

Symmetry-Constrained 3-D Interpolation of Viral X-Ray Crystallography Data

Yibin Zheng, *Member, IEEE*, Peter C. Doerschuk, *Member, IEEE*, and John E. Johnson

Abstract—A three-dimensional (3-D) interpolation problem that is important in viral X-ray crystallography is considered. The problem requires new methods because the function is known to have icosahedral symmetry, the data is corrupted by experimental errors and therefore lacks the symmetry, the problem is 3-D, the measurements are irregularly spaced, and the number of measurements is large (10^4). A least-squares approach is taken using two sets of basis functions: the functions implied by a minimum-energy bandlimited exact interpolation problem and a complete orthonormal set of bandlimited functions. A numerical example of the Cowpea Mosaic Virus is described.

I. INTRODUCTION

VIRUSES, like molecules, can sometimes be crystallized, and their three-dimensional (3-D) structure can then be determined by X-ray crystallography. This paper considers a 3-D interpolation problem that commonly arises during structure determination for spherical viruses. Spherical viruses are viruses with a shell of protein (the capsid) surrounding an inner core of nucleic acid. The capsid is “crystalline” in the sense that it is constructed from many repetitions of the same polypeptides, and the entire capsid is invariant under the rotational symmetries of the icosahedron. The icosahedron, as shown in Fig. 1, is constructed from 20 equilateral triangles and has 60 rotational symmetries: a five-fold axis where five triangles meet, a three-fold axis through the center of each triangle, and a two-fold axis at the midpoint of each edge between two triangles. A typical outer radius of the capsid is in the range 10^2 – 10^3 Å.

Let $\rho(\mathbf{x})$ [with Fourier transform $P(\mathbf{k})$] be the electron density in the crystal in real space or equivalently object space. $P(\mathbf{k})$ in reciprocal space, or, equivalently, Fourier space, is impulsive because $\rho(\mathbf{x})$ is periodic. The lattice of impulse locations is called the reciprocal lattice. The data in an X-ray crystallography experiment, which are called intensities, are the magnitude squared of the weights on the impulses of $P(\mathbf{k})$ [1]. One period of $\rho(\mathbf{x})$ is called the unit cell and occupies the volume S_u . Let $\rho_u(\mathbf{x})$ [with Fourier transform $P_u(\mathbf{k})$] be the electron density in the unit cell [i.e., $\rho_u(\mathbf{x}) = \rho(\mathbf{x})$ for $\mathbf{x} \in S_u$ and = 0 otherwise].

Manuscript received September 29, 1997; revised September 15, 1998. This work was supported by the National Science Foundation under Grants BIR-9513594 and DBI-9630497. The associate editor coordinating the review of this paper and approving it for publication was Dr. Phillip A. Regalia.

Y. Zheng is with Corporate R&D, General Electric, Schenectady, NY 12301 USA.

P. C. Doerschuk is with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907-1285 USA.

J. E. Johnson is with the Department of Molecular Biology, The Scripps Research Institute, La Jolla, CA 92037 USA.

Publisher Item Identifier S 1053-587X(00)00090-8.

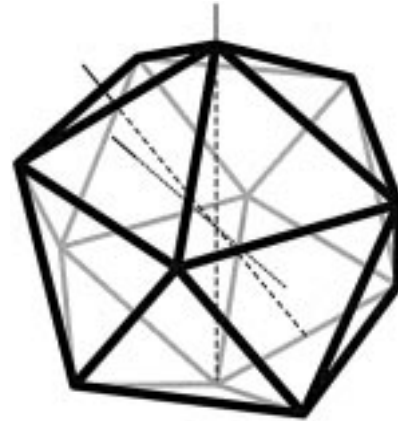


Fig. 1. Icosahedron. One symmetry axis of each type (two-, three-, and five-fold) is shown.

erwise], which is therefore bandlimited in real space. The intensities are proportional to samples of $G(\mathbf{k}) = (2\pi)^{3/2} |P_u(\mathbf{k})|^2$. Note that $G(\mathbf{k})$ is the Fourier transform of the autocorrelation function $g(\mathbf{x}) = \int \rho_u(\boldsymbol{\lambda}) \rho_u(\boldsymbol{\lambda} - \mathbf{x}) d\boldsymbol{\lambda}$, which is also bandlimited in real space.

A complete data set is one in which the intensities for all reciprocal lattice sites in some resolution range $k_0 \leq |\mathbf{k}| \leq k_1$ are measured. For experimental reasons, the data sets used to determine a virus structure are incomplete. However, in order to use standard algorithms (e.g., the electron density modification algorithm [2]), an essentially complete data set is required. The subject of this paper is one approach to estimating the missing intensities.

Because of the icosahedral symmetry of the viral particle in real space, the function $G(\mathbf{k})$ in reciprocal space also displays symmetry. This paper concerns crystals in which the unit cell either contains one virus particle or contains multiple virus particles in the same orientation. In these cases, the symmetry of $G(\mathbf{k})$ is icosahedral symmetry. We focus on the case where there is only one virus particle because in that case, we can more precisely model the contents of the unit cell in real space.

The interesting symmetry operations on $G(\mathbf{k})$ are operations that transform points on the reciprocal lattice (where measurements were made) to points off the lattice (where measurements cannot be made). Because the intensity [i.e., the value of $G(\mathbf{k})$] at the two points must be equal, measurement of the intensity at the point on the lattice determines the intensity at the point off the lattice. The objective of this paper is to create a complete data set by optimally interpolating the value of the intensity at reciprocal lattice points where no measurement was made from all the measurements (both actual measurements and the addi-

tional measurements implied by the symmetry). The intuition is that even if the intensity is measured for only one fifth of the reciprocal lattice points in the resolution range of interest (a fairly typical fraction), the fact that each measured value is really 60 measured values gives hope that the value of the intensity at reciprocal lattice sites where no measurement was made can be interpolated from the (60/5)-fold excess of irregularly sampled intensity values with sufficient accuracy to allow determination of the 3-D virus structure. The interpolation problem is difficult for the reasons listed in the Abstract. Note, however, that the interpolation only needs to be done once, at the beginning of the reconstruction process, so that substantial computation can be done.

The general problem is interpolation of a bandlimited function $G(\mathbf{k})$ from samples of the function. This problem has a long history in signal processing [3] and crystallography [4]. If an infinite number of samples are taken on a uniform rectangular lattice in an M -dimensional space and the support region of the Fourier transform is rectangular, then the well-known answer is given in terms of “ $\prod_{i=1}^M \text{sinc}(x_i)/x_i$.” In crystallography, such functions are called interference functions and are central to rotation function calculations [4]. Very few exact results giving explicit formulas for the interpolating function are known when an infinite number of samples are taken on a nonuniform lattice even in one dimension (e.g., [5], [6]), or the support region of the Fourier transform in multiple dimensions is not rectangular or spherical. A finite number of samples never leads to a unique interpolation without an additional constraint, and numerous constraints have been studied:

- polynomial interpolators [7], [8];
- nonlinear warpings of the independent variable [9];
- linear bandlimited interpolators that minimize the energy in the interpolated function [10], [11];
- interpolation motivated by the two-dimensional (2-D) problems that arise in computed tomography (see [12] and the references cited therein).

Interpolation is currently used to estimate the missing intensity values. However, the current methods are local and do not fully exploit the icosahedral symmetry of the intensity data and are therefore inaccurate. In this paper, we describe more sophisticated global interpolation algorithms based on a least-squares point of view that exploits the symmetry. This approach has two advantages: First, if the number of degrees of freedom is large enough and the basis functions are properly chosen, then least squares can provide exact interpolation of consistent data (i.e., data that exactly satisfies the symmetry condition) while providing an inexact least-squares fit to inconsistent data that takes into account weights (e.g., variances) that describe the accuracy of each data point. Second, it allows *a priori* information on virus structure (support and symmetry constraints) to be included by use of appropriate basis functions.

In order to demonstrate the generality of the least squares point of view, we show that proper choice of the basis functions and weighting matrices in least squares makes the least squares interpolator the same as interpolators derived from two alternative points of view. The first alternative point of view is to consider interpolation as a Bayesian estimation problem with Gaussian *a priori* model and Gaussian measurement noise. The

second alternative point of view, which is applicable only to data that exactly exhibits the symmetry (i.e., noise free data), is minimum-energy bandlimited exact interpolation (MEBLEI). The major benefit of these two alternative points of view is that they provide motivation for using particular families of basis functions in the least squares problem. In addition, for the MEBLEI problem, we show how to reduce the computation by the number of symmetry elements, which is 60 in the complete icosahedral group.

Several families of basis functions are considered: Families originating with the MEBLEI and Bayesian problems, and a family based on harmonic functions. The 3-D character and large number of measurements are central characteristics of the problem. The families originating with the MEBLEI and Bayesian problems have a number of basis functions equal to the number of measurements and no natural way to choose a subset of the basis functions. Therefore, they are not attractive for computation. The family based on harmonic functions has an infinite number of basis functions. However, there is a natural way to choose subsets based on bandwidth constraints, and by using subsets, we have successfully solved several problems.

Tobacco ring spot virus (TRSV) provides a typical example of the interpolation problem. The space group of the TRSV crystals is $C2$ with parameters

$$\begin{aligned} a &= 407.1 \text{ \AA}; \\ b &= 399.7 \text{ \AA}; \\ c &= 285.9 \text{ \AA}; \\ \beta &= 129.1^\circ. \end{aligned}$$

Sufficient virus was available to grow 27 crystals from which 71 imaging plates were recorded. In the resolution range of $0.02 \leq |\mathbf{k}| \leq 0.25$ cycles/Å, there were 66 594 measured unique reflections, which is roughly 22% of the total unique reflections in that resolution range. The quality of the data is determined by comparing reflections (in terms of the square root of the intensities) that should be equivalent due to the space group symmetry. The scaling R factor for only whole reflections with $I/\sigma(I) \geq 4$ after postrefinement was 8.4% (40 649 observations and 37 381 unique reflections). In this instance, the factor that prevents the recording of additional data is that it has not been possible to grow additional crystals that diffract to this resolution. Because the virus particle is positioned on a two-fold axis of the space group, the 60-fold redundancy is reduced to 30-fold. Since 22% of the data is available and the data is intrinsically 30-fold redundant, the available data is slightly more than six-fold redundant. This level of redundancy should be sufficient to solve the structure if the gaps in the data set could be accurately filled in by interpolation so that standard refinement programs would be able to converge.

The remainder of the paper is organized as follows. After defining notation (Section II), we list the properties of a viral particle and describe a low-resolution nominal model (Sections III and IV). The following three sections (Sections V–VII) describe the least squares, Bayesian, and minimum-energy bandlimited exact interpolators and show that the latter two are special cases of the first. In Section VIII, we describe two families

of basis functions. Finally, numerical results are described in Section IX, and the results are discussed in Section X.

II. NOTATION AND MATHEMATICAL PRELIMINARIES

\mathcal{Z}	Integer.
\mathcal{R}	Real numbers.
\mathcal{C}	Complex numbers.
*	Complex conjugation.
T	Transpose.
H	Hermitian transpose.
$-T$	Transpose of the inverse matrix.
$0_{i,j}$	$i \times j$ zero matrix.
I_n	$n \times n$ identity matrix.

“ >0 ” applied to a matrix means positive definite, and if $Q \in \mathcal{C}^{n \times n}$ with $Q^H = Q > 0$, then $Q^{1/2}$ denotes the matrix square root of Q , i.e., $Q = (Q^{1/2})^H Q^{1/2}$. The l th spherical Bessel function of the first and second kinds [13, Sec. 7.2.1] are denoted by $j_l(r)$ and $y_l(r)$, respectively. The Gaussian probability density function with mean m and covariance Q is denoted by $N(m, Q)$. The Dirac delta function is denoted by $\delta(x)$ or $\delta(\mathbf{x}, \mathbf{x}')$. The indicator function on the set S is denoted by $\mathbf{1}_S(\mathbf{x})$. The inner product of \mathbf{k} and \mathbf{x} is denoted by $\mathbf{k}^T \mathbf{x}$. Define $\tau = 1/(2\pi)^{3/2}$. The Fourier transform of a function $f(\mathbf{x})$ is defined by $F(\mathbf{k}) = \tau \int f(\mathbf{x}) \exp(-i\mathbf{k}^T \mathbf{x}) d\mathbf{x}$.

Let $\mathbf{R}_\beta \in \mathcal{R}^{3 \times 3}$ ($\beta \in \{0, \dots, 59\}$) be the orthonormal matrix with determinant +1 representing the β th rotation of the icosahedral group [14], [15]. To say a function $\rho(\mathbf{x})$ ($\mathbf{x} \in \mathcal{R}^3$) has icosahedral symmetry means $\rho(\mathbf{x}) = \rho(\mathbf{R}_\beta^{-1} \mathbf{x})$ for all $\mathbf{x} \in \mathcal{R}^3$ and $\beta \in \{0, \dots, 59\}$. The presence of the symmetry changes the natural notion of distance. Specifically, instead of measuring the distance between two points \mathbf{x} and \mathbf{x}' by $|\mathbf{x} - \mathbf{x}'| = [\sum_{n=1}^3 (x_n - x'_n)^2]^{1/2}$, it is now natural to use the metric $d(\mathbf{x}, \mathbf{x}') = \min_\beta |\mathbf{x} - \mathbf{R}_\beta \mathbf{x}'|$.

Integrals of the type $\zeta(\mathbf{k}; S) = \tau \int_S \exp(-i\mathbf{k}^T \mathbf{x}) d^3\mathbf{x}$ occur repeatedly and can be done analytically when S is a cube or a ball. Define the sets $S_c(\mathbf{s}) = [-s_1, s_1] \times [-s_2, s_2] \times [-s_3, s_3]$ and $S_b(r_0) = \{\mathbf{x}: |\mathbf{x}| \leq r_0\}$ and the functions $\zeta_c(\mathbf{k}; \mathbf{s}) = \tau \prod_{j=1}^3 2s_j j_0(k_j s_j)$ and $\zeta_b(k; r_0) = \tau \sqrt{4\pi} r_0^3 j_1(kr_0)/(kr_0)$. Then, $\zeta(\mathbf{k}; S_c(\mathbf{s})) = \zeta_c(\mathbf{k}; \mathbf{s})$, and $\zeta(\mathbf{k}; S_b(r_0)) = \zeta_b(|\mathbf{k}|; r_0)$.

III. CHARACTERISTICS OF THE VIRUS PARTICLE

Because we consider the case where there is one virus particle per unit cell, the electron density $\rho_u(\mathbf{x})$ is essentially the same (up to a rotation and translation) as the electron density of the virus particle, which is denoted by $\rho_v(\mathbf{x})$. The electron density $\rho_v(\mathbf{x})$ has several characteristics that we wish to incorporate into the interpolation process by correct choice of basis functions. The properties of $\rho_v(\mathbf{x})$ and their implications for $g(\mathbf{x})$ are as follows:

- 1) *Icosahedral Constraint:* $\rho_v(\mathbf{x})$ has icosahedral symmetry as defined in Section II. Therefore, $G(\mathbf{k})$ and $g(\mathbf{x})$ also have the symmetry.
- 2) *Support Constraint:* $\rho_v(\mathbf{x}) = 0$ for $|\mathbf{x}| \leq R_-$ and $|\mathbf{x}| \geq R_+$. Therefore, $g(\mathbf{x}) = 0$ for $|\mathbf{x}| \geq 2R_+$.
- 3) *Real-Valued Constraint:* $\rho_v(\mathbf{x})$, and therefore, $g(\mathbf{x})$ are real.

- 4) *Positivity Constraint:* $\rho_v(\mathbf{x}) \geq -\rho_0$ for all $\mathbf{x} \in \mathcal{R}^3$, where ρ_0 is the background solvent electron density.
- 5) *Inversion Constraint:* $\rho_v(\mathbf{x})$ real implies that $g(\mathbf{x}) = g(-\mathbf{x})$.

In addition, it is desirable to have a mathematical representation of $\rho_v(\mathbf{x})$ from which $P(\mathbf{k})$ can be computed analytically since this computation is a 3-D integral.

IV. A NOMINAL MODEL

A very low-resolution model for $\rho_v(\mathbf{x})$ in a spherical virus is [16]

$$\rho_0(\mathbf{x}) = \begin{cases} \rho_c, & 0 \leq |\mathbf{x}| < R_- \\ \rho_s, & R_- \leq |\mathbf{x}| < R_+ \\ 0, & R_+ \leq |\mathbf{x}|. \end{cases}$$

The functions $P_0(\mathbf{k})$ [the Fourier transform of $\rho_0(\mathbf{x})$], $g_0(\mathbf{x}) = \int \rho_0(\lambda) \rho_0^*(\lambda - \mathbf{x}) d\lambda$, and $G_0(\mathbf{k}) = (2\pi)^{3/2} |P_0(\mathbf{k})|^2$ [the Fourier transform of $g_0(\mathbf{x})$] can all be computed analytically.

Let S_{ρ_0} and S_{g_0} denote the support of $\rho_0(\mathbf{x})$ and $g_0(\mathbf{x})$, respectively. In general, $S_{\rho_0} = \{\mathbf{x}: |\mathbf{x}| < R_+\}$ and $S_{g_0} = \{\mathbf{x}: |\mathbf{x}| < 2R_+\}$. In some situations, a so-called empty virus is studied, in which case, $\rho_c = 0$. In these situations, $S_{\rho_0} = \{\mathbf{x}: R_- \leq |\mathbf{x}| < R_+\}$, but it is still true that $S_{g_0} = \{\mathbf{x}: |\mathbf{x}| < 2R_+\}$.

V. LEAST SQUARES

Let j_0 be the number of measurements that were made, let \mathbf{k}_j ($j \in \{1, \dots, j_0\}$) be the reciprocal space coordinates of the j th measurement, let G_j^d (“ d ” for “data”) be the value of the j th measurement, and let $G^d \in \mathcal{R}^{j_0}$ have components G_j^d . Let n_0 be the number of basis functions.

The autocorrelation function $g(\mathbf{x})$ or, equivalently, its Fourier transform $G(\mathbf{k})$ is described by $g(\mathbf{x}) = \sum_\alpha b_\alpha(\mathbf{x}) a_\alpha$ and $G(\mathbf{k}) = \sum_\alpha B_\alpha(\mathbf{k}) a_\alpha$, where the functions $b_\alpha(\mathbf{x})$ [with Fourier transforms $B_\alpha(\mathbf{k})$] are the basis functions, and a_α are the weights. Define the row-vector valued functions $b(\mathbf{x})$ and $B(\mathbf{k})$ with components $b_\alpha(\mathbf{x})$ and $B_\alpha(\mathbf{k})$, respectively, and the column vector a , with components a_α . Then, the interpolator can be written

$$g(\mathbf{x}) = b(\mathbf{x})a \quad (1)$$

$$G(\mathbf{k}) = B(\mathbf{k})a. \quad (2)$$

We consider least squares estimation of the unknown parameters $a \in \mathcal{C}^{n_0}$ from the data $G^d \in \mathcal{R}^{j_0}$ with a penalty on the deviation of the unknown parameters from a set of nominal parameters denoted by a_0 . Therefore, the nominal $g(\mathbf{x})$ and $G(\mathbf{k})$ are $g_0(\mathbf{x}) = b(\mathbf{x})a_0$ and $G_0(\mathbf{k}) = B(\mathbf{k})a_0$, respectively. Let $L \in \mathcal{C}^{j_0 \times n_0}$ be a matrix with rows $B(\mathbf{k}_j)$. Therefore, the vector La is the predicted value of the measurements given parameters a . The least squares cost function is

$$\chi_\lambda(a) = \frac{1}{2} (G^d - La)^H \Xi (G^d - La) + \lambda \frac{1}{2} (a - a_0)^H \Sigma (a - a_0) \quad (3)$$

where Ξ and Σ are the weights on the measurements and *a priori* model, respectively, and λ is a regularization parameter. The a that minimizes $\chi_\lambda(a)$ is described by Theorem 1.

Theorem 1: If $a \in \mathcal{C}^{n_0}$, $G^d \in \mathcal{C}^{j_0}$, $L \in \mathcal{C}^{j_0 \times n_0}$, $a_0 \in \mathcal{C}^{n_0}$, $\lambda \in \mathcal{R}$ with $\lambda \geq 0$, $\Xi \in \mathcal{C}^{j_0 \times j_0}$ with $\Xi^H = \Xi > 0$, $\Sigma \in \mathcal{C}^{n_0 \times n_0}$ with $\Sigma^H = \Sigma > 0$, and $\chi_\lambda: \mathcal{C}^{n_0} \rightarrow \mathcal{R}$ is defined by (3), then the global minimum of $\chi_\lambda(a)$ with respect to a occurs at

$$a = \Sigma^{-1/2} U M_\lambda V^H \Xi^{-1/2} (G^d - L a_0) + a_0 \quad (4)$$

where

$$M_\lambda = \begin{bmatrix} \text{diag} \left(\frac{\mu_1}{\mu_1^2 + \lambda}, \dots, \frac{\mu_\kappa}{\mu_\kappa^2 + \lambda} \right) & 0_{\kappa, j_0 - \kappa} \\ 0_{n_0 - \kappa, \kappa} & 0_{n_0 - \kappa, j_0 - \kappa} \end{bmatrix}$$

where κ is the rank of L , and μ_k ($k \in \{1, \dots, \kappa\}$), U , and V are the singular value decomposition of $A = \Xi^{1/2} L \Sigma^{-1/2}$: $Au = \mu v$ and $A^H v = \mu u$, where u and v are the columns of U and V , respectively.

By completion of the square in (3), the result of Theorem 1 can also be written:

$$a = (L^H \Xi L + \lambda \Sigma)^{-1} L^H \Xi (G^d - L a_0) + a_0 \quad (5)$$

but then, the limit $\lambda \rightarrow 0$ is difficult to compute unless $L^H \Xi L$ is full rank [i.e., $n_0 \leq j_0$ and $\text{rank}(L) = n_0$].

Using (4) or (5) in (1) and (2) and applying $g_0(\mathbf{x}) = b(\mathbf{x})a_0$ and $G_0(\mathbf{k}) = B(\mathbf{k})a_0$ provides two equivalent formulae for the least squares interpolator in real space and two more equivalent formulae for the interpolator in reciprocal space. For instance [(4) in (2)], one of the reciprocal space formulae is

$$G(\mathbf{k}) = G_0(\mathbf{k}) + B(\mathbf{k}) \Sigma^{-1/2} U M_\lambda V^H \Xi^{-1/2} (G^d - L a_0). \quad (6)$$

If it is desired to use a nominal model $G_0(\mathbf{k})$ that cannot be written in the form $B(\mathbf{k})a_0$, then the natural approach is to perform least squares on the residual $G^d - G_0$ (where $G_0 = (G_0(\mathbf{k}_1), \dots, G_0(\mathbf{k}_{j_0}))^T$) with $a_0 = 0_{j_0, 1}$. Such a least squares problem occurs in Section VI.

VI. RELATIONSHIP WITH GAUSSIAN BAYESIAN INTERPOLATION

In this section, we describe a Bayesian approach to interpolation that provides an alternative to the least-squares point of view since in the Bayesian approach, the basis functions are derived from the statistical assumptions. The idea is to assume a statistical *a priori* model for $g(\mathbf{x})$ and a statistical measurement model for G^d and then compute the conditional mean of $G(\mathbf{k})$ at the desired coordinate location \mathbf{k} conditional on the measured values G^d . We show that the Bayesian approach is a special case of the least-squares approach. However, in spite of this fact, the Bayesian approach is interesting as a method for choosing basis functions.

The *a priori* model is that $g(\mathbf{x})$ is a Gaussian stochastic process with mean $\bar{g}(\mathbf{x})$ and covariance $C_g(\mathbf{x}, \mathbf{x}') = E\{[g(\mathbf{x}) - \bar{g}(\mathbf{x})][g(\mathbf{x}') - \bar{g}(\mathbf{x}')]^*\}$. Assume that $G(\mathbf{k})$ is related to $g(\mathbf{x})$ by $G(\mathbf{k}) = \int \gamma(\mathbf{k}, \mathbf{x}) g(\mathbf{x}) d^3 \mathbf{x}$. For the

crystallographic problem, $\gamma(\mathbf{k}, \mathbf{x}) = \tau \exp(-i\mathbf{k}^T \mathbf{x})$, but the additional freedom of an arbitrary $\gamma(\mathbf{k}, \mathbf{x})$ is useful in Section VII. Then, $G(\mathbf{k})$ is a Gaussian stochastic process with mean $\bar{G}(\mathbf{k}) = \int \gamma(\mathbf{k}, \mathbf{x}) \bar{g}(\mathbf{x}) d^3 \mathbf{x}$ and covariance $C_G(\mathbf{k}, \mathbf{k}') = \iint C_g(\mathbf{x}, \mathbf{x}') \gamma(\mathbf{k}, \mathbf{x}) \gamma^*(\mathbf{k}', \mathbf{x}') d^3 \mathbf{x} d^3 \mathbf{x}'$. Let \mathbf{k} be the coordinate location at which an interpolated value is desired. For conditional mean estimation, unlike maximum likelihood estimation, it is only necessary to consider one location at a time because conditional mean estimation of a vector of random variables is the same as conditional mean estimation of each component in the vector. The measurement model is that $G^d = G + v$, where $G \in \mathcal{R}^{j_0}$ has components $G(\mathbf{k}_j)$, and $v \in \mathcal{R}^{j_0}$ is $N(0_{j_0, 1}, V)$ and independent of $g(\mathbf{x})$ (and, therefore, independent of G).

Let $m_{G(\mathbf{k})} \in \mathcal{R}^1$ and $m_{G^d} \in \mathcal{R}^{j_0}$ be vectors with components $\bar{G}(\mathbf{k})$ and $\bar{G}(\mathbf{k}_j)$ for $j \in \{1, \dots, j_0\}$, respectively. Let $\Sigma_{G(\mathbf{k}), G(\mathbf{k})} \in \mathcal{R}^{1 \times 1}$, $\Sigma_{G(\mathbf{k}), G^d} \in \mathcal{R}^{1 \times j_0}$, and $\Sigma_{G^d, G^d} \in \mathcal{R}^{j_0 \times j_0}$ be matrices with components $C_G(\mathbf{k}, \mathbf{k})$, $C_G(\mathbf{k}, \mathbf{k}_j)$ for $j \in \{1, \dots, j_0\}$, and $C_G(\mathbf{k}_j, \mathbf{k}_{j'}) + V_{j, j'}$ for $j, j' \in \{1, \dots, j_0\}$, respectively, where $V_{j, j'}$ is the (j, j') th element of V . Then, $(G(\mathbf{k})^T, G^d)^T$ is Gaussian with mean $(m_{G(\mathbf{k})}^T, m_{G^d}^T)^T$ and covariance

$$\begin{bmatrix} \Sigma_{G(\mathbf{k}), G(\mathbf{k})} & \Sigma_{G(\mathbf{k}), G^d} \\ \Sigma_{G(\mathbf{k}), G^d}^T & \Sigma_{G^d, G^d} \end{bmatrix}.$$

Therefore, the conditional probability density function on $G(\mathbf{k})$, given G^d , is Gaussian with mean and covariance

$$p(m_{G(\mathbf{k})} | G^d) = m_{G(\mathbf{k})} + \Sigma_{G(\mathbf{k}), G^d} \Sigma_{G^d, G^d}^{-1} (G^d - m_{G^d}) \\ \Sigma_{G(\mathbf{k}) | G^d} = \Sigma_{G(\mathbf{k}), G(\mathbf{k})} - \Sigma_{G(\mathbf{k}), G^d} \Sigma_{G^d, G^d}^{-1} \Sigma_{G^d, G(\mathbf{k})} \quad (7)$$

and (7) is the interpolator.

The Bayesian approach (7) is a special case of the least-squares approach (6). The parameters in the least-squares approach that make the least-squares interpolator equal to the Bayesian interpolator are the following: Take basis functions $B(\mathbf{k}) = \Sigma_{G(\mathbf{k}), G^d} = (C_G(\mathbf{k}, \mathbf{k}_1), \dots, C_G(\mathbf{k}, \mathbf{k}_{j_0}))$ so that L is square, has components $L_{j, j'} = C_G(\mathbf{k}_j, \mathbf{k}_{j'})$ ($j, j' \in \{1, \dots, j_0\}$), and, therefore, $L^H = L > 0$, so that L^{-1} exists. In addition, take measurement weights $\Xi = V^{-1}$, parameter weights $\Sigma = L^H L^{-1} L = L$, and regularization parameter $\lambda = 1$. If the mean $m_{G(\mathbf{k})}$ in the Bayesian problem is equal to $B(\mathbf{k})L^{-1}m_{G^d}$, then take nominal parameters $a_0 = L^{-1}m_{G^d}$, apply least squares to the data directly, and find that the interpolators for the least-squares and Bayesian approaches are identical by direct calculation. If the mean $m_{G(\mathbf{k})}$ in the Bayesian problem is not equal to $B(\mathbf{k})L^{-1}m_{G^d}$, then take nominal parameters $a_0 = 0_{j_0, 0}$, apply least squares to the residual $G^d - m_{G^d}$, and again find that the least squares and Bayesian approaches are identical.

As described previously, the Bayesian framework determines the basis functions in terms of the statistical assumptions, and the number of basis functions (i.e., n_0) is equal to the number of measurements (i.e., j_0). If $\gamma(\mathbf{k}, \mathbf{x}) = \tau \exp(-i\mathbf{k}^T \mathbf{x})$, then since the basis functions are $B(\mathbf{k}) = (C_G(\mathbf{k}, \mathbf{k}_1), \dots, C_G(\mathbf{k}, \mathbf{k}_{j_0}))$

in the reciprocal space, they are $b(\mathbf{x}) = (h_1(\mathbf{x}), \dots, h_{j_0}(\mathbf{x}))$ in real space, where

$$\begin{aligned} h_j(\mathbf{x}) &= \tau \int C_G(\mathbf{k}, \mathbf{k}_j) \exp(i\mathbf{k}^T \mathbf{x}) d^3 \mathbf{k} \\ &= \tau \int C_g(\mathbf{x}, \mathbf{x}') \exp(+i\mathbf{k}_j^T \mathbf{x}') d^3 \mathbf{x}'. \end{aligned}$$

Now, consider the case with exact measurements: $V = 0$. In this case, (7) is an exact interpolator, that is, $m_{G(\mathbf{k})|G^d} \mathbf{k} = \mathbf{k}_j = G(\mathbf{k}_j)$ because $\Sigma_{G(\mathbf{k}_j), G^d} \Sigma_{G^d, G^d}^{-1}$ is the j th row of $LL^{-1} = I_{j_0}$. The case of $V = 0$ is also equivalent to the MEBLEI interpolators of Section VII, as shown in Section VII.

Let $\gamma(\mathbf{k}, \mathbf{x}) = \tau \exp(-i\mathbf{k}^T \mathbf{x})$ for the remainder of this section. A natural choice for $\bar{G}(\mathbf{k})$, based on the nominal model of Section IV, is $\bar{G}(\mathbf{k}) = G_0(\mathbf{k})$. Two classes of model for the deviations of $g(\mathbf{x})$ around $g_0(\mathbf{x})$ are considered. In the first class, the deviations [denoted by $g_u(\mathbf{x})$] are a space-varying white stochastic process with covariance $C_{g_u}(\mathbf{x}, \mathbf{x}') = \omega_u(\mathbf{x})\delta(\mathbf{x}, \mathbf{x}')$, where $\omega_u(\mathbf{x})$ is any positive function. This class allows modeling of the support constraint of the virus by having $\omega_u(\mathbf{x}) = 0$ on the complement of the support set. It does not enforce the symmetry constraint and, hence, the subscript “ u ” for “unsymmetric.” In the second class, the deviations [denoted by $g_s(\mathbf{x})$] have icosahedral symmetry. Let $g_s(\mathbf{x})$ be defined by

$$g_s(\mathbf{x}) = (1/60) \sum_{\beta=0}^{59} g_u(\mathbf{R}_\beta \mathbf{x})$$

which is symmetric by construction and, hence, the subscript “ s ” for “symmetric.” By direct calculation exploiting the group structure of the matrices \mathbf{R}_β , it can be shown that the covariance of $g_s(\mathbf{x})$ is

$$C_{g_s}(\mathbf{x}, \mathbf{x}') = \omega_s(\mathbf{x})\Delta(\mathbf{x}, \mathbf{x}')$$

where

$$\begin{aligned} \omega_s(\mathbf{x}) &= (1/60) \sum_{\beta=0}^{59} \omega_u(\mathbf{R}_\beta \mathbf{x}) \\ \Delta(\mathbf{x}, \mathbf{x}') &= (1/60) \sum_{\beta=0}^{59} \delta(\mathbf{x} - \mathbf{R}_\beta \mathbf{x}'). \end{aligned}$$

($\Delta(\mathbf{x}, \mathbf{x}')$ is called a symmetrized delta function [14], [15]). By considering $\omega_u(\mathbf{x})$ that are nonzero only on one asymmetric unit of the group, it immediately follows that $\omega_s(\mathbf{x})$ can be any symmetric function. Therefore, in the second class, the deviations have covariance $C_{g_s}(\mathbf{x}, \mathbf{x}') = \omega_s(\mathbf{x})\Delta(\mathbf{x}, \mathbf{x}')$, where $\omega_s(\mathbf{x})$ is any positive symmetric function.

Define $\Omega_u(\mathbf{k})$ and $\Omega_s(\mathbf{k})$ to be the Fourier transforms of $\omega_u(\mathbf{x})$ and $\omega_s(\mathbf{x})$, respectively. Then

$$\begin{aligned} C_{G_u}(\mathbf{k}, \mathbf{k}') &= \tau \Omega_u(\mathbf{k} - \mathbf{k}') \\ C_{G_s}(\mathbf{k}, \mathbf{k}') &= (\tau/60) \sum_{\beta'=0}^{59} \Omega_s(\mathbf{k} - \mathbf{R}_{\beta'} \mathbf{k}') \end{aligned} \quad (8)$$

where the calculation of $C_{G_s}(\mathbf{k}, \mathbf{k}')$ exploits the group structure of the rotation matrices \mathbf{R}_β . Therefore, the reciprocal-space basis functions for unsymmetric and symmetric classes are $B_j(\mathbf{k}) = C_G(\mathbf{k}, \mathbf{k}_j) = \tau \Omega_u(\mathbf{k} - \mathbf{k}_j)$ and $= (\tau/60) \sum_{\beta=0}^{59} \Omega_s(\mathbf{k} - \mathbf{R}_\beta \mathbf{k}_j)$, respectively, and the real-space basis functions for unsymmetric and symmetric classes are $b_j(\mathbf{x}) = \tau \int B_j(\mathbf{k}) \exp(i\mathbf{k}^T \mathbf{x}) d^3 \mathbf{k} = \tau \omega_u(\mathbf{x}) \exp(i\mathbf{k}_j^T \mathbf{x})$ and $= \omega_s(\mathbf{x}) (\tau/60) \sum_{\beta=0}^{59} \exp(i\mathbf{R}_\beta \mathbf{k}_j^T \mathbf{x})$, respectively. The simplest choices for $\omega_u(\mathbf{x})$ and $\omega_s(\mathbf{x})$ are $\omega_u(\mathbf{x}) = \mathbf{1}_{S_c(\mathbf{s})}(\mathbf{x})$ (which is not symmetric), $\omega_u(\mathbf{x}) = \mathbf{1}_{S_{g_0}}(\mathbf{x})$ (which is symmetric), and $\omega_s(\mathbf{x}) = \mathbf{1}_{S_{g_0}}(\mathbf{x})$ (which is symmetric), which result in the basis functions

$$\begin{aligned} B_j(\mathbf{k}) &= \begin{cases} \tau \zeta_c(\mathbf{k} - \mathbf{k}_j; \mathbf{s}) \\ \tau \zeta_b(|\mathbf{k} - \mathbf{k}_j|; 2R_+) \\ (\tau/60) \sum_{\beta=0}^{59} \zeta_b(|\mathbf{k} - \mathbf{R}_\beta \mathbf{k}_j|; 2R_+) \end{cases} \quad (9) \\ b_j(\mathbf{x}) &= \begin{cases} \tau \mathbf{1}_{S_c(\mathbf{s})}(\mathbf{x}) \exp(i\mathbf{k}_j^T \mathbf{x}) \\ \tau \mathbf{1}_{S_{g_0}}(\mathbf{x}) \exp(i\mathbf{k}_j^T \mathbf{x}) \\ \mathbf{1}_{S_{g_0}}(\mathbf{x}) (\tau/60) \sum_{\beta=0}^{59} \exp(i\mathbf{R}_\beta \mathbf{k}_j^T \mathbf{x}) \end{cases} \quad (10) \end{aligned}$$

respectively. Other choices of $\omega(\mathbf{x})$ that cannot be written in terms of linear combinations of these choices of $\omega(\mathbf{x})$ lead to 3-D Fourier transforms that are difficult to evaluate analytically.

VII. RELATIONSHIP WITH MINIMUM-ENERGY BAND-LIMITED EXACT INTERPOLATION

As described in Theorems 2 and 3 below, the basis functions in the minimum-energy bandlimited exact interpolation (MEBLEI) problem are determined by the Fourier transform relationship between $G(\mathbf{k})$ and $g(\mathbf{x})$ and by the support constraint. The resulting basis functions are basis functions that are also naturally generated by the Bayesian approach. Therefore, the MEBLEI approach is not a method for generating novel basis functions. In theory, the MEBLEI can be found for an arbitrary support constraint, but the calculations require the evaluation of 3-D integrals that can be done analytically only for cubic $S_c(\mathbf{s})$ and spherical $S_{g_0} = S_b(2R_+)$ support regions.

Consider an invertible linear operator with kernel $w(\mathbf{x}, \mathbf{x}')$ [i.e., if $u(\mathbf{x})$ is the input and $v(\mathbf{x})$ is the output of the linear operator, then $v(\mathbf{x}) = \int w(\mathbf{x}, \mathbf{x}') u(\mathbf{x}') d^3 \mathbf{x}'$], and denote the kernel of the inverse operator by $w_{\text{inv}}(\mathbf{x}, \mathbf{x}')$. Define $W(\mathbf{x}, \mathbf{x}') = \int w^*(\xi, \mathbf{x}) w(\xi, \mathbf{x}') d^3 \xi$. The linear operator having W as its kernel is also invertible, and the inverse has kernel $W_{\text{inv}}(\xi, \mathbf{x}) = \int w_{\text{inv}}(\mathbf{x}, \xi) w_{\text{inv}}^*(\mathbf{x}', \xi) d^3 \xi$. The MEBLEI problem is summarized in Theorem 2.

Theorem 2: Let $W(\mathbf{x}, \mathbf{x}')$, $W_{\text{inv}}(\mathbf{x}, \mathbf{x}')$, $w(\mathbf{x}, \mathbf{x}')$, and $w_{\text{inv}}(\mathbf{x}, \mathbf{x}')$ be related as above. Consider the interpolation problem

$$\min \frac{1}{2} \int_{S_g} \int_{S_g} [g(\mathbf{x}) - g_0(\mathbf{x})]^* W(\mathbf{x}, \mathbf{x}') [g(\mathbf{x}') - g_0(\mathbf{x}')] \cdot d^3 \mathbf{x} d^3 \mathbf{x}'$$

subject to

$$\alpha_j = \int_{S_g} b_j^*(\mathbf{x})g(\mathbf{x})d^3\mathbf{x}$$

for $j \in \{1, \dots, j_0\}$. Define

$$\begin{aligned} \Lambda_{j,j'} &= \iint b_j^*(\mathbf{x})W_{\text{inv}}(\mathbf{x}, \mathbf{x}')b_{j'}(\mathbf{x}')d^3\mathbf{x}d^3\mathbf{x}' \\ \tilde{\alpha}_j &= \alpha_j - \int b_j^*(\mathbf{x})g_0(\mathbf{x})d^3\mathbf{x}. \end{aligned} \quad (11)$$

Define $\Lambda \in \mathcal{C}^{j_0 \times j_0}$ to have elements $\Lambda_{j,j'}$, $\tilde{\alpha} \in \mathcal{C}^{j_0}$ to have components $\tilde{\alpha}_j$, and $\lambda \in \mathcal{C}^{j_0}$ by $\tilde{\alpha} = \Lambda\lambda$. Then, the solution is

$$g(\mathbf{x}) = g_0(\mathbf{x}) + \sum_{j=1}^{j_0} \left[\int W_{\text{inv}}(\mathbf{x}, \mathbf{x}')b_j(\mathbf{x}')d\mathbf{x}' \right] \lambda_j.$$

The interpolator of Theorem 2 corresponds to a subclass of the Bayesian interpolators described by (7), specifically, those Bayesian interpolators with $V = 0$, which makes them exact interpolators. If the Bayesian problem is given, then the corresponding MEBLEI problem is

$$\begin{aligned} g_0(\mathbf{x}) &= \bar{g}(\mathbf{x}) \\ W_{\text{inv}}(\mathbf{x}, \mathbf{x}') &= C_g(\mathbf{x}, \mathbf{x}') \\ b_j(\mathbf{x}) &= \gamma(\mathbf{k}_j, \mathbf{x}) \end{aligned}$$

which can be verified by directly checking that the basis functions, weights, and nominal model are identical. Since the Bayesian problem is a special case of the least squares problem (Section VI), this result also describes the correspondence between the MEBLEI and least squares problems.

For the crystallography problem, $b_j(\mathbf{x}) = \tau \exp(i\mathbf{k}_j^T \mathbf{x})$. Because the integral in (11) is 3-D, analytical forms are important, and the two cases for which analytical results are available are $S_g = S_c(\mathbf{s})$ (a cube) and $S_g = S_{g_0} = S_b(2R_+)$ (a sphere). Assume that $W(\mathbf{x}, \mathbf{x}') = \delta(\mathbf{x} - \mathbf{x}')$. Then, for $S_g = S_c(\mathbf{s})$, it can be computed that $\Lambda_{j,j'} = \zeta_c(\mathbf{k}_j - \mathbf{k}_{j'}; \mathbf{s})$, and the basis functions are exactly the first cases in (9) and (10), whereas for $S_g = S_{g_0} = S_b(2R_+)$, it can be computed that $\Lambda_{j,j'} = \zeta_b(|\mathbf{k}_j - \mathbf{k}_{j'}|; 2R_+)$, and the basis functions are exactly the second cases in (9) and (10).

Now, consider the effect of symmetry.

Definition 1: A set S is *symmetric* if $\mathbf{x} \in S$ implies $\mathbf{R}_\beta \mathbf{x} \in S$ for all $\beta \in \{0, \dots, 59\}$.

Definition 2: A data set for interpolation is *complete* if for each $j \in \{1, \dots, j_0\}$ and $\beta \in \{0, \dots, 59\}$ there exists a $j' \in \{1, \dots, j_0\}$ such that $\mathbf{R}_\beta \mathbf{k}_j = \mathbf{k}_{j'}$.

Definition 3: A data set for interpolation is *consistent* if $\mathbf{R}_\beta \mathbf{k}_j = \mathbf{k}_{j'}$ implies $\alpha_j = \alpha_{j'}$. Let $\bar{j}_0 = j_0/60$. Assume that the data set is both complete and consistent. Let $\boldsymbol{\kappa}_m \in \mathcal{R}^3$ ($m \in \{1, \dots, \bar{j}_0\}$) be the measurement locations in one asymmetric unit of the icosahedral group. Then, for each $j \in \{1, \dots, j_0\}$, there exists $l \in \{1, \dots, \bar{j}_0\}$ and $\beta \in \{0, \dots, 59\}$ such that $\mathbf{k}_j = \mathbf{R}_\beta \boldsymbol{\kappa}_l$. Furthermore, let $\tilde{\alpha}_m$ (for $m \in \{1, \dots, \bar{j}_0\}$) be the values of the measurements at the locations $\boldsymbol{\kappa}_m$. Then, for each $j \in \{1, \dots, j_0\}$, there exists $l \in \{1, \dots, \bar{j}_0\}$ such that $\alpha_j = \tilde{\alpha}_l$.

By examination of $\tilde{\alpha} = \Lambda\lambda$ in Theorem 2, the following result can be shown.

Theorem 3: Let $Q: \mathcal{R}^3 \rightarrow \mathcal{R}$ satisfy $Q(\mathbf{x}) > 0$ for all $\mathbf{x} \in \mathcal{R}^3$ and satisfy $Q(\mathbf{R}_\beta^{-1} \mathbf{x}) = Q(\mathbf{x})$ for all $\beta \in \{0, \dots, 59\}$ and for all $\mathbf{x} \in \mathcal{R}^3$. Assume that there exists a function $f: \mathcal{R} \rightarrow \mathcal{C}$ such that $b_j(\mathbf{x}) = f(\mathbf{k}_j^T \mathbf{x})$. Assume that $g_0(\mathbf{R}_\beta \mathbf{x}) = g_0(\mathbf{x})$ for all $\beta \in \{0, \dots, 59\}$ and for all $\mathbf{x} \in \mathcal{R}^3$. Assume that S_g is symmetric. Assume that the data set is complete and consistent. Consider the interpolation problem

$$\min \frac{1}{2} \int_{S_g} |g(\mathbf{x}) - g_0(\mathbf{x})|^2 Q(\mathbf{x}) d^3\mathbf{x}$$

subject to

$$\alpha_j = \int_{S_g} f^*(\mathbf{k}_j^T \mathbf{x})g(\mathbf{x})d^3\mathbf{x}$$

for $j \in \{1, \dots, j_0\}$. Define

$$\begin{aligned} \bar{b}_l(\mathbf{x}) &= \sum_{\beta=0}^{59} f(\mathbf{R}_\beta \boldsymbol{\kappa}_l^T \mathbf{x}) \\ \bar{\Lambda}_{l,\nu} &= \int_{S_g} \frac{f^*(\boldsymbol{\kappa}_l^T \mathbf{x})\bar{b}_\nu(\mathbf{x})}{Q(\mathbf{x})} d^3\mathbf{x} \\ \bar{\alpha}_l &= \tilde{\alpha}_l - \int f^*(\boldsymbol{\kappa}_l^T \mathbf{x})g_0(\mathbf{x})d^3\mathbf{x}. \end{aligned}$$

Define $\bar{\Lambda} \in \mathcal{C}^{\bar{j}_0 \times \bar{j}_0}$ to have elements $\bar{\Lambda}_{l,\nu}$, $\bar{\alpha} \in \mathcal{C}^{\bar{j}_0}$ to have components $\bar{\alpha}_l$, and $\bar{\lambda} \in \mathcal{C}^{\bar{j}_0}$ by $\bar{\alpha} = \bar{\Lambda}\bar{\lambda}$. Then, the solution is

$$g(\mathbf{x}) = g_0(\mathbf{x}) + \sum_{l=1}^{\bar{j}_0} \frac{1}{Q(\mathbf{x})} \bar{b}_l(\mathbf{x})\bar{\lambda}_l.$$

Continue with the example of $b_j(\mathbf{x}) = \tau \exp(i\mathbf{k}_j^T \mathbf{x})$ for which $f(r) = \tau \exp(ir)$. The cube $S_g = S_c(\mathbf{s})$ is not symmetric and cannot be considered. However, the sphere $S_g = S_{g_0} = S_b(2R_+)$ is symmetric and, for the case $Q(\mathbf{x}) = 1$, it can be computed that $\bar{\Lambda}_{l,\nu} = \sum_{\beta=0}^{59} \zeta_b(|\boldsymbol{\kappa}_l - \mathbf{R}_\beta \boldsymbol{\kappa}_\nu|; 2R_+)$, and the basis functions are exactly the third cases in (9) and (10).

VIII. BASIS FUNCTION FAMILIES

Two families of basis functions that guarantee that the interpolator has various of the properties discussed in Section III are described in this section. Typically, it is not possible to do exact interpolation with a symmetric interpolator because the data do not exhibit the necessary symmetry. However, it is still possible to do least squares using the basis functions that are determined by the exact interpolation problem. Since $g(\mathbf{x})$ is real and $g(\mathbf{x}) = g(-\mathbf{x})$, it follows that both $\tau \int g(\mathbf{x}) \cos(\mathbf{k}^T \mathbf{x}) d\mathbf{x}$ and $\tau \int g(\mathbf{x}) \exp(i\mathbf{k}^T \mathbf{x}) d\mathbf{x}$ are valid expressions for $G(\mathbf{k})$ [the Fourier transform of $g(\mathbf{x})$]. The cosine form has practical advantages because it yields real basis functions and, if the weights are also real, it guarantees that the interpolator is real. Therefore, the first family of basis functions are the third cases in (9) and (10) modified by using the cosine definition. This choice guarantees that the interpolated $g(\mathbf{x})$ is symmetric, has the correct support, and is real.

In this paragraph, a set of basis functions that are orthonormal (unlike those of the prior paragraph) and for which the number of basis functions is naturally under the user's control are described. Because of the icosahedral symmetry (a rotational symmetry) and the maximum radius for the

region in which $g(\mathbf{x})$ may be nonzero, it is natural to use spherical coordinates in both real ($\mathbf{x} = (r, \theta, \phi)$) and reciprocal ($\mathbf{k} = (k, \theta', \phi')$) spaces. In order to easily compute the Fourier transform relating $g(\mathbf{x})$ to $G(\mathbf{k})$, it is natural to define real-space basis functions that are products of harmonic angular functions and spherical-Bessel radial functions. The angular function determines the rotational symmetry of the basis function. Spherical harmonics [17, eq. (3.53)], which are denoted by $Y_{l,m}(\theta, \phi)$ ($l \in \{0, 1, \dots\}$, $m \in \{-l, \dots, +l\}$) are a complete orthonormal (CON) basis for L_2 functions on the sphere. However, $g(\mathbf{x})$ must have icosahedral symmetry, and therefore, we only need a CON basis for the subspace of icosahedrally symmetric L_2 functions on the sphere. Such a basis is provided by icosahedral harmonics [14], [15], [18]–[20], which is denoted by $T_{l,n}(\theta, \phi)$ ($l \in \{0, 1, \dots\}$, $n \in \{0, 1, \dots, N_l - 1\}$). Use of icosahedral rather than spherical harmonics means that any superposition of these basis functions has icosahedral symmetry and that the number of basis functions for each l is markedly decreased, specifically, N_l (for which formulas are known [18]) versus $2l + 1$. Because $g(\mathbf{x}) = g(-\mathbf{x})$, there is an additional simplification: Only l even is required. For the radial functions, which determine the support of $g(\mathbf{x})$, it is straightforward to derive a CON basis on $(0, R_+)$ constructed from spherical Bessel functions by using Sturm–Liouville theory [21, ch. XVIII]. [Here, R_+ is twice the value used in the nominal model of Section IV for $\rho(\mathbf{x})$ because $g(\mathbf{x})$ is the autocorrelation function of $\rho(\mathbf{x})$]. The result of these calculations are the real, and reciprocal, space basis functions $b_{l,n,p}(\mathbf{x}) = H_{l,p}(r)T_{l,n}(\theta, \phi)$ and $B_{l,n,p}(\mathbf{k}) = h_{l,p}(k)T_{l,n}(\theta', \phi')$ ($l \in \{0, 2, 4, \dots\}$, $p \in \{1, 2, 3, \dots\}$, $n \in \{0, 1, \dots, N_l - 1\}$), where $H_{l,p}(r) = j_l(\gamma_{l,p}r)/n_{l,p}$, $h_{l,p}(k) = R_+^2 H'_{l,p}(R_+) j_l(kR_+) / (k^2 - \gamma_{l,p}^2)$, $\gamma_{l,p}$ is the p th root in ascending order of $j_l(\gamma_{l,p}R_+) = 0$, $n_{l,p} = \sqrt{R_+}/2 R_+ j'_l(\gamma_{l,p}R_+)$, $H'_{l,p}$ is the derivative of $H_{l,p}$, and $T_{l,n}(\theta, \phi) = \sum_{m=-l}^{+l} b_{l,n,m} Y_{l,m}(\theta, \phi)$ for known coefficients $b_{l,n,m}$. There exists an efficient numerical algorithm that can compute $j_l(x)$, $y_l(x)$, $j'_l(x)$, and $y'_l(x)$ simultaneously so that $h_{l,p}(k)$ can be computed without numerical derivatives. The number of basis functions used is determined by the truncation of the l sum (an angular bandwidth constraint) and p sum (a radial bandwidth constraint) and is therefore under user control.

IX. NUMERICAL METHODS AND RESULTS

In this section, we describe numerical results on synthetic data computed from the atomic-resolution structure of Cowpea Mosaic Virus (CpMV) [22], which contains 60×4341 nonhydrogen atoms. Each numerical experiment consists of several steps:

- 1) Compute synthetic data from the atomic-resolution structure. One viral particle is placed in each unit cell of a cubic P Bravais lattice where the unit cell of the lattice measures $a = 317.053 \text{ \AA}$ on a side. The symmetries of the icosahedral group are related to the coordinate axes of the lattice in the following fashion: The icosahedral group has three orthogonal two-fold symmetry axes. One of these symmetry axes is aligned with the z coordinate axis. The second and third of these symmetry axes,

which could be aligned with the x and y coordinate axes, respectively, are instead rotated 10° off of the x and y coordinate axes. This relationship between the icosahedral group and the coordinate axes of the lattice (which is really the space group of the crystal) implies that the only symmetry-related data points are $\mathbf{k} = (k_1, k_2, k_3)^T$ and $\mathbf{k}' = (-k_1, k_2, -k_3)^T$; therefore, the data is 30-fold redundant. The synthetic data are calculated by

$$G(\mathbf{k}) = \left| \sum_{n=1}^N \sum_{\beta=0}^{59} f_n \exp(-i\mathbf{k}^T \mathcal{R}_\beta \mathbf{x}_n) \right|^2$$

where \mathbf{x}_n and f_n are the location and the atomic scattering factor, respectively, of the n th atom. The specific \mathbf{k} used are \mathbf{k} on the reciprocal lattice [i.e., $\mathbf{k} = (2\pi/a)\mathbf{n}$ for $\mathbf{n} \in \mathcal{Z}^3$] such that $0 < |\mathbf{k}| \leq k_1$, where $k_1 = 2\pi \times 0.1 \text{ rad/\AA}$ or $2\pi \times 0.05 \text{ rad/\AA}$. For the 0.1 (0.05) data set, there are 7073 (56 540) measurements that are not symmetry related.

- 2) Approximately 80% of the computed data is randomly deleted. Two mechanisms are considered: a) In the first mechanism, independent and identically distributed Bernoulli trials with $p = 0.2$ are performed (one trial for each measurement), and if the trial is a success, then the measurement is retained. In the 0.05 cycles/ \AA realization used in these calculations, the number of retained and deleted points are 1399 and 5674, respectively, whereas the corresponding numbers for the 0.1 cycles/ \AA realization are 11 245 and 45 295. b) The second mechanism models the actual data collection process. Data are collected in a sequence of double-wedge shaped regions intersecting the origin, and data sets are incomplete because an insufficient number of regions are collected rather than because the data in an individual region are incomplete. This process is simulated by randomly generating two unit vectors denoted by $\hat{\mathbf{n}}_1$ and $\hat{\mathbf{n}}_2$, where the angle between them is 3° , and defining one boundary plane by requiring it to be normal to $\hat{\mathbf{n}}_1$ and to intersect the origin and likewise for the second plane and $\hat{\mathbf{n}}_2$. The double wedge is then the region between these two planes that is the region of reciprocal space such that $(\mathbf{k}^T \hat{\mathbf{n}}_1)(\mathbf{k}^T \hat{\mathbf{n}}_2) < 0$. Regions are added until the number of points falling within some region is roughly 20% of the total. In the 0.05 cycles/ \AA realization used in these calculations, there are 14 regions, and the number of retained and deleted points is 1405 and 5668, respectively.

- 3) The parameters that minimize the least squares cost function (3) are computed either by singular value decomposition (SVD) (4), [23, Sec. 2.6] or by direct solution of the normal equations by Gauss–Jordan elimination (GJE) (5), [23, Sec. 2.1]. The SVD is a superior numerical technique but requires more computation and storage than the GJE, in particular, storage proportional to $n_0 j_0$ rather than n_0^2 , and we do not have the resources required for the SVD approach for the larger calculations described here. Least squares is applied to the residual after subtraction of the nominal model $G_0(\mathbf{k})$ with $R_- = 95 \text{ \AA}$, $R_+ = 160 \text{ \AA}$, $\rho_c = 0$, and ρ_s determined from the data by least squares. Interpolation on the data directly rather than on the residuals leads to only slightly inferior performance. The parameters in the cost function are $\Xi = I_{j_0}$ (measurement variance) and $\lambda = 0$ (regularization parameter). The choice of $\lambda = 0$ means that there is no nominal model (a_0 and Σ) in

the least squares procedure. We focus on the harmonic basis functions ($R_+ = 320 \text{ \AA}$) because the number of parameters (j_0) can then be controlled. (Recall that for the exact-interpolation basis functions, the number of parameters is always equal to the number of observations, there is no natural way to delete basis functions, and the number of observations in the problem of real interest is large, e.g., 66 594 for TRSV). The calculations are carried out for a subset of the infinite set of harmonic basis functions, and the subset used is $l \in \{0, \dots, L\}$, $n \in \{0, \dots, N_l - 1\}$, and $p \in \{1, \dots, P = \lfloor k_1 R_+ \rfloor\}$, where L is variable and is described in the following.

4) The interpolator determined in Step 3) is evaluated at the \mathbf{k} values of the deleted measurements, and the results are compared with the values of the deleted measurements. Let $G(\mathbf{k})$ and $\hat{G}(\mathbf{k})$ be the values of the true and interpolated data at location \mathbf{k} . Three measures of performance used by crystallographers are computed:

$$R_2 = \frac{\sum'_{j=1}^{j_0} [G(\mathbf{k}_j) - \hat{G}(\mathbf{k}_j)]^2}{\sum'_{j=1}^{j_0} [G(\mathbf{k}_j)]^2}$$

$$R_1 = \frac{\sum'_{j=1}^{j_0} \left| \sqrt{G(\mathbf{k}_j)} - \sqrt{\hat{G}(\mathbf{k}_j)} \right|}{\sum'_{j=1}^{j_0} \sqrt{G(\mathbf{k}_j)}}$$

$$C = \frac{\sum'_{j=1}^{j_0} \sqrt{G(\mathbf{k}_j)} \sqrt{\hat{G}(\mathbf{k}_j)}}{\left[\sum'_{j=1}^{j_0} G(\mathbf{k}_j) \right]^{1/2} \left[\sum'_{j=1}^{j_0} \hat{G}(\mathbf{k}_j) \right]^{1/2}}$$

where \sum' indicates a sum over only those measurements that were deleted. We also compute a modified version of R_2 (which is denoted by R_2^{self}) in which the sums are over the retained rather than the deleted measurements, which is a normalized version of the least squares cost function at the minimum.

We first consider the case $k_1 = 2\pi \times 0.05 \text{ rad/\AA}$. Results using the SVD computation are shown in Tables I and II for a variety of L . Current unpublished algorithms due to Johnson and collaborators make only local interpolations and operate without fully exploiting the icosahedral symmetry. Let \mathbf{k}_0 be the location at which an interpolated value is desired. The 59 symmetry-related locations are computed by $\mathbf{k}_\beta = \mathbf{R}_\beta \mathbf{k}_0$. At each \mathbf{k}_β , an interpolation is performed using local values, where local is defined in the sense of $|\mathbf{k} - \mathbf{k}_\beta|$ rather than $d(\mathbf{k}, \mathbf{k}_\beta)$. Then, the 60 values from local interpolation are averaged to give the final value. Using such algorithms yields results $R_1 = 0.25$ and $C = 0.8$, and therefore, the algorithm described here, which computes global interpolations and fully exploits the icosahedral symmetry, is already improving performance by a factor of 2 when only 312 parameters are used. This interpolator is nearly an exact interpolator since R_2^{self} is small. Since the two deletion models give very similar results, we only considered the Bernoulli model in later computations.

We now consider the case $k_1 = 2\pi \times 0.1 \text{ rad/\AA}$, which is the size of problem that is of biological interest. Results using the SVD and GJE computations are shown in Table III for a variety of L . The larger problems are computed exclusively with GJE because we do not have the computational resources necessary to compute the SVD's, e.g., the SVD of a $j_0 = 11\,245$ by

TABLE I
RESULTS FOR $k_1 = 2\pi \times 0.05 \text{ RAD/\AA}$
USING THE BERNOULLI DELETION MODEL

	$L = 10$	$L = 20$	$L = 30$
R_2^{self}	0.000201	0.000198	0.0000273
R_1	0.315	0.223	0.145
R_2	0.00880	0.000920	0.000127
C	0.974	0.988	0.995
n_0	88	183	299

TABLE II
RESULTS FOR $k_1 = 2\pi \times 0.05 \text{ RADIANS/\AA}$ USING THE DOUBLE-WEDGE
DELETION MODEL

	$L = 10$	$L = 20$	$L = 30$
R_2^{self}	0.00184	0.000229	0.0000416
R_1	0.359	0.222	0.148
R_2	0.0787	0.000994	0.000196
C	0.957	0.989	0.994
n_0	88	183	299

TABLE III
RESULTS FOR $k_1 = 2\pi \times 0.1 \text{ RADIANS/\AA}$ USING THE BERNOULLI
DELETION MODEL

	$L = 36 \text{ SVD}$	$L = 36 \text{ GJE}$	$L = 60 \text{ GJE}$	$L = 75 \text{ GJE}$	$L = 85 \text{ GJE}$
R_2^{self}	0.0000998	0.0000998	0.0000380	0.0000239	0.0000183
R_1	0.312	0.312	0.239	0.203	0.184
R_2	0.000342	0.000342	0.000162	0.0000853	0.0000713
C	0.978	0.978	0.986	0.989	0.991
n_0	900	900	1797	2277	2573

$n_0 = 1836$ matrix for $L = 60$. Similarly, we stop at $L = 85$ because our current software for the computation of the $T_{l,n}(\theta, \phi)$ functions, which is based on recursions, becomes inaccurate for larger l . Even with these limitations, we have achieved useful interpolations since a typical refined atomic-resolution macromolecular structure only has an R_1 in the low to mid 20%'s.

X. DISCUSSION

In this paper, we consider 3-D interpolation in the presence of symmetry. The particular symmetry involved is the icosahedral symmetry of viral X-ray crystallography (which also occurs in quasi crystals and fullerenes) but any rotational symmetry would have equivalent results. We consider the problem from three points of view—least squares, Bayesian estimation, and minimum-energy bandlimited exact interpolation—and show that MEBLEI is a special case of Bayesian, which in turn is a special case of least squares. However, the MEBLEI and Bayesian approaches are still of interest because they suggest basis function families for use in least squares. An important aspect of the problem is that although a symmetric interpolator is desired, the data are not symmetric due to the presence of noise, and this aspect is naturally dealt with by the least

squares approach. Examples of interpolation in the region $0 < |\mathbf{k}| \leq 2\pi \times 0.1 \text{ rad/\AA}$ are described.

In order to definitively treat problems such as the tobacco ring spot virus problem described in Section I, several software issues must be addressed; it would be desirable to use SVD methods, but that requires the practical computation of the SVD of a large matrix (e.g., 6×10^4 by 3×10^3), stable software for high-order icosahedral harmonics (i.e., $l > 85$) must be developed, and more general space groups and more general positioning of the viral particle within the space group need to be implemented.

In this paper, we have considered interpolation when the intensities have icosahedral symmetry. A second related problem of great importance in crystallography is determination of symmetry. Specifically, if multiple copies of an object are present in the unit cell then, based on the intensities, what are the spatial relationships between the copies? These relationships are traditionally determined by so-called rotation function methods [4], which essentially search for the maximum of the correlation between the Patterson function (the inverse Fourier transform of the samples of $G(\mathbf{k})$ [1]) and the rotated and translated Patterson function. If the objects themselves have a symmetry, then the so-called locked rotation function method [24] can exploit the symmetry. When the object has icosahedral symmetry, a very efficient locked rotation function method can be based on the same icosahedral harmonics tools used in this paper.

REFERENCES

- [1] R. P. Millane, "Phase retrieval in crystallography and optics," *J. Opt. Soc. Amer. A*, vol. 7, no. 3, pp. 394–411, 1990.
- [2] B.-C. Wang, "Resolution of phase ambiguity in macromolecular crystallography," in *Methods in Enzymology*. New York: Academic, 1985, vol. 115, pp. 90–112.
- [3] A. J. Jerri, "The Shannon sampling theorem—Its various extensions and applications: A tutorial review," *Proc. IEEE*, vol. 65, pp. 1565–1596, Nov. 1977.
- [4] M. G. Rossman and D. M. Blow, "The detection of sub-units within the crystallographic asymmetric unit," *Acta Cryst.*, vol. 15, pp. 24–31, 1962.
- [5] J. R. Higgins, "A sampling theorem for irregularly spaced sample points," *IEEE Trans. Inform. Theory*, vol. IT-22, pp. 621–622, Sept. 1976.
- [6] J. L. Yen, "On nonuniform sampling of bandwidth-limited signals," *IRE Trans. Circuit Theory*, vol. CT-5, pp. 251–257, Dec. 1956.
- [7] F. J. Buetler, "Recovery of randomly sampled signals by simple interpolators," *Inform. Contr.*, vol. 26, pp. 313–340, 1974.
- [8] J. Jimenez and J. C. Agui, "Approximate reconstruction of randomly sampled signals," *Signal Process.*, vol. 12, pp. 153–168, 1987.
- [9] J. J. Clark, M. R. Palmer, and P. D. Lawrence, "A transformation method for the reconstruction of functions from nonuniformly spaced samples," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 1151–1165, Oct. 1985.
- [10] D. S. Chen and J. P. Allebach, "Analysis of error in reconstruction of two-dimensional signals from irregularly spaced samples," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, pp. 173–180, Feb. 1987.
- [11] L. Levi, "Fitting a bandlimited signal to given points," *IEEE Trans. Inform. Theory*, vol. IT-11, pp. 372–376, July 1965.
- [12] H. Fan and J. L. C. Sanz, "Comments on "Direct Fourier reconstruction in computer tomography,"" *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 446–449, Apr. 1985.
- [13] A. Erdelyi, Ed., *Higher Transcendental Functions*, New York: McGraw-Hill, 1953.
- [14] Y. Zheng and P. C. Doerschuk, "Explicit orthonormal fixed bases for spaces of functions that are totally symmetric under the rotational symmetries of a platonic solid," *Acta Cryst.*, vol. A52, pp. 221–235, 1996.
- [15] —, "Symbolic symmetry verification for harmonic functions invariant under polyhedral symmetries," *Comput. Phys.*, vol. 9, no. 4, pp. 433–437, July/Aug. 1995.
- [16] T. Schmidt, J. E. Johnson, and W. E. Phillips, "The spherically averaged structures of cowpea mosaic virus components by X-ray solution scattering," *Virology*, vol. 127, pp. 65–73, 1983.
- [17] J. D. Jackson, *Classical Electrodynamics*, 2nd ed. New York: Wiley, 1975.
- [18] O. Laporte, "Polyhedral harmonics," *Z. Naturforsch.*, vol. 3a, pp. 447–456, 1948.
- [19] J. Raynal, "On a labeling for point group harmonics—II: Icosahedral group," *J. Math. Phys.*, vol. 26, no. 10, pp. 2441–2456, Oct. 1985.
- [20] Y. Zheng and P. C. Doerschuk, "Iterative reconstruction of three-dimensional objects from averaged Fourier-transform magnitude: Solution and fiber X-ray scattering problems," *J. Opt. Soc. Amer. A*, vol. 13, no. 7, pp. 1483–1494, July 1996.
- [21] G. N. Watson, *A Treatise on the Theory of Bessel Functions*. London, U.K.: Cambridge Univ. Press, 1944.
- [22] Z. Chen, C. V. Stauffacher, and J. E. Johnson, "Capsid structure and RNA packaging in comoviruses," *Semin. Virol.*, vol. 1, pp. 453–466, 1990.
- [23] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C: The Art of Scientific Computing*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 1992.
- [24] L. Tong and M. G. Rossman, "The locked rotation function," *Acta Cryst.*, vol. A46, pp. 783–792, 1990.
- [25] P. M. Morse and H. Feshbach, *Methods of Theoretical Physics*. New York: McGraw-Hill, 1953.
- [26] N. V. Cohan, "The spherical harmonics with the symmetry of the icosahedral group," in *Proc. Camb. Philos. Soc.*, vol. 54, 1958, pp. 28–38.
- [27] A. Jack and S. C. Harrison, "On the interpretation of small-angle X-ray solution scattering from spherical viruses," *J. Mol. Biol.*, vol. 99, pp. 15–25, 1975.
- [28] J. Raynal, "Determination of point group harmonics for arbitrary j by a projection method—II: Icosahedral group, quantization along an axis of order 5," *J. Math. Phys.*, vol. 25, no. 5, pp. 1187–1194, May 1984.



Yibin Zheng (M'96) received the B.S.E.E. degree from Zhongshan University, China, in 1988, the M.A. and Ph.D. degrees in physics from the State University of New York, Buffalo, in 1992, and the Ph.D. degree in electrical engineering from Purdue University, West Lafayette, IN, in 1996.

He is currently with Corporate Research and Development, General Electric Company, Schenectady, NY. His research interests include statistical signal processing, numerical and symbolic computing for science and engineering, and mathematical physics.



Peter C. Doerschuk (M'86) received the B.S., M.S., and Ph.D. degrees in electrical engineering from the Massachusetts Institute of Technology (MIT), Cambridge, in 1977, 1979, and 1985, respectively, and the M.D. degree from Harvard Medical School, Cambridge, in 1987.

After postgraduate training at Brigham and Womens' Hospital, Boston, MA, he held a post-doctoral appointment at the Laboratory for Information and Decision Systems at MIT from January 1988 to August 1990. Since August 1990, he has been on the faculty in the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN.

John E. Johnson received the Ph.D. degree in physical chemistry from Iowa State University, Ames, in 1972.

After a post-doctoral appointment with the Department of Biological Sciences, Purdue University, West Lafayette, IN, he joined the Department as a faculty member in 1975, where he remained until 1995, when he joined The Scripps Research Institute, La Jolla, CA, as a Professor in the Department of Molecular Biology. His research is focused on the structural biology of viruses.