Efficient 3D Macromolecular Reconstruction with Electron Cryomicroscopy

Marcus A. Brubaker Ali Punjani David J. Fleet
University of Toronto
{mbrubake,alipunjani,fleet}@cs.toronto.edu

Introduction Discovering the 3D atomic structure of molecules such as proteins and viruses is a fundamental research problem in biology and medicine. The ability to routinely determine the 3D structure of such molecules would potentially revolutionize the process of drug development and accelerate research on fundamental biological processes. Electron Cryomicroscopy (Cryo-EM) is a vision-based approach to 3D macromolecular structure determination which works with medium to large-sized molecules in their native state.

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). The problem is similar in spirit to multi-view scene carving [2, 7], large-scale, uncalibrated multi-view reconstruction [1] and computed tomography (CT) [6, 4] which uses a similar imaging model (orthographic integral projection). However in CT the projection direction of each image is known whereas with single particle Cryo-EM this is unknown.

Existing Cryo-EM techniques, *e.g.*, [5, 12], can be extremely slow and require good initialization to converge to good solutions. 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 and insensitive to initialization, providing useful low resolution density estimates in an hour on a single workstation. To demonstrate our method, we perform reconstructions on two real datasets.

A Framework for 3D Density Estimation In Cryo-EM, particle images are formed as orthographic, integral projections of the electron density of a molecule, $\mathcal{V} \in \mathbb{R}^{D^3}$. In each image, the density is oriented in an unknown pose, $\mathbf{R} \in \mathcal{SO}(3)$, and shifted from the center of the image by an unknown translation $\mathbf{t} \in \mathbb{R}^2$. This image formation model is linear and can be represented by the matrix $\mathbf{P}_{\mathbf{R},\mathbf{t}} \in \mathbb{R}^{D^2 \times D^3}$. Noise in Cryo-EM images is modelled using an IID Gaussian distribution, and the unknown pose parameters, $\mathbf{R}$ and $\mathbf{t}$, are marginalized out. Thus, the conditional distribution of a particle image, $\mathcal{I} \in \mathbb{R}^{D^2}$, is

$$p(\mathcal{I} \mid \mathcal{V}) = \int_{\mathbb{R}^2} \int_{\mathcal{SO}(3)} \mathcal{N}(\mathcal{I} \mid \mathbf{P}_{\mathbf{R},\mathbf{t}}\mathcal{V}, \sigma^2\mathbf{I}) d\mathbf{R} d\mathbf{t} \qquad (1)$$

where σ is the standard deviation of the noise, $\mathcal{N}(\cdot|\mu, \Sigma)$ is the multivariate normal distribution, and $p(\mathbf{R})$ and $p(\mathbf{t})$ are (uniform) priors over the pose parameters. This double integral is not analytically tractable, so numerical approaches must be used.

Given a set of K images $\mathfrak{D} = \{\mathcal{I}_i\}_{i=1}^K$ and assuming conditional independence of the images, the posterior probability of a density $\mathcal{V}$ is

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

where $p(\mathcal{V})$ is an exponential prior over 3D voxels. Estimating the density consists of finding $\mathcal{V}$ which maximizes Equation (2). Taking the negative log and dropping constant factors, the optimization problem becomes $\arg\min_{\mathcal{V} \in \mathbb{R}_+^{D^3}} f(\mathcal{V})$,

$$f(\mathcal{V}) = -\log p(\mathcal{V}) - \sum_{i=1}^{K} \log p(\mathcal{I}_i|\mathcal{V}) \qquad (3)$$

where $\mathcal{V}$ is restricted to be positive.

To efficiently cope with the large number of particle images in a typical dataset (tens or hundreds of thousands), we propose to use stochastic optimization. Stochastic optimization exploits the large amount of redundancy in most datasets by only considering subsets of data at each iteration by rewriting the objective as $f(\mathcal{V}) = \sum_k f_k(\mathcal{V})$ where each $f_k(\mathcal{V})$ evaluates only a subset of data. This allows for fast progress to be made before a batch optimization algorithm would be able to take a single step. We specifically use Stochastic Average Gradient Descent (SAGD) [9] because it requires minimal tuning and is designed for the finite data case, allowing for faster convergence. To enforce the positivity of density, negative values of $\mathcal{V}$ are truncated to zero after each iteration. To reduce the cost of computing the

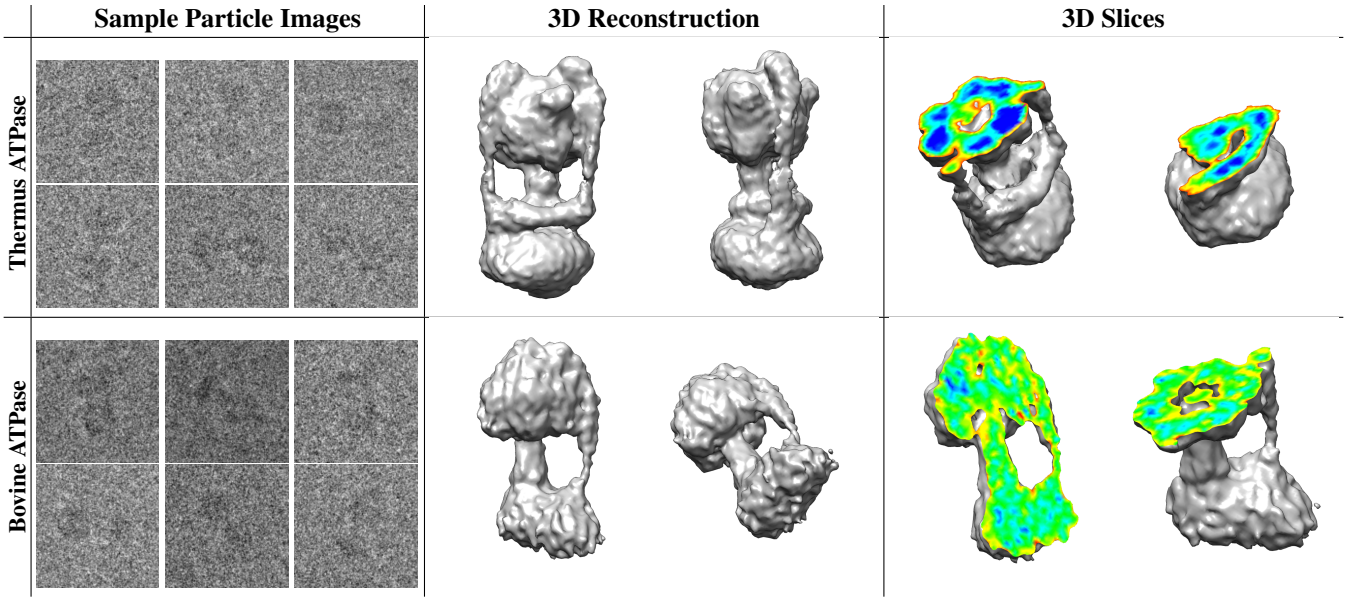

Figure 1: Sample particle images (left), an isosurface of the reconstructed 3D density (middle) and slices through the 3D density with colour indicating relative density (right) for, thermus thermophilus ATPase Lau and Rubinstein [8] (top) and bovine mitochondrial ATPase [11] (bottom). Reconstructions took a day or less on a 16 core workstation.

gradient for each particle image we use importance sampling to approximate the integrals in Eq. (1). This is critical and produces speedups of five orders of magnitude.

Results and Conclusions The proposed method was applied to two experimental datasets and the results are shown in Fig. 1. 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 on a single workstation. Full details of the method can be found in [3]. 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 advances [10] suggests that atomic resolution reconstructions for arbitrary molecules may soon be feasible, a landmark result for biological and medical research.

References

[1] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski, "Building rome in a day," in *ICCV*, 2009. 1

[2] R. Bhotika, D. Fleet, and K. Kutulakos, "A probabilistic theory of occupancy and emptiness," in *ECCV*, 2002. 1

[3] M. A. Brubaker, A. Punjani, and D. J. Fleet, "Building Proteins in a Day: Efficient 3D Molecular Reconstruction," in *Proc. CVPR*, 2015. 2

[4] 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. 1

[5] N. Grigorieff, "Frealign: high-resolution refinement of single particle structures," *J Struct Biol*, vol. 157, no. 1, p. 117–125, Jan 2007. 1

[6] J. Hsieh, *Computed Tomography: Principles, Design, Artifacts, and Recent Advances*. SPIE, 2003. 1

[7] K. N. Kutulakos and S. M. Seitz, "A theory of shape by space carving," *IJCV*, vol. 38, no. 3, pp. 199–218, 2000. 1

[8] 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. 2

[9] 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. 1

[10] 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. 2

[11] 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. 2

[12] 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. 1