3D *AB INITIO* MODELING IN CRYO-EM BY AUTOCORRELATION ANALYSIS

Eitan Levin

Advisors: Tamir Bendory, Nicolas Boumal, Joe Kileel, Amit Singer

ABSTRACT

Single-Particle Reconstruction (SPR) in Cryo-Electron Microscopy (cryo-EM) is the task of estimating the 3D structure of a molecule from a set of noisy 2D projections, taken from unknown viewing directions. Many algorithms for SPR start from an initial reference molecule, and alternate between refining the estimated viewing angles given the molecule, and refining the molecule given the viewing angles. This scheme is called iterative refinement. Reliance on an initial, user-chosen reference introduces model bias, and poor initialization can lead to slow convergence. Furthermore, since no ground truth is available for an unsolved molecule, it is difficult to validate the obtained results. This creates the need for high quality *ab initio* models that can be quickly obtained from experimental data with minimal priors, and which can also be used for validation. We propose a procedure to obtain such an *ab initio* model directly from raw data using Kam's autocorrelation method. Kam's method has been known since 1980, but it leads to an underdetermined system, with missing orthogonal matrices. Until now, this system has been solved only for special cases, such as highly symmetric molecules or molecules for which a homologous structure was already available. In this paper, we show that knowledge of just two clean projections is sufficient to guarantee a unique solution to the system. This system is solved by an optimization-based heuristic. For the first time, we are then able to obtain a low-resolution *ab initio* model of an asymmetric molecule directly from raw data, without 2D class averaging and without tilting. Numerical results are presented on both synthetic and experimental data.

Index Terms— cryo-EM, single particle reconstruction, Kam's method, autocorrelation analysis, *ab initio* modeling, orthogonal matrix retrieval, Riemannian optimization

1. INTRODUCTION

Cryo-EM is an increasingly popular method for determining the 3D structure of molecules, especially those that resist crystalization [15, 3, 9]. Advances in this technique were recognized by the 2017 Nobel Prize in Chemistry [1]. For SPR in cryo-EM, a sample containing many (ideally) identical molecules in unknown orientations are frozen in a sheet of ice. An electron microscope produces a top view of the sample in one image, called a micrograph, from which projection images of individual molecules are extracted in a process called particle picking. In order to limit radiation damage to the organic molecules caused by the electron beam, the electron dosage must be kept low, resulting in a low signal-to-noise ratio (SNR) in each of the projections. In addition, the images are affected by the Contrast Transfer Function (CTF) of the microscope, causing further aberrations. The goal is to estimate the 3D structure of the molecule from a large set of projections selected from multiple micrographs.

Typical approaches to SPR use iterative refinement procedures that start from an initial guess of the 3D structure, apply a low-pass filter, and then refine it by alternating between estimation of the viewing directions of the projections given the molecule and vice versa [23, 4, 18]. Since these algorithms solve a non-convex problem, the quality of their output as well as the speed of their convergence depend on the initialization, particularly at low SNR or with small particles [5, 7].

In contrast, *ab initio* methods do not require an initial model. Currently, few *ab initio* methods are available. The random conical tilt method [19] requires the molecule to have a strongly preferred orientation. Methods that do not involve tilting are either based on moments [22, 10] or common lines [25, 26]. However, these approaches typically fail to recover the 3D structure from non-averaged experimental images due to the low SNR.

We present a new method called *orthogonal matrix retrieval by projection matching*, based on Kam's autocorrelation analysis [13, 14]. Unlike the above mentioned methods for *ab initio* modeling, Kam's method completely sidesteps estimation of particle orientations. It only requires the covariance matrix of the projection images, which can be estimated accurately for any SNR given sufficiently many particle images. Kam's analysis recovers the expansion coefficients of the structure, up to a sequence of missing orthogonal matrices. It assumes the viewing directions are uniformly distributed over the sphere. Recently, there have been numerous attempts to apply Kam's method to XFEL [20, 21, 16] and to cryo-EM [5, 7]. Restrictingly, the first make either strong symmetry assumptions on the molecule or limit the rotations to a single axis, while the latter assume that the structure of a similar molecule is already available.

In this work, we apply Kam's method to resolve the molecular structure directly from raw experimental images

without estimating viewing directions, for the first time. We use the method of [6] to estimate the covariance matrix of the projections from raw data. We then recover the missing orthogonal matrices by matching to two clean or denoised images, via Riemannian optimization. The computational complexity of our algorithm is linear in the number of images. As an information-theoretic guarantee, we prove that 2D covariance together with merely two clean images uniquely determine the 3D molecular structure. For reproducibility, a Matlab implementation of our method is available at https://github.com/eitangl/kam_cryo.

The rest of this paper is organized as follows. In Section 2, we describe the image formation model in cryo-EM. In Section 3, we describe Kam's autocorrelation analysis and formulate the orthogonal matrix retrieval problem. Section 4 describes our procedure for solving the orthogonal matrix retrieval problem, which enables us to recover the molecular structure, and provides an information-theoretic guarantee. In Section 5, we show the performance of our method on synthetic and experimental datasets. Finally, in Section 6, we discuss possible extensions of the method for future work.

2. IMAGE FORMATION MODEL

Let $\phi : \mathbb{R}^3 \to \mathbb{R}$ be the Coulomb potential representing the molecular structure we wish to estimate. The j^{th} projection image $I_j : \mathbb{R}^2 \to \mathbb{R}$ is modeled as

$$I_j = H_j * \mathcal{P}_j[\phi] + \varepsilon_j, \quad j = 1, \dots, n. \tag{1}$$

Here, $H_j : \mathbb{R}^2 \to \mathbb{R}$ corresponds to the CTF affecting the j^{th} image by convolution, ε_j is noise and $\mathcal{P}_j$ is the tomographic projection operator given by

$$\mathcal{P}_j[\phi](x, y) = \int_{-\infty}^{\infty} \phi(R_j^T r)\, dz, \tag{2}$$

where $r = (x, y, z)^T$ are Cartesian coordinates and $R_j \in \mathrm{SO}(3)$ is the orientation of the j^{th} particle. This formation model is more neatly expressed in the Fourier domain. Owing to the Fourier-slice theorem [17, pp. 11], the 2D Fourier transform of a 2D projection image is the restriction of the 3D Fourier transform of ϕ to the plane passing through the origin perpendicular to the viewing direction. Denoting Fourier transforms by hats, we can rewrite the formation model as

$$\widehat{I}_j(k_x, k_y) = \widehat{H}_j \cdot \widehat{\phi}\big(R_j^T (k_x, k_y, 0)^T\big) + \widehat{\varepsilon}_j, \tag{3}$$

where k_x, k_y are Cartesian coordinates in 2D Fourier space.

3. KAM'S AUTOCORRELATION ANALYSIS

We assume that the structure ϕ is essentially compactly supported and bandlimited with bandlimit c. We expand the Fourier transform of the density in the eigenfunctions of the Laplacian with Dirichlet boundary conditions over the radius c ball in $\mathbb{R}^3$, working in spherical coordinates:

$$\widehat{\phi}(k, \theta, \varphi) = \sum_{l=0}^{L} \sum_{m=-l}^{l} \sum_{s=1}^{S(l)} a_{lms} Y_{lm}(\theta, \varphi) j_{ls}(k). \tag{4}$$

Here, Y_{lm} are the real spherical harmonics and

$$j_{ls}(k) = \frac{\sqrt{2}}{c^{3/2} |j_{l+1}(u_{l,s})|}\, j_l(u_{l,s} k / c), \tag{5}$$

where j_l is the spherical Bessel function of order l, the scalar $u_{l,s}$ is the s^{th} positive zero of j_l.

We shall assume a bandlimit c smaller than the Nyquist frequency and a finite expansion of the above form, since we focus on recovering a low-resolution version of the molecule, suitable for an *ab initio* estimate. The truncation limit $S(l)$ is chosen by the sampling criterion proposed in [7, Eq. (8)], enforcing essentially compact support in real space, and $S(l)$ is a monotonically decreasing function of l.

Our goal is to estimate the coefficients a_{lms}. In a seminal paper [13], Kam showed that the matrices

$$C_l(s_1, s_2) = \sum_{m=-l}^{l} a_{lms_1} \overline{a_{lms_2}}, \quad l = 0, \dots, L, \tag{6}$$

can be recovered directly from the noisy projections, provided that the viewing directions are uniformly distributed over the sphere.

Defining the $S(l) \times (2l+1)$ matrix of coefficients A_l indexed as $A_l(s, m) = a_{lms}$ for fixed l, the $S(l) \times S(l)$ matrices C_l in Eq. (6) satisfy the relation

$$C_l = A_l A_l^*, \tag{7}$$

where A^* denotes the Hermitian conjugate of A. Since the molecular density ϕ is real-valued, its Fourier transform is conjugate-symmetric, and hence the matrices A_l are purely real for even l, and purely imaginary for odd l. Therefore, Eq. (7) determines A_l uniquely up to an orthogonal matrix of size $(2l+1) \times (2l+1)$.

Formally, we take a Cholesky decomposition of the estimated C_l to obtain $S(l) \times (2l+1)$ matrices F_l. Accordingly, $A_l = F_l O_l$ for some unknown $(2l+1) \times (2l+1)$ orthogonal matrices O_l. This is the missing orthogonal matrix problem in Kam's method, which we aim to solve with minimal priors on the molecule. This would then allow us to recover the 3D structure.

4. ORTHOGONAL MATRIX RETRIEVAL BY PROJECTION MATCHING

We begin by noting that the matrix O_0 is just a sign ± 1, and can be easily recovered from the average image of the dataset.

Specifically, we take the radially-isotropic average of all the projections, and note that this average is determined only by the $l = 0$ component, proportional to $\sum_{s=1}^{S(0)} a_{00s} j_{0s}(k)$. This determines the coefficients for $l = 0$.

The main contribution of this paper is the observation that the remaining $\{O_l\}_{l=1}^L$ may be retrieved by matching to merely two clean or denoised projections. These projections can be obtained for example by using the Wiener filter-based method of [6] to denoise and CTF-correct individual projection images. To see how a known clean projection constrains the missing $\{O_l\}$, we write Eq. (4) in matrix form to get

$$\widehat{\phi}(\{O_l\}) = \sum_{l=0}^{L} j_l F_l O_l Y_l, \tag{8}$$

where matrices are indexed as $[j_l]_{k,s} = j_{ls}(k)$, $[Y_l]_{m,(\theta,\varphi)} = Y_{lm}(\theta,\varphi)$ and F_l is obtained from a Cholesky decomposition of C_l mentioned in Section 3.

Without loss of generality, the first clean projection is the restriction to the $k_x k_y$-plane of $\widehat{\phi}$, as the molecule can only be estimated up to a global rotation and reflection. Now, let the orientation of the particle in the second clean image be given by an unknown $R \in \mathrm{SO}(3)$. Since rotating real spherical harmonics of degree l may be expressed with matrix multiplication [12], the second clean image also imposes linear constraints on $\{O_l\}$. Writing $D_l^{(R)}$ for the $(2l+1)\times(2l+1)$ Wigner D-matrix of R in this irreducible representation of $\mathrm{SO}(3)$, the second projection imposes the constraints

$$\widehat{\phi}(\{O_l\}, R) = \sum_{l=0}^{L} j_l F_l O_l D_l^{(R)} Y_l. \tag{9}$$

From Rodrigues' formula for associated Legendre polynomials, restricting Y_{lm} to $\theta = \pi/2$ (the $k_x k_y$-plane) sets all the rows of Y_l for which $l \not\equiv m \pmod 2$ to zero. Thus the clean projections constrain every other column of $\{O_l\}$ and $\{O_l D_l^{(R)}\}$, respectively. It can be shown (proof omitted here due to space limitations) that under mild technical conditions these linear constraints in fact uniquely determine the orthogonal matrices $\{O_l\}$ and the rotation R, and hence the 3D structure itself.

In practice, given two clean or denoised images I_1^c, I_2^c, we begin by matching to each image separately. To do this, we assume both projections lie on the $k_x k_y$-plane corresponding to the 3×3 identity rotation matrix $\mathcal{I}_3$, let $\widehat{\phi}(\{O_l\})_{k_x k_y}$ denote the restriction of Eq. 8 to the $k_x k_y$ plane, and obtain estimates for every other column in two sets of orthogonal matrices:

$$\begin{aligned}
\{o_{l;1}\}_{l=1}^L &= \operatorname*{argmin}_{\substack{o_l \in \mathbb{R}^{(2l+1)\times(l+1)} \\ o_l^T o_l = \mathcal{I}_{l+1}}} ||\widehat{\phi}(\{o_l\})_{k_x k_y} - \widehat{I_1^c}||_F^2, \\
\{o_{l;2}\}_{l=1}^L &= \operatorname*{argmin}_{\substack{o_l \in \mathbb{R}^{(2l+1)\times(l+1)} \\ o_l^T o_l = \mathcal{I}_{l+1}}} ||\widehat{\phi}(\{o_l\})_{k_x k_y} - \widehat{I_2^c}||_F^2.
\end{aligned} \tag{10}$$

We estimate $\{o_{l;1}\}, \{o_{l;2}\}$ via optimization over the appropriate product of manifolds using Manopt [8]. Note that Riemannian gradient descent is only guaranteed to converge to critical points of the cost function [2]. However, for the purposes of *ab initio* modeling, our implementation performs satisfactorily, as seen empirically in Section 5.

Continuing the algorithm, we then merge results from the two images together. Writing O_l for the missing orthogonal matrices, taking the first image to have identity orientation and the second R, it follows that every other column of O_l should equal columns of $o_{l;1}$ while every other column of $O_l D_l^{(R)}$ should equal columns of $o_{l;2}$. We solve for R and $\{O_l\}$ by making these as consistent as possible. Formally, for each $l = 1, \ldots, L$, we form the matrices $D_l = \left[\widetilde{\mathcal{I}}_{2l+1} \,|\, \widetilde{D}_l^{(R)}\right]$ and $B_l = [o_{l;1} \,|\, o_{l;2}]$, where we denote by $\widetilde{A} \in \mathbb{R}^{(2l+1)\times(l+1)}$ the matrix obtained from $A \in \mathbb{R}^{(2l+1)\times(2l+1)}$ by taking only every other column including the first and last, and where $[X \,|\, Y]$ denotes the horizontal concatenation of matrices X and Y. We then solve the minimization

$$\min_{R\in\mathrm{SO}(3)} \sum_{l=1}^{L} \min_{O_l \in O(2l+1)} ||O_l D_l - B_l||_F^2. \tag{11}$$

This is done by densely sampling $R \in \mathrm{SO}(3)$ and noting that for R fixed the minimization $\min_{O_l\in O(2l+1)} ||O_l D_l - B_l||_F^2$ is an instance of the orthogonal Procrustes problem, which has a closed form solution via SVD of $B_l D_l^T$ [24]. Finally, we further refine our estimates of $\{O_l\}$ and R using Manopt.

5. NUMERICAL EXAMPLES

We begin with results on a synthetic dataset consisting of 5×10^4 noisy projections of size 109×109 with SNR $= 1/10$ from uniformly random viewing directions of the 70S ribosome with P-site tRNA, available in the Protein Data Bank (EMDB) as EMD-5360. The images are divided into 100 defocus groups, and are centered. The bandlimit is assumed to be $c = 1/4$ (half the Nyquist frequency), and the truncation for the expansion of the structure is set to $L = 10$. The two images for the reconstruction were chosen uniformly at random. The reconstruction results are presented in Fig. 1 (a) and (c), where we also show the Fourier Shell Correlation (FSC) [11] of our reconstruction with the low-resolution ground truth. The resolution of the reconstruction is 19 Å using the FSC $= 0.5$ criterion.

We also present results on an experimental dataset consisting of 3.5×10^5 projections of size 330×330 of the yeast mitochondrial ribosome, available in the Electron Microscopy Public Image Archive (EMPIAR) as EMPIAR-10107, out of which we chose 2×10^5 random projections for implementation reasons. We pre-processed the data only by whitening the projections using the method described in [6, Sec. 2.2], and directly applied the method of [6] to estimate the covariance matrix of the projections, from which we obtained $\{C_l\}$.

Here we set $c = 1/4$ and $L = 7$. Once again, the two projections for the reconstruction were chosen randomly. We ran the algorithm on a machine with 60 cores, running at 2.3 GHz, with total RAM of 1.5TB. The pre-processing then took $\sim$5 hours, while the reconstruction itself took $\sim$15 minutes. The resolution of the reconstruction is 89 Å using the FSC $= 0.5$ criterion. For the ground truth, we took the RELION reconstruction available as EMD-3551, and expanded it in a finite expansion of the form Eq. 4 with the same truncation as for our own reconstruction. The original EMD-3551 is presented alongside its finite expansion for comparison, slightly low-passed with a Gaussian filter to remove noise artifacts present in the reconstruction.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented a new method to obtain *ab initio* low-resolution 3D molecular structures directly from raw cryo-EM data. We rely on Kam's autocorrelation analysis, which recovers the expansion coefficients of the molecule from the covariance matrix of its projections, up to a set of missing orthogonal matrices. We retrieve these matrices using two clean or denoised projections, and showed that two clean projections determine the structure under mild assumptions. Finally, we demonstrated the performance of the method on both synthetic and experimental datasets. This is the first application of Kam's method to *ab initio* modeling of asymmetric molecules from raw experimental data without any averaging.

Nevertheless, we observe in practice that our method is only capable of recovering a low-resolution version of the molecule. While sufficient to initialize iterative refinement algorithms and validate their output, we would like to improve the method to obtain higher-resolution reconstructions, while keeping the computational cost low. We believe our resolution limitation stems from several features present in real datasets, but which our formulation currently ignores. First, because individual projection images are picked from extremely noisy micrographs, the projections may not be centered. Second, while we assume all the molecules in the sample are identical, in practice they may appear in different conformations, and so the projections would come from several different molecules. Third, Kam's method as stated here assumes the viewing angles of the projections are uniformly distributed over the sphere. In practice, molecules have preferred orientations, which skews the distribution of viewing angles. For symmetric molecules, it can be shown that a single clean image is sufficient to determine the missing orthogonal matrices, in which case our method may be simplified and improved. We intend to extend Kam's method to account for these features. Finally, it may be possible to avoid the need for clean projections in the first place by using higher-order correlations in addition to the covariance matrix, as originally suggested by Kam [13].

7. ACKNOWLEDGEMENTS

We thank Joakim Andén, Tejal Bhamre, and Yoel Shkolniskly for productive discussions and help with the code.

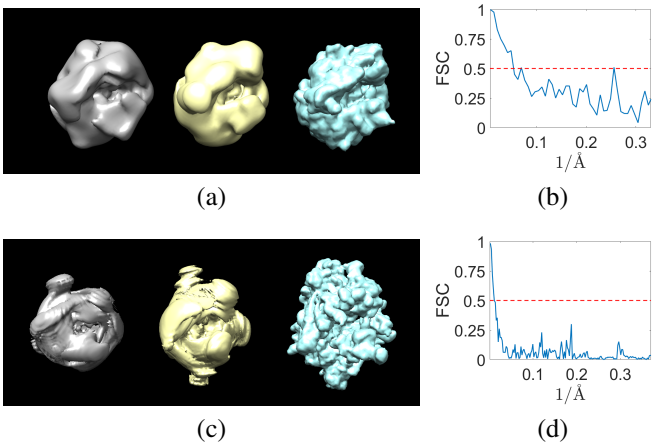

Fig. 1. *Reconstruction results.* (a) Synthetic data results: the reconstruction (grey), the ground truth (yellow), and the original (blue). (b) FSC curve for synthetic data. (c) Raw data results for yeast mitochondrial ribosome (EMPIAR-10107): the reconstruction (grey), the ground truth, taken as the corresponding low-resolution EMD-3551 (yellow) and the original EMD-3551, reconstructed using RELION, slightly low-passed with a Gaussian filter to remove noise artifacts (blue). (d) FSC curve for raw data. Note that the reconstruction was measured with respect to the low-resolution ground truth in both (b) and (d).

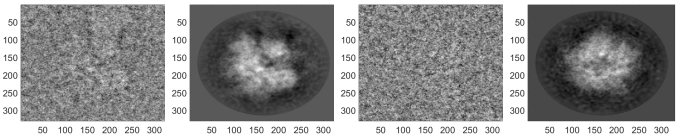

Fig. 2. *Images used for raw data reconstruction.* (a) and (c) are the original noisy images and (b) and (d) are the corresponding denoised images used for the optimization.

8. REFERENCES

[1] https://www.nobelprize.org/nobel_prizes/chemistry/laureates/2017/.

[2] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2008.

[3] X. Bai, G. McMullan, and S. H. W. Scheres. How cryo-EM is revolutionizing structural biology. *Trends Biochem. Sci.*, 40(1):49 – 57, 2015.

[4] A. Barnett, L. Greengard, A. Pataki, and M. Spivak. Rapid Solution of the Cryo-EM Reconstruction Problem by Frequency Marching. *SIAM J. Imaging Sci.*, 10(3):1170–1195, 2017.

[5] T. Bhamre, T. Zhang, and A. Singer. Orthogonal matrix retrieval in cryo-electron microscopy. In *2015 IEEE 12th Int. Symp. Biomed. Imaging*, pages 1048–1052. IEEE, 2015.

[6] T. Bhamre, T. Zhang, and A. Singer. Denoising and covariance estimation of single particle cryo-EM images. *J. Struct. Biol.*, 195(1):72–81, 2016.

[7] T. Bhamre, T. Zhang, and A. Singer. Anisotropic twicing for single particle reconstruction using autocorrelation analysis. *ArXiv e-prints*, 2017, 1704.07969.

[8] N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre. Manopt, a Matlab Toolbox for Optimization on Manifolds. *J. Mach. Learn. Res.*, 15:1455–1459, 2014.

[9] J. Frank. *Three-Dimensional Electron Microscopy of Macromolecular Assemblies: Visualization of Biological Molecules in Their Native State*. Oxford University Press, New York, 2006.

[10] A. B. Goncharov. Integral geometry and three-dimensional reconstruction of randomly oriented identical particles from their electron microphotos. *Acta Appl. Math.*, 11(3):199–211, 1988.

[11] G. Harauz and M. van Heel. Exact filters for general geometry three dimensional reconstruction. *Optik (Stuttg).*, 73:146–156, 1986.

[12] J. Ivanic and K. Ruedenberg. Rotation Matrices for Real Spherical Harmonics. Direct Determination by Recursion. *J. Phys. Chem.*, 100(15):6342–6347, 1996.

[13] Z. Kam. The reconstruction of structure from electron micrographs of randomly oriented particles. *J. Theor. Biol.*, 82(1):15 – 39, 1980.

[14] Z. Kam and I. Gafni. Three-dimensional reconstruction of the shape of human wart virus using spatial correlations. *Ultramicroscopy*, 17(3):251–262, 1985.

[15] W. Kühlbrandt. The resolution revolution. *Science*, 343:1443–1444, 2014.

[16] R. P. Kurta, J. J. Donatelli, C. H. Yoon, P. Berntsen, J. Bielecki, B. J. Daurer, H. DeMirci, P. Fromme, M. F. Hantke, F. R. Maia, et al. Correlations in scattered x-ray laser pulses reveal nanoscale structural features of viruses. *Physical review letters*, 119(15):158102, 2017.

[17] F. Natterer. *The Mathematics of Computerized Tomography*. Society for Industrial and Applied Mathematics, 2001.

[18] A. Punjani, J. L. Rubinstein, D. J. Fleet, and M. A. Brubaker. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Meth.*, 14(3):290–296, 2017.

[19] M. Radermacher, T. Wagenknecht, A. Verschoor, and J. Frank. Three-dimensional structure of the large ribosomal subunit from Escherichia coli. *EMBO J.*, 6(4):1107–14, 1987.

[20] D. K. Saldin, H.-C. Poon, P. Schwander, M. Uddin, and M. Schmidt. Reconstructing an icosahedral virus from single-particle diffraction experiments. *Opt. Express*, 19(18):17318, aug 2011.

[21] D. K. Saldin, V. L. Shneerson, M. R. Howells, S. Marchesini, H. N. Chapman, M. Bogan, D. Shapiro, R. A. Kirian, U. Weierstall, K. E. Schmidt, et al. Structure of a single particle from scattering by many particles randomly oriented about an axis: toward structure solution without crystallization? *New J. Phys.*, 12(3):035014, 2010.

[22] D. B. Salzman. A method of general moments for orienting 2D projections of unknown 3D objects. *Comput. Vision Graph.*, 50:129–156, 1990.

[23] S. H. W. Scheres. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.*, 180(3):519 – 530, 2012.

[24] P. H. Schnemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

[25] B. Vainshtein and A. Goncharov. Determination of the spatial orientation of arbitrarily arranged identical particles of unknown structure from their projections. *Soviet Physics Doklady*, 31:278, 1986.

[26] M. van Heel. Angular reconstitution: A posteriori assignment of projection directions for 3d reconstruction. *Ultramicroscopy*, 21(2):111 – 123, 1987.