d}$. The phase shift corresponds to the translation between particle and microscope coordinates in the plane normal to $\mathbf{d}$.

The projection P is subject to defocus and microscope aberration, which is characterized as the modulation of a contrast transfer function (CTF) in Fourier domain, and denoted $\hat{c}(\boldsymbol{\omega}; \theta)$ with parameters θ. The standard parametric CTF model is derived as part of the weak phase object approximation. The specific form is given by Eq. (3) in [33]), and there exists freely available and widely used software (called CTFFIND3) for parameter estimation of θ. As depicted in Figure 2, the CTF for an electron microscope oscillates with periodic phase reversals, not typical of traditional light cameras. As a consequence, EM images are not strictly positive, and important frequency structure in the neighborhoods of the zeros is lost. Finally, it is important to note that the locations of zeros vary with experimental settings of the microscope, and are often varied from micrograph to micrograph to ensure that all frequencies are represented among the the set of particle images. (We refer the interested read to [37] for a a more complete treatment of the CTF.)

The final image is also contaminated by noise, two sources of which are Poisson distributed electron noise, and absorption due to varying thickness of the ice in which the particles are suspended. As is common in the cryo-EM literature, we assume a Gaussian approximation to the Poisson electron noise, and constant ice density in each particle image. The mean electron noise and the local ice absorption are estimated around the border of each particle image (since particles are roughly centered by the particle picker), and subtracted from the particle images. This is a process commonly known as "floating" in the cryo-EM literature. As a consequence, we assume a mean-zero, white Gaussian noise process with variance σ^2 and zero density outside the support of the particle.

Putting the above together, the image in the microscope can be expressed, in the spatial domain, as follows

$$I(\mathbf{x}) \;=\; (c * P)\,(\mathbf{x}; \theta, \mathbf{R}, \mathbf{t}) \;+\; \nu(\mathbf{x}) \;, \qquad (28)$$

where $*$ denotes convolution, $c(\mathbf{x}; \theta)$ is the real-space form of the CTF (the point spread function), and ν denotes the white Gaussian noise process. In the Fourier domain,

$$\hat{I}(\boldsymbol{\omega}) \;=\; \hat{c}(\boldsymbol{\omega}; \theta)\, e^{-\iota 2\pi\, \boldsymbol{\omega}^\mathsf{T}\mathbf{t}}\, \hat{V}(\omega_1 \mathbf{n}_1 + \omega_2 \mathbf{n}_2) \;+\; \hat{\nu}(\boldsymbol{\omega}) \;. \qquad (29)$$

Because the Fourier transform is unitary,[4] the noise remains additive and Gaussian in the Fourier domain. That is, the distribution of noise at frequency $\boldsymbol{\omega}$ is proportional to a mean-zero, isotropic complex normal random variable with variance $2\sigma^2$. The proportionality and rescaled variance is due to the fact that the noise is constrained to be purely real, resulting in a Hermitian Fourier transform.

In practice we discretize the density V to obtain a 3D grid, $\mathcal{V}$, with density represented at each of D^3 voxels. Let $\hat{\mathcal{V}}$ denote the 3D discrete Fourier transform (DFT) of $\mathcal{V}$. The discrete particle images, denoted $\mathcal{I}$, formed through orthographic projection, comprise D^2 pixels, sharing the same resolution as the 3D voxels. The 2D DFT of $\mathcal{I}$, denoted $\hat{\mathcal{I}}$, is defined at frequencies $\boldsymbol{\omega}$ in $\Omega \equiv \{(n/D, m/D)^\mathsf{T}\}_{0 \le n,m < D}$. Given the Gaussian noise model, the Fourier coefficients are independent so the joint likelihood can be written as the product; *i.e.*,

$$p(\hat{\mathcal{I}} \,|\, \theta, \mathbf{R}, \mathbf{t}, \mathcal{V}, \sigma^2) \propto \prod_{\boldsymbol{\omega} \in \Omega} \mathcal{CN}(\hat{\mathcal{I}}[\boldsymbol{\omega}]; \mu(\boldsymbol{\omega}), 2\sigma^2) \qquad (30)$$

where $\hat{\mathcal{I}}[\boldsymbol{\omega}]$ denotes the DFT coeficient at frequency $\boldsymbol{\omega}$, and $\mu(\boldsymbol{\omega}) = \hat{c}(\boldsymbol{\omega}; \theta)\, e^{-\iota 2\pi\, \boldsymbol{\omega}^\mathsf{T}\mathbf{t}}\, \hat{\mathcal{V}}(\omega_1 \mathbf{n}_1 + \omega_2 \mathbf{n}_2)$ and $\mathcal{CN}(\cdot)$ denotes the Complex Normal distribution. To evaluate the Fourier coefficients in (30) we must interpolate $\hat{\mathcal{V}}$. To do this we use trilinear interpolation with premultiplication, however more advanced approaches are possible. See [31] for a thorough discussion of issues around interpolation in the Fourier domain for volume rendering.

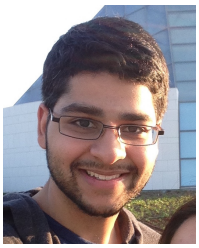
One key advantage of expressing the likelihood over Fourier coefficients is that we can easily adapt the likelihood to perform low resolution estimation with significant computational savings. In particular, it is easy to include only low frequency term, *e.g.*, up to a frequency cut-off, ω_*:

$$p(\hat{\mathcal{I}} \,|\, \theta, \mathbf{R}, \mathbf{t}, \mathcal{V}, \sigma^2) \propto \prod_{\substack{\boldsymbol{\omega} \in \Omega \\ ||\boldsymbol{\omega}|| < \omega_*}} \mathcal{CN}(\hat{\mathcal{I}}[\boldsymbol{\omega}]; \mu(\boldsymbol{\omega}), 2\sigma^2) \qquad (31)$$

The savings stem from the fact that the number of Fourier coefficients increases quadratically with frequency.

Optimization entails computation of the gradient of the likelihood with respect to the density $\mathcal{V}$, and in doing so, the CTF and pose parameters, and any terms that depend on them, are treated as fixed. Computation of the gradients with respect to $\hat{\mathcal{V}}$ is straightforward because of the linearity of interpolation and the image formation model. The gradient with respect to $\mathcal{V}$ can then be found through the chain rule to be the (unitary) inverse Fourier transform of the gradient with respect to $\hat{\mathcal{V}}$. This is possible because of the linear, unitary nature of the Fourier transform.

4. Note that while the standard definition of a continuous Fourier transform is unitary, most implementations of the discrete Fourier transform are not. In practice the DFT must be rescaled or the equations here must be scaled appropriately.



Ali Punjani received the B.A.Sc. degree in aerospace engineering from the University of Toronto, Canada, in 2012, M.S. degree in computer science from the University of California, Berkeley, USA, and is now a Ph.D. student at the University of Toronto where he works on computer vision and machine learning. His research interests include deep learning, optimization methods, and computational biology. He has also worked on large-scale data visualization, robotics and autonomous control. He is a recipient of the NSERC Canada Graduate Scholarship.

Marcus A. Brubaker received his Ph.D. in Computer Science from the University of Toronto in 2011 and from 2011 to 2016 he was a postdoctoral fellow at TTI-Chicago and University of Toronto. He is currently an Assistant Professor at York University in Toronto, Canada and consults with Cadre Research Labs. His research interests span statistics, machine learning and computer vision and includes applications in computational biology, robotics and forensics.

David J. Fleet David J Fleet received his PhD in Computer Science from the University of Toronto in 1991. Following 8 years on faculty at Queen's University, and then 5 years at the Palo Alto Research Center (PARC), he joined the University of Toronto as Professor of Computer Science. He is Senior Fellow of the Canadian Institute for Advanced Research.

His research interests include computer vision, image processing, visual perception, and visual neuroscience. He has published numerous research articles on a broad range of topics, including optical flow, motion perception and human stereopsis, image-based tracking, 3D hand and human pose tracking, latent variable models, physics-based models for motion analysis, and large-scale image retrieval.

He received the Alfred P. Sloan Research Fellowship in 1996. He has won several research awards, including the 2010 Koenderink Prize for his work with Michael Black and Hedvig Sidenbladh on human pose tracking. He has served as Area Chair for numerous computer vision and machine learning conference. He was Program Co-chair for the CVPR 2003 and ECCV 2014. He has been Associate Editor, and Associate Editor-in-Chief for IEEE TPAMI, and currently serves on the TPAMI Advisory Board.