Statistical characterization of ensembles of symmetric virus particles: 3-D stochastic signal reconstruction from electron microscope images *

Nan Xu[1], Peter C. Doerschuk[2]

Abstract— **Stochastic models of nano-biomachines have been studied by 3-D reconstruction from cryo electron microscopy images in recent years. The image data is the projection of many heterogeneous instances of the object under study (e.g., a virus). Initial reconstruction algorithms require different instances of the object, while still heterogeneous, to have the same symmetry. This paper presents a maximum likelihood reconstruction approach which allows each object to lack symmetry while constraining the *statistics* of the ensemble of objects to have symmetry. This algorithm is demonstrated on bacteriophage HK97 and is contrasted with the former algorithm. Reconstruction results show that the proposed algorithm provides estimates that make more biological sense.**

I. INTRODUCTION

In virology, there is a large class of plant and animal viruses called "spherical" viruses, in which the particle has a shell of protein, called a "capsid", surrounding a cavity containing the viral genome. Typical sizes and molecular weights of the virus particles are 10^2–10^3Å and 10MDa. The capsid is constructed of many repetitions of the same peptide molecule in geometric arrays similar to human-constructed geodesic domes. The geometry of these spherical viruses is important to their lifecycle. Cryo electron microscopy (cryo EM), which has become an important technique for determining the geometry of a particle, leads to 3-D image reconstruction problems for these biological nanomachines. Specifically, $10^3 - 10^6$ virus particles are flash frozen to cryogeneic temperatures and imaged. Each image is basically a highly-noisy (SNR < 0.1) 2-D projection of the 3-D electron scattering intensity distribution of the particle. Only one such image can be taken per particle and its projection orientation is unknown. Therefore, the information from many such images is fused to compute a 3-D reconstruction [1].

In x-ray crystallography, the capsid of spherical virsus has a geometric symmetry, of which the most common is icosahedral symmetry–the symmetry group of an icosahedron– which is invariant under a total of 60 rotational symmetry operations. Standard reconstruction approaches [2]–[5] assume that the particles are either identical or are members of a small number of discrete classes where every particle in a class is identical. In [6], [7], a reconstruction approach was presented, which includes both discrete classes and continuous variability of the particles within each class. Within one class, the basic assumption on the virus particles [2]–[5] is

that each instance of the virus particle is identical and exactly obeys the icosahedral symmetry. An improved assumption introduced in [6], [7] is that instances are different from each other (for example, due to the inherent flexibility of such a large molecular complex) but still each instance exactly obeys the icosahedral symmetry.

In contrast to this previous work, this paper considers a more realistic and sophisticated view: The different instances of the particle are different and it is the *statistics* of the particle that obey the symmetry not the individual instances. In particular, the mean and covariance of the 3-D electron scattering intensity distribution of the particle obey [8]:

$$\bar{\rho}(R_g^{-1}\mathbf{x}) = \bar{\rho}(\mathbf{x}) \text{ and } C_\rho(R_g^{-1}\mathbf{x}_1, R_g^{-1}\mathbf{x}_2) = C_\rho(\mathbf{x}_1, \mathbf{x}_2), \tag{1}$$

where $\bar{\rho}(\cdot)$ is the mean function and $C_\rho(\cdot, \cdot)$ is the covariance function of the electron scattering intensity $\rho(\cdot)$, $R_g \in \mathbb{R}^{3\times 3}$ is the gth rotation matrix of the symmetry group, and $\mathbf{x}, \mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^3$ are real-space coordinates.

Conditional on class membership, each particle can be described as a linear combination of basis functions where the weights in the linear combination are Gaussian random variables [6], [7]:

$$\rho(\mathbf{x}) = \sum_\zeta F_\zeta(\mathbf{x}) c_\zeta \tag{2}$$

where $F_\zeta(\mathbf{x})$ is the basis function and c_ζ is the weight. Because $\rho(\cdot)$ is real, it is convenient to have real-valued basis functions and weights. To describe the situation in which particles have non symmetric structures but symmetric statistics, a complete orthonormal set of real-valued basis functions with specific rotational properties under the operation of the icosahedral group is required (Section II).

The maximum likelihood (ML) reconstruction algorithm can estimate the mean vector and covariance matrix of the vector of weights c_ζ, which is a generalization of classical ML Gaussian mixture parameter estimation [9]. After choosing appropriate basis functions, the estimates of the mean vector and covariance matrix need to be constrained in order to realize the two statistical symmetries (Section III). The reconstruction results are demonstrated and discussed in Section IV.

II. REAL-VALUED BASIS FUNCTIONS

A complete orthonormal set of real-valued basis functions is needed to describe both the angular and the radial behaviors. Each real-valued basis function $F_\zeta(\mathbf{x})$ is a product of an angular basis function and a radial basis function, and both angular basis and radial basis functions are complete real-valued orthonormal sets on their respective spaces.

[1]Nan Xu is Ph.D. student of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14853, USA nx25@cornell.edu

[2]Peter C. Doerschuk is Professor of Biomedical Engineering, Electrical and Computer Engineering, Cornell University, Ithaca, NY 14853, USA pd83@cornell.edu

The real-valued angular basis functions [8] characterize the behavior around the surface of a sphere and include the symmetry properties of the particle, since the symmetries are all rotational. A key concept used in this approach is the group irreducible representations (irred reps), which are sets of matrices that are homomorphic under matrix multiplication to the elements of an abstract group [10]. Let the matrices in the pth irred rep be denoted by $\Gamma^p(g)$ where $p \in \{1, \ldots, N_{\text{rep}}\}$ (N_{rep} is the number of irred reps) and $g \in \{1, \ldots, N_g\}$ (N_g is the number of elements in the group). For the icosahedral group, $N_{\text{rep}} = 5$ and $N_g = 60$. The dimension of the matrices in the pth irred rep are $d_p \times d_p$ where, for the icosahedral group, $d_p = 1, 3, 3, 4, 5$ for $p = 1, 2, 3, 4, 5$, respectively. The resulting angular basis functions described as linear combinations of spherical harmonics of fixed index $l \in \mathbb{N}$, denoted by $I_{p;l,n}(\mathbf{x}/x)$ for $n \in \{1, \ldots, N_{p;l}\}$, have the following properties.

1) Each $I_{p;l,n}$ is a d_p-dimensional real-valued vector function, i.e., $I_{p;l,n} \in \mathbb{R}^{d_p}$.
2) The $I_{p;l,n}$ functions are orthonormal on the surface of the sphere.
3) The subspace of square integrable functions on the surface of the sphere defined by spherical harmonics of index l, contains a set of $I_{p;l,n}$ functions with a total of $2l + 1$ components.
4) The family of $I_{p;l,n}$ is a complete basis for square integrable functions on the surface of the sphere.
5) Each $I_{p;l,n}$ function has a specific transformation property under rotations from the icosahedral group [10, p. 20], in particular,

$$I_{p;l,n}(R_g^{-1}\mathbf{x}/x) = (\Gamma^p(g))^T I_{p;l,n}(\mathbf{x}/x) \qquad (3)$$

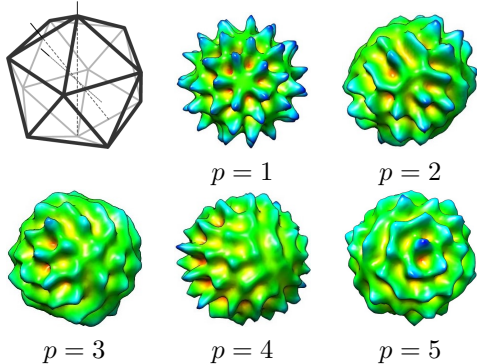where ^{T} is transpose not Hermitian transpose.

Fig. 1: An icosahedron with one of each type of symmetry axis (2-, 3-, and 5-fold) shown and example angular basis functions with $l = 15$ and $p \in \{1, \ldots N_{\text{rep}}\}$. The surfaces of 3-D objects are defined by $\xi(\mathbf{x}) = 1$ for $x \leq \kappa_1 + \kappa_2 I_{p;l,n}(\mathbf{x}/x)$ and 0 otherwise, where κ_1 and κ_2 are chosen so that $\kappa_1 + \kappa_2 I_{p;l,n}(\mathbf{x}/x)$ varies between 0.5 and 1. These 3-D surfaces are visualized by UCSF Chimera where the color indicates the distance from the center of the object.

Examples of $I_{p,j;l,n}$ functions for $l = 15$ are visualized in Figure 1. Note that the $p = 1$ exhibits all of the symmetries of an icosahedron. The previous algorithm, which used the $I_{p=1,j;l,n}$ angular basis functions only [6], [11], requires each

virus object to obey the icosahedral symmetry. The range of behavior of a non-symmetric particle that leads to symmetric statistics is illustrated by the $p \in \{2, \ldots, 5\}$ plots.

Let $I_{p,j;l,n}$ be the components of $I_{p;l,n}$ for $j \in \{1, \ldots, d_p\}$. The real-valued basis function being used is $F_{p,j;l,n,q}(\mathbf{x}) = I_{p,j;l,n}(\mathbf{x}/x)h_{l,q}(x)$, where the radial basis functions $h_{l,q}(x)$ for $q \in \{1, 2, \ldots\}$ are exactly the family of Bessel functions used in [12], which form a complete orthonormal set on $\mathbb{R}^+ \cup \{0\}$.

III. STATISTICAL CHARACTERIZATION OF SYMMETRICAL PARTICLES

The symmetrical virus particles are statistically characterized by Eq. 1. The constraints on the mean vector and covariance matrix of the Gaussian weights $c_{p,j;l,n,q}$ in Eq. 2 induced by these two symmetry requirements are derived and applied to the maximum likelihood reconstruction algorithm.

Specifically, to achieve $\bar{\rho}(R_g^{-1}\mathbf{x}) = \bar{\rho}(\mathbf{x})$, a constraint on the mean of $c_{p,j;l,n,q}$, denoted by $\bar{c}_{p,j;l,n,q}$, is required:

$$\bar{c}_{p,j;l,n,q} = \bar{c}_{l,n,q}\delta_{p,1}. \qquad (4)$$

Note that because $p = 1$ implies $j = 1$ [7], j is not needed on the right-hand side of Eq. 4.

The second order statistics of $c_{p,j;l,n,q}$, in particular, the covariance matrix V with components $V_{p_1,j_1;l_1,n_1,q_1:p_2,j_2;l_2,n_2,q_1}$, are constrained by the symmetry on the covariance function (Eq. 1). After expanding $C_\rho(R_g^{-1}\mathbf{x}, R_g^{-1}\mathbf{x}') = C_\rho(\mathbf{x}, \mathbf{x}')$ by applying the orthonormal expansion (Eq. 2) and the icosahedral symmetry property (Eq. 3), the simplified result becomes a linear system in the covariance matrix V, where the coefficients are the irred rep matrices $\Gamma^p(g)$. The solution of the covariance matrix V is constrained by the group properties of the irred rep matrices $\Gamma^p(g)$. In particular, Schur's Lemma [10, Theorem I and II, Section 4.5, p. 80] and related results imply that the V matrix has the structure

$$\begin{aligned} V_{p_1,j_1;l_1,n_1,q_1:p_2,j_2;l_2,n_2,q_2} \\ = \delta_{p_1,p_2}\delta_{j_1,j_2}v_{p_1}(l_1, n_1, q_1 : l_2, n_2, q_2) \end{aligned} \qquad (5)$$

where $v_{p_1}(l_1, n_1, q_1 : l_2, n_2, q_2)$ is an arbitrary function for the six indices $q_1, q_2 \in \{1, 2, \ldots\}$, $p_1, p_2 \in \{1, \ldots, N_{\text{rep}}\}$, $j_1 \in \{0, 1, \ldots d_{p_1}\}$, $j_2 \in \{0, 1, \ldots d_{p_2}\}$ $l_1, l_2 \in \{0, 1, \ldots\}$, $n_1 \in \{0, \ldots, N_{p_1;l_1} - 1\}$, and $n_2 \in \{0, \ldots, N_{p_2;l_2} - 1\}$.

Now suppose that V is diagonal. Let $\nu = \text{diag}(V)$. Then $\nu_{p,j;l,n,q} = \alpha_{p;l,n,q}$, where $\alpha_{p;l,n,q}$ has arbitrary dependence on the four indices. In other words, ν does not depend on the j index. In previous algorithms [6], [11], which require every individual particle to exactly obey the symmetry, only the $p = 1$ functions are allowed. That means that only the case of $d_{p=1} = 1$ is used, so that $j = 1$ and $\Gamma^{p=1}(g) = 1$. In this case, $\nu_{p=1,j=1;l,n,q} = \alpha_{p=1;l,n,q}$, so that the covariance matrix V is not constrained.

Following [7], a maximum likelihood estimator is used to determine the mean vector and covariance matrix of $c_{p,j;l,n,q}$. The maximum likelihood estimator is computed by an expectation-maximization algorithm in which the nuisance parameters are the projection direction orientation for

each image and the class of particle shown in each image. In our problem, two quantities, the mean and the covariance of the weights $c_{p,j:l,n,q}$, are updated alternatively in the expectation-maximization algorithm. A previous algorithm, which used a subset of the proposed angular basis functions (i.e., just $p = 1$) [6], [11], used Matlab's `fmincon` (option "trust-region-reflective") with symbolic cost, gradient of the cost, and Hessian of the cost. We desired to modify the software to include the full set of basis functions described in Section II. The constraint on the mean vector (Eq. 4) requires its entries to be zero if $p \neq 1$. The constraints on the covariance matrix V (Eq. 5) require certain matrix elements to be equal. To implement the above constraints, the method of computing the gradient and Hessian of the cost is modified based on the chain rule: if f is the cost then

$$\frac{\partial f}{\partial \zeta_{p;l,n,q}} = \sum_{j=1}^{d_p} \frac{\partial f}{\partial \nu_{p,j;l,n,q}},$$

$$\frac{\partial^2 f}{\partial \zeta_{p_1;l_1,n_1,q_1} \partial \zeta_{p_2;l_2,n_2,q_2}} = \sum_{j_1=1}^{d_{p_1}} \sum_{j_2=1}^{d_{p_2}} \frac{\partial^2 f}{\partial \nu_{p_1,j_1;l_1,n_1,q_1} \partial \nu_{p_2,j_2;l_2,n_2,q_2}}.$$

These equations compute the necessary gradient and Hessian in terms of larger vectors and matrices which are then reduced in size. While this approach fits the software easily, more efficient approaches may be possible.

IV. RECONSTRUCTION OF VIRUS PARICLE

The proposed algorithm, heterogeneous reconstruction with statistical icosahedral symmetry (HRSIS), is compared with a previous algorithm, heterogeneous reconstructions with individual icosahedral symmetry (HRIIS) [6], [11]. The proposed algorithm (HRSIS) strictly contains the previous algorithm (HRIIS) as a special case, because HRIIS uses only the $p = 1$ angular basis functions and has no constraints on the mean vector or covariance matrix of the weights, whereas HRSIS uses all $p \in \{1, 2, 3, 4, 5\}$ angular basis functions and has the constraint on the mean vector (Eq. 4) and the covariance matrix (Eq. 5).

These two reconstruction algorithms were tested on a Virus Like Particle (VLP) derived from bacteriophage HK97 Prohead I^{+pro}. Specifically, the VLP is bacteriophage minus the bacteriophage's tail leaving only the icosahedrally symmetric capsid. The average outer radius of the capsid is 254Å. A stack of 1200 cryo-EM images was randomly selected. Each 2-D cryo-EM image measuring 200×200 pixels with a pixel size of 2.76Å contains one bacteriophage HK97 Prohead I^{+pro} particle. Following the preprocessing procedure in Refs. [6], [7], these 1200 cryo-EM images were transformed into 1200 2-D images in reciprocal space, which became the input to reconstruction algorithms.

The reconstruction was computed in two steps. First, the Gaussian weights $c_{p,j;l,n,q}$ in Eq. 2 were computed by a homogeneous reconstruction algorithm described in [7] which used only the $p = 1$ angular basis functions and had the covariance $V = 0$. Second, using the computed Gaussian weights as the initial condition, the new weights $c_{p,j;l,n,q}$ and the electron scattering intensity $\rho(\mathbf{x})$ were computed by heterogeneous reconstruction algorithms, HRIIS and HRSIS.
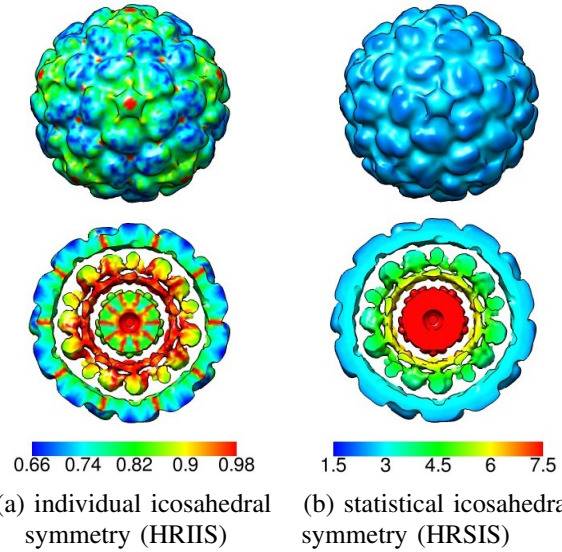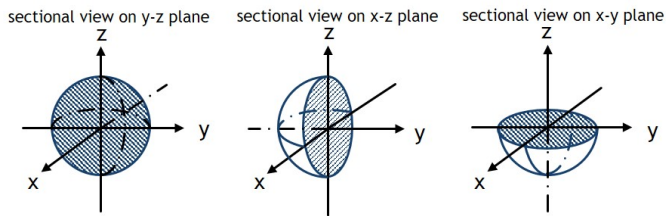

(a) individual icosahedral symmetry (HRIIS) (b) statistical icosahedral symmetry (HRSIS)

Fig. 2: 3-D reconstructions of $\bar{\rho}$ for HK97 Prohead I^{+pro} (different colormap). The shape is a surface of constant intensity (8×10^{-5}) of $\bar{\rho}(\mathbf{x})$ colored by the standard deviation $s_\rho(\mathbf{x})$, which is visualized by UCSF Chimera [13]. The HRIIS and HRSIS reconstructions use different color maps. All markings are scaled by 10^{-3}.

From the mean vector and covariance matrix of $c_{p,j;l,n,q}$, we are able to compute the mean function [7, Eq. 16], $\bar{\rho}(\mathbf{x}) = E[\rho(\mathbf{x})]$, and the covariance function [7, Eq. 18], $C_\rho(\mathbf{x}_1, \mathbf{x}_2) = E[(\rho(\mathbf{x}_1) - \bar{\rho}(\mathbf{x}_1))(\rho(\mathbf{x}_2) - \bar{\rho}(\mathbf{x}_2))]$, of the electron scattering intensity $\rho(\mathbf{x})$ of the particle. The standard deviation $s_\rho(\mathbf{x}) = \sqrt{C_\rho(\mathbf{x}, \mathbf{x})}$, which has the same units as $\bar{\rho}(\mathbf{x})$, is also computed for visualization purposes. In Fig. 2, $\bar{\rho}(\mathbf{x})$ and $s_\rho(\mathbf{x})$ estimated by HRIIS and HRSIS are jointly visualized. Specifically, a 3-D surface of constant value of $\bar{\rho}(\mathbf{x})$ is colored by the value of $s_\rho(\mathbf{x})$. In Fig. 3, three perpendicular cross sections of the standard deviation function $s_\rho(\mathbf{x})$ estimated by HRIIS and HRSIS are shown.

Two primary distinctions are observed. First, HRIIS gives a covariance function that is organized in radially-directed rays (Fig. 2(a)), whereas HRSIS gives a covariance function that is organized in *annular* structures (Fig. 2(b)). Such an annular structure, as is obtained by HRSIS, matches the physical structure of the particle: the outer protein shell and the inner core of nucleic acid. This makes more biological sense than the estimates from HRIIS. Such a distinction can also be seen in Fig. 3, in which the second row has dark blue radial lines, whereas the third row shows the detailed annular structure of the virus. In the previous analyses using HRIIS [14], features like high values located along radially directed lines were averaged out leaving interpretable information, but as we try to increase the spatial detail of our interpretation, such behavior is difficult to understand.

Second, HRSIS gives a covariance function with larger dynamic range (0–18.37×10^{-3}) than the earlier HRIIS (0–5.28×10^{-3}). This can be clearly seen in Fig. 3 where the same color map is used for both HRSIS and HRIIS. The standard deviation estimated by HRIIS is essentially pure blue at the lowest level of the color map, while HRSIS

Cross sectional view geometry
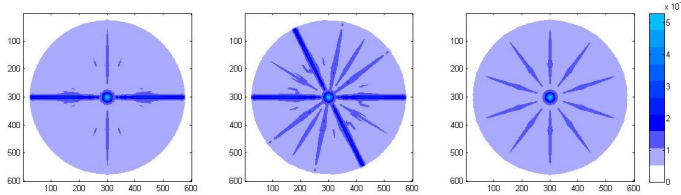
Individual icosahedral symmetry (HRIIS)
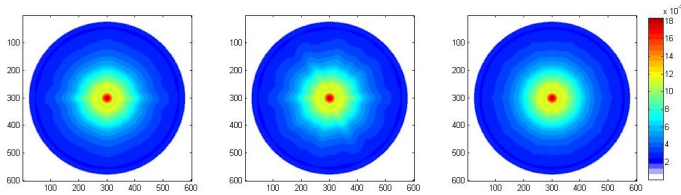
Statistical icosahedral symmetry (HRSIS)

Fig. 3: Cross section of the standard deviation function $s_\rho(\mathbf{x})$ for HK97 Prohead I$^{+\mathrm{pro}}$ using both HRISS and HRSIS and displayed with a common color map.

contributes to a broader range of colors. This is due to the fact that HRSIS uses a larger family of basis functions than is used by HRIIS and therefore can represent features that HRIIS averages away.

HK97 [15] is a bacteriophage with tens of protein constituents that self assemble and then undergo a complicated maturation process leading to an increase in diameter of the particle by about 100Å. Most structural biology is performed on VLPs which have subsets of the constituents, are not infectious virus particles, and are produced in bacterial hosts. In collaboration with Profs. J. E. Johnson (TSRI) and D. Veesler (U. Wash.) we are comparing two VLPs. Both VLPs have the capsid but lack the tail of the native particle. The first is basically just the 420 copies of the capsid protein and stops maturation at the Prohead I phase because it cannot cleave the capsid peptides which is the next step in maturation. The second is basically just the 420 copies plus about 60 copies of the protease which cleaves the capsid peptide. However, the protease is defective so maturation stops at the same Prohead I phase. While maturation stops at the same phase in both particles, the particles are quite different because the protease (located on the inner surface of the capsid) stabilizes the capsid [14]. The biological goal of these calculations is to understand how the binding of the protease on the inner surface of the capsid influences the entire capsid via understanding the covariance function $C_\rho(\mathbf{x}_1, \mathbf{x}_2)$ for $\mathbf{x}_1 \neq \mathbf{x}_2$ and especially with the two locations at different distances from the center of the particle. This could not be achieved previously [14] because the covariance

function $C_\rho(\mathbf{x}_1, \mathbf{x}_2)$ of $\rho(\cdot)$ at different locations could not be interpreted do to the limited number of angular basis functions that were used.

REFERENCES

[1] Grant J. Jensen, Ed., *Cryo-EM, Parts A–C*, vol. 481–483 of *Methods in Enzymology*, Elsevier Inc., 2010.

[2] Peter C. Doerschuk and John E. Johnson, "*Ab initio* reconstruction and experimental design for cryo electron microscopy," *IEEE Transactions on Information Theory*, vol. 46, no. 5, pp. 1714–1729, Aug. 2000, http://dx.doi.org/10.1109/18.857786.

[3] Zhye Yin, Yili Zheng, Peter C. Doerschuk, Padmaja Natarajan, and John E. Johnson, "A statistical approach to computer processing of cryo electron microscope images: Virion classification and 3-D reconstruction," *J. Struct. Biol.*, vol. 144, no. 1/2, pp. 24–50, 2003, http://www.ncbi.nlm.nih.gov/pubmed/14643207.

[4] Sjors H. W. Scheres, Haixiao Gao, Mikel Valle, Gabor T. Herman, Paul P. B. Eggermont, Joachim Frank, and Jose-Maria Carazo, "Disentangling conformational states of macromolecules in 3D-EM through likelihood optimization," *Nature Methods*, vol. 4, no. 1, pp. 27–29, Jan. 2007.

[5] Junghoon Lee, Peter C. Doerschuk, and John E. Johnson, "Exact reduced-complexity maximum likelihood reconstruction of multiple 3-D objects from unlabeled unoriented 2-D projections and electron microscopy of viruses," *IEEE Transactions on Image Processing*, vol. 16, no. 11, pp. 2865–2878, Nov. 2007, http://www.ncbi.nlm.nih.gov/pubmed/18092587.

[6] Qiu Wang, Tsutomu Matsui, Tatiana Domitrovic, Yili Zheng, Peter C. Doerschuk, and John E. Johnson, "Dynamics in cryo EM reconstructions visualized with maximum-likelihood derived variance maps," *J. Struct. Biol.*, vol. 181, no. 3, pp. 195–206, Mar. 2013.

[7] Yili Zheng, Qiu Wang, and Peter C. Doerschuk, "3-D reconstruction of the statistics of heterogeneous objects from a collection of one projection image of each object," *J. Opt. Soc. Am. A*, vol. 29, no. 6, pp. 959–970, June 2012.

[8] Nan Xu and Peter C. Doerschuk, "Reconstruction for stochastic 3-d signals with symmetric statistics in noise: Electron microscopy of virus particles," in *2015 IEEE International Conference on Image Processing, ICIP 2015, Quebec City, QC, Canada, September 27-30, 2015*, 2015, pp. 1444–1447.

[9] Richard A. Redner and Homer F. Walker, "Mixture densities, maximum likelihood and the EM algorithm," *SIAM Review*, vol. 26, no. 2, pp. 195–239, Apr. 1984.

[10] J. F. Cornwell, *Group Theory in Physics*, vol. 1, Academic Press, London, 1984.

[11] Jinghua Tang, Bradley M. Kearney, Qiu Wang, Peter C. Doerschuk, Timothy S. Baker, and John E. Johnson, "Dynamic and geometric analyses of *Nudaurelia capensis* ω virus maturation reveal the energy landscape of particle transitions," *J. Molecular Recognition*, vol. 27, no. 4, pp. 230–237, 10 February 2014, http://www.ncbi.nlm.nih.gov/pubmed/24591180.

[12] Yibin Zheng and Peter C. Doerschuk, "3D image reconstruction from averaged Fourier transform magnitude by parameter estimation," *IEEE Transactions on Image Processing*, vol. 7, no. 11, pp. 1561–1570, Nov. 1998.

[13] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin, "UCSF Chimera—A visualization system for exploratory research and analysis," *J. Comput. Chem.*, vol. 25, no. 13, pp. 1605–1612, 2004.

[14] Yunye Gong, David Veesler, Peter C Doerschuk, and John E Johnson, "Effect of the viral protease on the dynamics of bacteriophage HK97 maturation intermediates characterized by variance analysis of cryo em particle ensembles," *Journal of structural biology*, 2015.

[15] David Veesler and John E. Johnson, "Virus maturation," *Annual Review of Biophysics*, vol. 41, no. 1, pp. 473–496, 2012, PMID: 22404678.