Acta Crystallographica Section A
Foundations and Advances
ISSN 2053-2733

Received 25 March 2013
Accepted 23 November 2013

Direct phasing in femtosecond nanocrystallography. I. Diffraction characteristics

Joe P. J. Chen,[a] John C. H. Spence[b] and Rick P. Millane[a]*

[a]Computational Imaging Group, Department of Electrical and Computer Engineering, University of Canterbury, Christchurch, New Zealand, and [b]Department of Physics, Arizona State University, Tempe, AZ 85287, USA. Correspondence e-mail: rick.millane@canterbury.ac.nz

X-ray free-electron lasers solve a number of difficulties in protein crystallography by providing intense but ultra-short pulses of X-rays, allowing collection of useful diffraction data from nanocrystals. Whereas the diffraction from large crystals corresponds only to samples of the Fourier amplitude of the molecular transform at the Bragg peaks, diffraction from very small crystals allows measurement of the diffraction amplitudes between the Bragg samples. Although highly attenuated, these additional samples offer the possibility of iterative phase retrieval without the use of ancillary experimental data [Spence *et al.* (2011). *Opt. Express*, **19**, 2866–2873]. This first of a series of two papers examines in detail the characteristics of diffraction patterns from collections of nanocrystals, estimation of the molecular transform and the noise characteristics of the measurements. The second paper [Chen *et al.* (2014). *Acta Cryst.* A**70**, 154–161] examines iterative phase-retrieval methods for reconstructing molecular structures in the presence of the variable noise levels in such data.

1. Introduction

The three major problems plaguing crystallographers since the dawn of X-ray crystallography are that of crystal preparation, radiation damage and phase determination. The first problem is significant as complex molecules are difficult to crystallize into well diffracting macrocrystals – membrane proteins being one prominent example (DeLucas, 2009). The second problem resides in the fact that the incident X-ray flux must be increased to obtain accurate high-resolution diffraction data, leading to a greater radiation dose that can damage the sample and counteract the gain in resolution. The final obstacle derives from the inability to measure the phase information of the incoming wavefront, thus preventing a direct inversion of the diffraction pattern *via* the inverse Fourier transform. This last difficulty is a result of crystal periodicity giving only discrete Bragg reflections that sample the amplitude of the molecular transform with no information between the Bragg samples. These data undersample the Fourier amplitude, leaving the phase problem underdetermined (Millane, 1990).

The recent development of a new type of X-ray source, the X-ray free-electron laser (XFEL), has the potential to solve all three problems *via* its ability to generate intense but extremely brief X-ray pulses. Appropriate signal levels can be attained whilst sidestepping radiation damage, as the duration of these X-ray pulses can be arranged such that they terminate before significant development of the damaging photoelectron cascade (Barty *et al.*, 2012), meaning that diffraction patterns from a crystal can be obtained before the molecular structure is significantly perturbed (Neutze *et al.*, 2000). The high-intensity X-ray pulse enables measurable diffraction data to be obtained from crystals only a few unit cells across, so-called 'nanocrystals'. Such small crystals have the advantage that they can aid the solution to the phase problem by providing information between the Bragg peaks (Sayre, 1952; Miao & Sayre, 2000), and at the same time being free of difficulties surrounding the growth of macroscopic protein crystals. The coherent diffraction imaging of the molecular density *via* these nanocrystals is thus referred to as 'nanocrystallography'.

The proposed approach using nanocrystals for direct phasing can be considered an alternative to approaches based on diffraction by noncrystalline specimens for which the diffraction is inherently stronger between the Bragg positions. There are two such approaches. The first approach is one-molecule-per-shot methods, which have shown some success for a very large virus particle for which projections of the electron density were reconstructed at low resolution (Seibert *et al.*, 2011). However, the diffraction from single, smaller molecules is extremely weak, whereas the use of nanocrystals concentrates the scattering and the coherent amplification of Bragg diffraction brings the signal far above the background. Furthermore, the use of nanocrystals also has the advantage that the diffraction patterns can be brought accurately into alignment and merged by indexing the Bragg spots, unlike single-particle studies, where determination of the relative orientations of successive molecules is more difficult than crystallographic indexing. The second approach uses diffraction of XFEL pulses by a collection of a large number of

randomly oriented single molecules (as in a solution) and the electron density is determined by analysis of the angular correlations using a method first proposed by Kam (1977). The potential of the Kam method for imaging using XFELs was described by Saldin *et al.* (2011) and subsequently demonstrated for simple test particles (Starodub *et al.*, 2012). While this method has the advantage of not having to utilize the suppressed diffraction between the Bragg positions for the case of nanocrystals and has considerable potential, it too has yet to be demonstrated for more complex, lower-symmetry particles or molecules. We also note that inelastic Compton scattering is negligible at low beam energies, but will form a smoothly varying background comparable with other forms of noise such as readout and electronic noise at higher energies.

The Linac Coherent Light Source (LCLS) located at the Stanford Linear Accelerator Center currently generates pulses of hard X-rays with pulse lengths between 10 and 300 fs with approximately 10^{12} photons per pulse (Chapman, 2009; Spence *et al.*, 2012). A liquid microjet injection set-up (Weierstall *et al.*, 2012) as shown pictorially in Fig. 1 is used to introduce fully hydrated nanocrystals into the pulsing XFEL beam. In the experiments undertaken so far, diffraction patterns are recorded from a succession of nanocrystals of different sizes in random orientations at the rate of about 100 snapshot diffraction patterns per second (Chapman *et al.*, 2011). The X-ray pulses are so brief that they terminate before the effects of radiation damage develop. This approach has been applied to the plant protein group Photosystem I (Kirian *et al.*, 2011) and hen egg white lysozyme (Boutet *et al.*, 2012), where the assembled data have been shown to be consistent with their known structures. Recently, the previously unknown structure of *Trypanosoma brucei* cysteine protease cathepsin B has also been solved using femtosecond nanocrystallography in conjunction with molecular replacement for phase determination (Redecke *et al.*, 2013). A method of phasing non-identical, randomly oriented nanocrystal snapshot diffraction patterns would thus address three of the most important problems of protein crystallography – crystal quality, radiation damage and the phase problem.

It is well known that if one has access to the continuous diffraction amplitude, *i.e.* between, as well as at, the reciprocal-lattice points, then the phase problem can be solved without any ancillary experimental data (Sayre, 1952; Millane, 1990; Miao *et al.*, 1999; Miao & Sayre, 2000). Using this fact, Spence *et al.* (2011) have proposed a novel phasing method using XFEL data from nanocrystals. Their strategy extracts an estimate of the amplitude of the continuous molecular transform from the millions of snapshot diffraction patterns from nanocrystals of varying sizes injected into the path of the femtosecond XFEL beam. The data from size-varying nanocrystals are indexed and merged in a three-dimensional diffraction volume before any phasing is attempted – individual nanocrystals are not phased. Averaging the intensity data around each Bragg peak allows the finite-size effect of the nanocrystals to be estimated without any prior knowledge of their size distribution, which is then used, together with the diffraction data, to estimate the molecular transform. Iterative phase retrieval can then in principle be used to reconstruct the protein electron density from the molecular transform. The inherent difficulty with this method, however, is that the diffraction signal between the Bragg reflections is weak and so the derived continuous molecular transform suffers from a low signal-to-noise ratio (SNR) in those regions, negatively affecting the success of phase retrieval. Furthermore, there is a trade-off in regards to nanocrystal size, since while smaller crystals give less amplification of the noise between the Bragg reflections, the overall diffracted intensity is weaker.

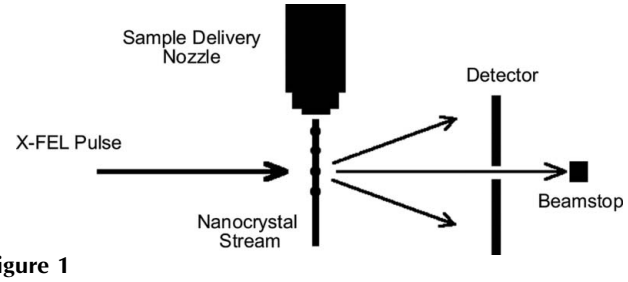

Figure 1
Experimental set-up of diffraction-pattern collection from a stream of nanocrystals.

Recently, Elser (2013) has proposed an alternative approach that uses the intensities as well as their gradients at the Bragg reflections to reconstruct the electron density from XFEL nanocrystal diffraction data. This avoids use of the weak amplitudes between the Bragg reflections but requires a rather noise-sensitive estimate of the intensity gradients.

In this paper we examine in detail the characteristics of diffraction patterns from a collection of nanocrystals of different sizes and shapes, estimation of the molecular transform from such diffraction patterns and the signal-to-noise characteristics of the molecular transform estimate that would be used for phasing. In the second paper (Chen *et al.*, 2014) we describe how iterative phasing methods can be adapted to ameliorate the effects of the particular distribution of signal-to-noise found in this problem.

2. Diffraction by finite crystals

We begin by considering diffraction by small crystals (see, *e.g.*, Vartanyants & Robinson, 2001; Als-Nielsen & McMorrow, 2011). We assume here that the nanocrystal lattice is well ordered. It is possible that disorder may be significant as a result of the large surface-to-volume ratio of nanocrystals however (Dilanian *et al.*, 2013). There are a variety of possible kinds of disorder that could be present, but there is currently little information available on disorder in protein nanocrystals (apart from some results from atomic force microscopy on nanocrystals larger than those we consider). However, the fact that diffraction patterns from nanocrystals as small as nine unit cells across show strong interference fringes (Chapman *et al.*, 2011) indicates that the effects of disorder are quite small, maybe because the crystals are smaller than the mosaic domains. Highly mobile molecules on the nanocrystal surface will contribute only diffuse background diffraction which would have little impact on the proposed phasing method.

Likewise, disorder that is dominated by uncorrelated distortions of the crystal lattice will decrease the amplitudes of the Bragg reflections with increasing resolution but will not change their profile shapes, and will add a slowly varying diffuse background (Stroud & Millane, 1996; Welberry, 2004), neither of which will have an effect on the methods proposed. If correlated distortions of sufficient magnitude are present, then the Bragg profiles would broaden with increasing resolution (Stroud & Millane, 1996; Welberry, 2004), which would require modification of the methods proposed. However, as mentioned above, there is little evidence of such distortions in diffraction patterns obtained so far from nanocrystals.

Consider first a one-dimensional molecule embedded inside a unit cell of length a on the interval $(-a/2, a/2)$ and denote the electron density inside this cell by $f(x)$. The unit cell is repeated N times along the x direction and centered on the origin, giving a one-dimensional crystal of size N as shown in Fig. 2. In this paper we consider only crystals consisting of an integral number of a single kind of unit cell. Effects due to different kinds of and incomplete unit cells that can occur when there is more than one molecule in the unit cell are discussed elsewhere (Chen & Millane, 2013; Liu *et al.*, 2014). The electron density of the crystal, $g(N, x)$, can be expressed as the convolution of $f(x)$ with a comb function consisting of a train of N delta functions spaced by a,

$$g(N, x) = f(x) \otimes \sum_{h=0}^{N-1} \delta\left(x - ha + \frac{1}{2}(N-1)a\right), \qquad (1)$$

where $\otimes$ denotes convolution. Note that since we have chosen the origin to be centered on the crystal, the position of the central unit cell(s) about the origin is different for N odd and N even, as shown in Fig. 2.

A general expression for hard X-ray scattering from nanocrystals in three dimensions is given by Kirian *et al.* (2010). In the one-dimensional case, the complex diffraction pattern in the far-field is proportional to $G(N, u)$, the Fourier transform of $g(N, x)$ (Als-Nielsen & McMorrow, 2011), so that

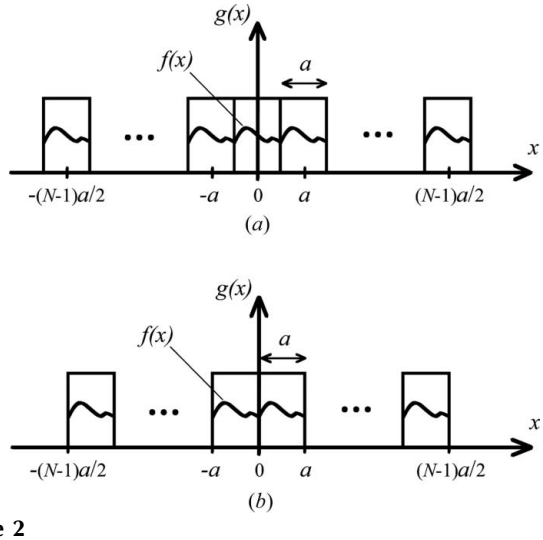

Figure 2
A one-dimensional crystal with N unit cells for (*a*) N odd and (*b*) N even.

$$G(N, u) = F(u) \exp\left(i\pi(N-1)au\right) \sum_{h=0}^{N-1} \exp(-i2\pi hau), \qquad (2)$$

with $F(u)$ being the Fourier transform of one unit cell, referred to henceforth as the molecular transform. Note that nanocrystal diffraction data are averaged over thousands of nanocrystals in different orientations in a procedure called Monte Carlo integration (Kirian *et al.*, 2010). This has the effect of averaging out different partialities for samples of the continuous intensity so that curvature of the Ewald sphere is automatically accounted for. Evaluating the geometric series in equation (2) gives

$$G(N, u) = F(u)S(N, u), \qquad (3)$$

where

$$S(N, u) = \frac{\sin(\pi Nau)}{\sin(\pi au)}. \qquad (4)$$

The intensity of the diffraction pattern from a crystal of this size, $I(N, u)$, is equal to $|G(N, u)|^2$, or

$$I(N, u) = \left|F(u)\right|^2 \left|S(N, u)\right|^2, \qquad (5)$$

such that the crystal-size-dependent function that modulates the intensities of the molecular transform is

$$\left|S(N, u)\right|^2 = \frac{\sin^2(\pi Nau)}{\sin^2(\pi au)}, \qquad (6)$$

and is referred to here as the 'shape transform' for a one-dimensional crystal of size N. The observed diffraction intensity from a crystal is therefore the squared magnitude of the molecular transform modulated by the shape transform. The effect of this modulation is illustrated in Fig. 3(*a*), where the reciprocal-space axis is normalized such that the Bragg peaks lie on integer values. As N tends to infinity, $|S(N, u)|^2$ converges to a comb function, modeling the situation in conventional crystallography with macrocrystals where only the Bragg reflections are observed.

An alternative description for diffraction by a finite crystal uses a truncated infinite crystal. The diffraction is then expressed as a convolution of the Bragg reflections from an infinite crystal with a shape transform that is different to $S(N, u)$. The relationship between these two descriptions is described in the Appendix.

2.1. A collection of small crystals

In femtosecond nanocrystallography, diffraction patterns are collected from many crystals of varying sizes as they pass randomly across a pulsing XFEL beam. Assuming that the orientation of each pattern can be determined by automated Miller indexing (Kirian *et al.*, 2010; White *et al.*, 2012), the averaged diffracted intensity can be obtained. This averaged diffracted intensity is derived as follows. Let angle brackets denote averaging over the corresponding subscript variables and let $I(N_n, u)$, $S(N_n, u)$ and N_n be the diffracted intensity, shape transform and the number of unit cells, for the nth nanocrystal, respectively. Writing $I(N_n, u)$ as $I_n(u)$ and $S(N_n, u)$ as $S_n(u)$, the diffracted intensity averaged over a large

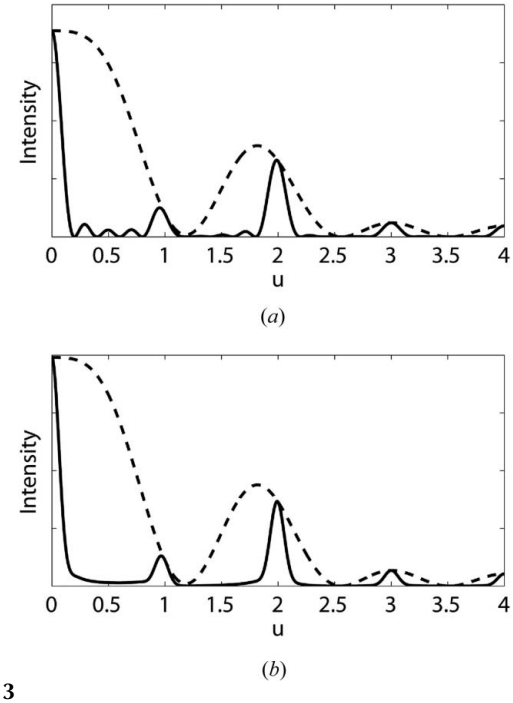

Figure 3
Diffracted intensity (solid line) (*a*) from a one-dimensional crystal of size five unit cells, and (*b*) averaged over a collection of one-dimensional crystals with mean size of five unit cells and standard deviation of 5/3 unit cells. The envelope (dashed line) is the squared amplitude of the molecular transform.

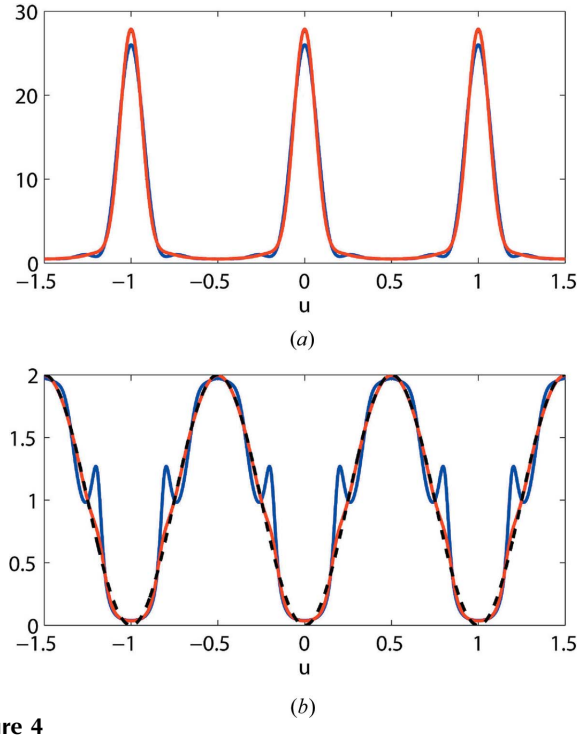

Figure 4
(*a*) The averaged shape transform and (*b*) its inverse for two Gaussian size distributions both with a mean of five unit cells but one with a standard deviation of 5/3 unit cells (red curve) and the other with a standard deviation of one unit cell (blue curve). The sinusoidal approximation for the inverse averaged shape transform given by equation (15) is shown by the dashed line in (*b*).

number of diffraction patterns from crystals of different sizes is then

$$\langle I_n(u)\rangle_n = \sum_N P(N) I(N,u), \tag{7}$$

where $P(N)$ is the probability density function (p.d.f.) of the distribution of nanocrystal sizes N. Substituting for the intensity of a single crystal from equation (5) gives

$$\langle I_n(u)\rangle_n = |F(u)|^2 Q^2(u), \tag{8}$$

where $Q^2(u)$ is the averaged shape transform function, *i.e.*

$$Q^2(u) = \langle |S_n(u)|^2\rangle_n = \sum_N P(N)|S(N,u)|^2$$
$$= \sum_N P(N)\frac{\sin^2(\pi N a u)}{\sin^2(\pi a u)}. \tag{9}$$

Fig. 3(*b*) shows a plot of $\langle I_n(u)\rangle_n$ with the sizes of the nanocrystals following a truncated Gaussian distribution with a mean of five unit cells and a standard deviation of 5/3 unit cells. The truncation is due to the requirement that $N \geq 1$ and in this case σ is chosen to be small enough that the size distribution is essentially Gaussian. Note that by summing over the shape transform for various sized crystals, the zeroes of $|S(N,u)|^2$ have been averaged out, making $Q^2(u)$ strictly positive for all u.

The maxima of the averaged shape transform occur at the Bragg peaks at $u_b = k/a$, where k is an integer. Evaluating $Q^2(u)$ *via* equation (9) at these positions and noting that $|S(N,u_b)|^2 = N^2$ yields

$$Q^2(u_b) = \sum_N P(N) N^2, \tag{10}$$

i.e. $Q^2(u)$ at the Bragg peaks is equal to the second moment of $P(N)$. For a Gaussian distribution of crystallite sizes, taking the continuum limit for N, we therefore have that

$$Q^2(u_b) \simeq \mu^2 + \sigma^2, \tag{11}$$

where μ and σ are the mean and standard deviation of the crystallite size distribution, respectively. For example, for the case above, $\sigma = \mu/3$, which results in $Q^2(u_b) \simeq 1.1\mu^2$. In general, unless the crystallite size distribution is quite broad, we have that $Q^2(u_b) \simeq \mu^2$.

The minimum values of the averaged shape transform occur halfway between adjacent Bragg peaks at $u_{bh} = (2k-1)/2a$. Substituting this value of u into equation (9) gives

$$Q^2(u_{bh}) = \sum_{N\,\mathrm{odd}} P(N). \tag{12}$$

It is reasonable to assume that $\sum_{N\,\mathrm{odd}} P(N) \simeq \sum_{N\,\mathrm{even}} P(N)$ and since $\sum_N P(N) = 1$, the minimum value of the averaged shape transform is

$$Q^2(u_{bh}) \simeq 1/2 \tag{13}$$

for any $P(N)$.

2.2. The inverse averaged shape transform

The inverse of the averaged shape transform is of relevance later in this paper. Using equation (9) shows that

$$\frac{1}{Q^2(u)} = 2\sin^2(\pi a u)\left[1 - \sum_N P(N)\cos(2\pi N a u)\right]^{-1}. \quad (14)$$

The summation in equation (14) approximates a Poisson sum and therefore tends to vanish except when the distribution $P(N)$ is so narrow such that there are an insufficient number of cosine terms in the summation for cancellation. Therefore, we expect that except for narrow crystallite size distributions

$$\frac{1}{Q^2(u)} \simeq 2\sin^2(\pi a u). \quad (15)$$

The function $1/Q^2(u)$ and the approximation [equation (15)] are shown in Fig. 4(*b*) for two crystal size distributions with the same mean but different standard deviations. The inverse averaged shape transform is closely approximated by equation (15), except for deviations occurring for the narrower size distribution.

3. The three-dimensional case

In reality the crystals are three dimensional. Following the derivations above, for a crystal of dimension $N_1 \times N_2 \times N_3$ unit cells, the diffracted complex amplitude is given by

$$G(\mathbf{N}, \mathbf{u}) = F(\mathbf{u})S(\mathbf{N}, \mathbf{u}), \quad (16)$$

where $\mathbf{u}$ is the position vector in reciprocal space, $\mathbf{N} = (N_1, N_2, N_3)$ is a column vector of the number of unit cells in each crystal direction, $\mathbf{a}_1$, $\mathbf{a}_2$ and $\mathbf{a}_3$ are the crystal axis of the unit cell, and $S(\mathbf{N}, \mathbf{u})$ is the three-dimensional shape transform, *i.e.*

$$S(\mathbf{N}, \mathbf{u}) = \frac{\sin(\pi N_1 \mathbf{a}_1 \cdot \mathbf{u})}{\sin(\pi \mathbf{a}_1 \cdot \mathbf{u})} \cdot \frac{\sin(\pi N_2 \mathbf{a}_2 \cdot \mathbf{u})}{\sin(\pi \mathbf{a}_2 \cdot \mathbf{u})} \cdot \frac{\sin(\pi N_3 \mathbf{a}_3 \cdot \mathbf{u})}{\sin(\pi \mathbf{a}_3 \cdot \mathbf{u})}. \quad (17)$$

The diffracted intensity is then given by

$$I(\mathbf{N}, \mathbf{u}) = |F(\mathbf{u})|^2 |S(\mathbf{N}, \mathbf{u})|^2. \quad (18)$$

Consider now a distribution of crystal sizes with p.d.f. $P(\mathbf{N}) = P(N_1, N_2, N_3)$ such that $P(\mathbf{N})\mathrm{d}\mathbf{N}$ is the probability of a crystal size falling within the interval $(N_1, N_1 + \mathrm{d}N_1)$, $(N_2, N_2 + \mathrm{d}N_2)$ and $(N_3, N_3 + \mathrm{d}N_3)$. A wide range of forms for $P(\mathbf{N})$ is theoretically possible but we assume here, for simplicity, that the p.d.f. is jointly normal, *i.e.*

$$P(\mathbf{N}) = \frac{1}{[(2\pi)^3 \det(C)]^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{N} - \boldsymbol{\mu})^{\mathrm{T}} C^{-1} (\mathbf{N} - \boldsymbol{\mu})\right), \quad (19)$$

where $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)$ is a column vector of the mean crystal sizes in each crystal axis direction, C is the covariance matrix given by

$$C = \begin{pmatrix} \sigma_1^2 & \sigma_{12}^2 & \sigma_{13}^2 \\ \sigma_{12}^2 & \sigma_2^2 & \sigma_{23}^2 \\ \sigma_{13}^2 & \sigma_{23}^2 & \sigma_3^2 \end{pmatrix}, \quad (20)$$

and $\det(C)$ and C^{-1} denote the determinant and inverse of C, respectively. If the crystal lengths N_1, N_2 and N_3 are uncorrelated with each other, *i.e.* the correlation coefficients $\rho_{ij} = \sigma_{ij}^2/\sigma_i\sigma_j$ satisfy $\rho_{12} = \rho_{13} = \rho_{23} = 0$, then the covariance matrix is diagonal. Because we assume a Gaussian distribution, the length of a crystal in one direction when the correlation coefficients are zero is independent of the length in the other directions. In this case, the p.d.f. factorizes as

$$P(\mathbf{N}) = P(N_1)P(N_2)P(N_3), \quad (21)$$

where

$$P(N_i) = \frac{1}{(2\pi)^{1/2}\sigma_i} \exp\left[-\frac{(N_i - \mu_i)^2}{2\sigma_i^2}\right], \quad (22)$$

for $i = 1, 2, 3$. In general, we may suppose that complete independence of the crystal lengths is relatively unlikely since a crystal that grows in one direction is likely to grow in the other directions as well, *i.e.* needle-like or high aspect ratio crystals are special. Therefore, it is more probable that the crystal sizes N_1, N_2 and N_3 are positively correlated.

At the other extreme, consider the case where the crystal sizes in each direction are completely correlated, *i.e.* the correlation coefficients satisfy $\rho_{12} = \rho_{13} = \rho_{23} = 1$. The joint density is then degenerate with only one degree of freedom, *i.e.* an instance of one random variable fixes the other two. Without loss of generality, assume that N_1 is known, drawn from the single-variable Gaussian distribution given by equation (22) for $i = 1$, then the numbers of unit cells in the other two directions are given by

$$N_i = \frac{\sigma_i}{\sigma_1}(N_1 - \mu_1) + \mu_i, \quad (23)$$

for $i = 2, 3$. Equation (23) can be written as $(N_i - \mu_i)/(N_1 - \mu_1) = \sigma_i/\sigma_1$ which shows that the ratio of the difference between the length of each crystal edge and its mean is constant.

However, it is also unlikely in practice that the length of one crystal edge will fix the other two edges, so it is reasonable that the correlation coefficients satisfy $0 < \rho_{ij} < 1$. Regardless of the detailed form of $P(\mathbf{N})$, for a collection of crystals the averaged shape transform in three dimensions is given by

$$Q^2(\mathbf{u}) = \sum_{\mathbf{N}} P(\mathbf{N})|S(\mathbf{N}, \mathbf{u})|^2. \quad (24)$$

If the three crystallite edge lengths are independent, equation (24) is separable and can be written as the product of three one-dimensional averaged shape transforms, *i.e.*

$$Q^2(\mathbf{u}) = \prod_{i=1}^{3} Q^2(u_i) = \prod_{i=1}^{3} \sum_{N_i} P(N_i)|S(N_i, u_i)|^2. \quad (25)$$

The results derived in §2.2 for the maxima and minima of the one-dimensional averaged shape transform are easily extended to three dimensions. With $\mathbf{u}_\mathrm{b}$ and $\mathbf{u}_\mathrm{bh}$ denoting the Bragg and half-integer Bragg positions in three dimensions, respectively, the maxima of $Q^2(\mathbf{u})$ have the value

$$Q^2(\mathbf{u}_\mathrm{b}) = \sum_{\mathbf{N}} P(\mathbf{N}) N_1^2 N_2^2 N_3^2, \quad (26)$$

which is approximately equal to $\mu_1^2\mu_2^2\mu_3^2$ for a Gaussian distribution that is not too broad. The minima of $Q^2(\mathbf{u})$ have the value

$$Q^2(\mathbf{u}_{\rm bh}) \simeq 1/8. \tag{27}$$

Example crystallite size p.d.f.s for a two-dimensional crystal and the resulting averaged shape transforms are shown in Fig. 5 for $\mu_1 = \mu_2 = 10$, $\sigma_1 = \sigma_2 = 3$, and for three values of the correlation coefficient. Fig. 5 shows that increasing the correlation between the number of unit cells along each side of the crystal has only a small effect on the averaged shape transform. This is shown more clearly in Fig. 6, which shows a cut along $u_2 = 1$ through the center of the shape transform. Increasing the correlation slightly sharpens the interference function, but the effect is small.

The specific form of $S(\mathbf{N}, \mathbf{u})$ in equation (17) assumes that the crystals are all parallelepipeds. In reality, the crystals will adopt a variety of different sizes and shapes. This is of no significance in practice, however, since $Q^2(\mathbf{u})$ in equation (24) is the average over the appropriate set of shape transforms $S(\mathbf{N}, \mathbf{u})$, and $Q^2(\mathbf{u})$ is estimated from the diffraction data and not calculated using equation (24). As described by Dilanian *et al.* (2013), other effects such as crystal disorder are also automatically incorporated into $Q^2(\mathbf{u})$ that is estimated from the diffraction data.

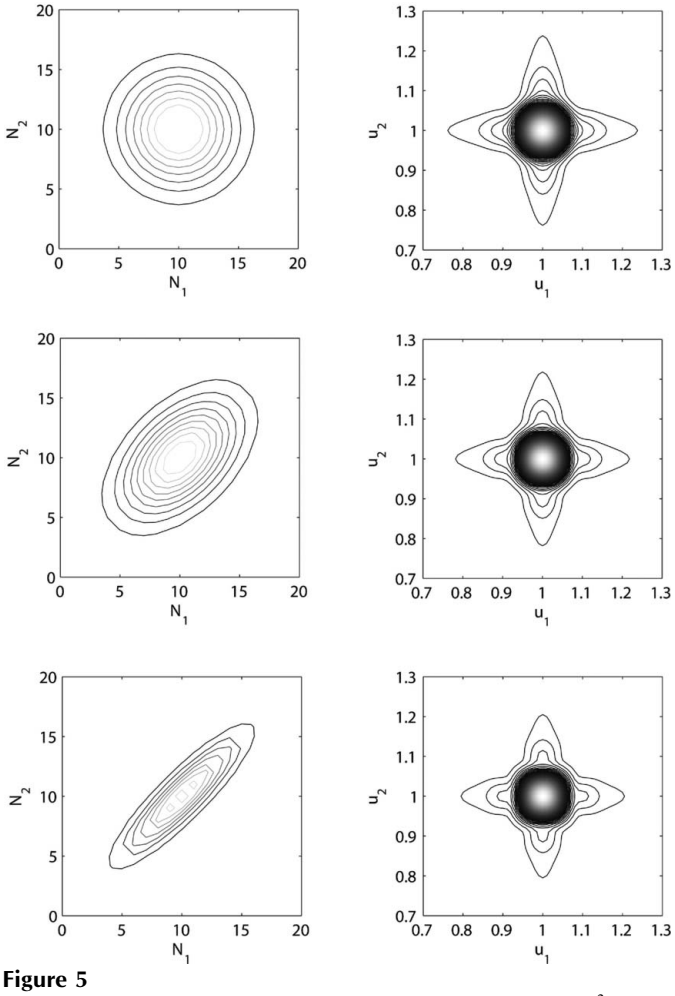

Figure 5
P.d.f.s (left) and corresponding averaged shape transforms $Q^2(u_1, u_2)$ (right) for $\rho_{12} = 0$ (top), $\rho_{12} = 0.5$ (middle), $\rho_{12} = 0.9$ (bottom).

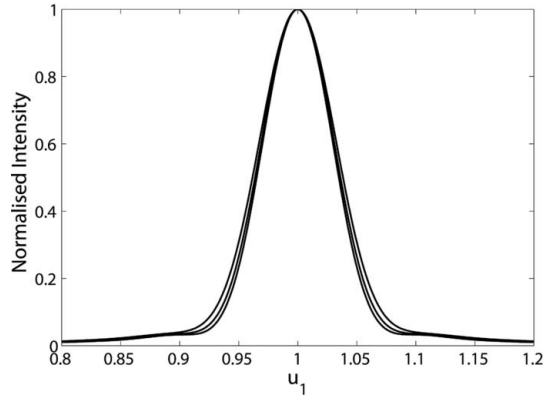

Figure 6
Normalized averaged shape transform profiles $Q^2(u_1, 1)$ for $\rho = 0$ (top curve), $\rho = 0.5$ (center curve) and $\rho = 0.9$ (bottom curve).

4. Estimating the molecular transform

It is clear from equation (8) that the magnitude of the molecular transform can be obtained by dividing the averaged measured intensity by the averaged shape transform, *i.e.* in the one-dimensional case

$$\left|F(u)\right|^2 = \frac{\langle I_n(u)\rangle_n}{Q^2(u)}. \tag{28}$$

In general, the nanocrystal size distribution $P(N)$, and thence $Q^2(u)$, is unknown. However, Spence *et al.* (2011) describe a method for estimating $Q^2(u)$ directly from the diffraction data using the fact that the shape transform is the same around each Bragg reflection, whereas the molecular transform is different but slow-varying in the vicinity of a Bragg peak. Averaging the diffraction intensities over fixed regions around many different reflections should preserve the shape transform but encourage the molecular transform to tend towards a constant value, *i.e.* smoothing it out to generate an average proportional to the shape transform for that particular diffraction pattern. Thus averaging over all Bragg peaks in many patterns from crystals of varying sizes gives an estimate of the averaged shape transform around one Bragg peak, which can be translated periodically throughout the reciprocal lattice to yield an estimate of the averaged shape transform function, as shown in Fig. 7. Spence *et al.* (2011) have simulated this process for realistic experimental conditions and shown that it is feasible to generate an estimate of the molecular transform magnitude using equation (28) by using a sufficiently large number of diffraction patterns.

We show here that such a procedure does converge to the averaged shape transform. Define a 'Bragg region' associated with each Bragg peak as a region with boundaries equidistant between that peak and the nearest neighboring peaks. The diffracted intensity in each Bragg region is then averaged over all Bragg regions. Let $\langle\rangle_{\rm b}$ denote the average over the Bragg regions and $\langle\rangle_{{\rm b},n}$ the average over the Bragg regions from all

diffraction patterns, then referring to equation (5), averaging over all Bragg regions from all patterns gives

$$\left\langle I_n(u-u_\mathrm{b})\right\rangle_{\mathrm{b},n} = \left\langle \left|F(u-u_\mathrm{b})\right|^2 \left|S_n(u-u_\mathrm{b})\right|^2 \right\rangle_{\mathrm{b},n}. \quad (29)$$

Since the shape transform is invariant under translations u_b, $|S_n(u-u_\mathrm{b})|^2 = |S_n(u)|^2$, and since the molecular transform is the same for all patterns, it is independent of n. Furthermore, the molecular transform and the shape transform are uncorrelated, so that equation (29) reduces to

$$\left\langle I_n(u-u_\mathrm{b})\right\rangle_{\mathrm{b},n} = \left\langle \left|F(u-u_\mathrm{b})\right|^2 \right\rangle_\mathrm{b} Q^2(u). \quad (30)$$

Consider now the average over the molecular transform, $\langle |F(u-u_\mathrm{b})|^2\rangle_\mathrm{b}$, in equation (30). Denote the sum of $|F(u)|^2$ over B Bragg regions, after shifting them to the origin, by $K_B(u)$, so that

$$K_B(u) = \sum_b^{(B)} \left| F\left(u - \frac{b}{a}\right)\right|^2, \quad (31)$$

where $\sum_b^{(B)}$ denotes the sum over B Bragg regions indexed by b. Equation (31) can be written as

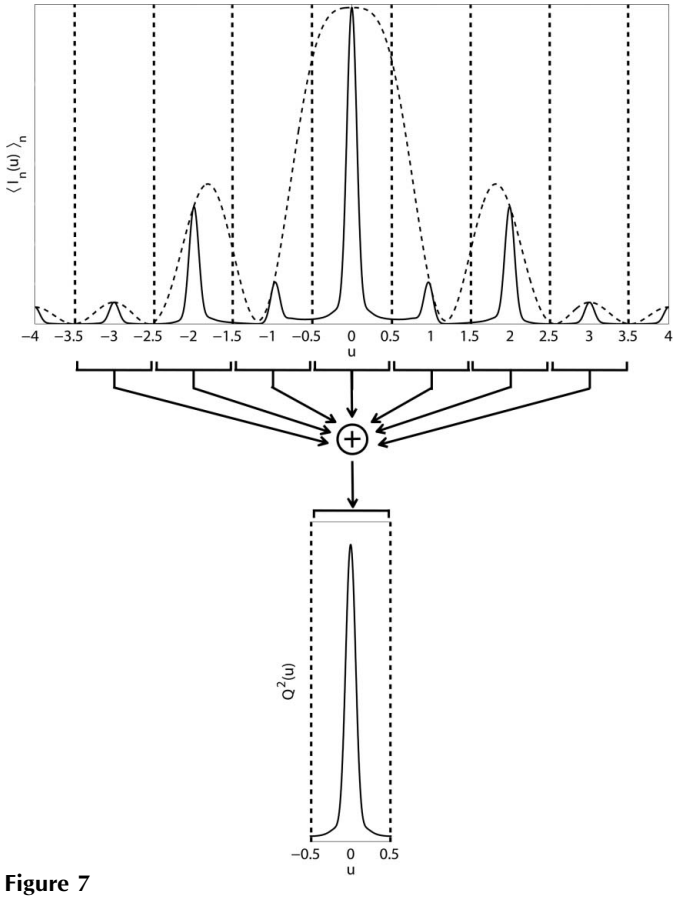

Figure 7
The averaged shape transform recovery process. The Bragg regions of the averaged diffracted intensity are bounded by vertical dotted lines which are fed into an averager represented by a circle with a plus symbol. The output yields the averaged shape transform around one Bragg region.

$$K_B(u) = |F(u)|^2 \otimes \sum_b^{(B)} \delta\left(u - \frac{b}{a}\right). \quad (32)$$

Letting $B \to \infty$ and taking the inverse Fourier transform, equation (32) becomes

$$k_\infty(x) = aA(x) \sum_{b=-\infty}^{\infty} \delta(x - ba), \quad (33)$$

where $A(x) = f(x) \otimes f(-x)$ is the autocorrelation of the molecular density $f(x)$ and $k_\infty(x)$ is the inverse Fourier transform of $K_\infty(u)$. Since $f(x)$ is zero outside $(-a/2, a/2)$, $A(x)$ is zero outside $(-a, a)$, so that equation (33) collapses to

$$k_\infty(x) = aA(0)\delta(x) = a \int_{-\infty}^{\infty} f^2(x)\,\mathrm{d}x\,\delta(x). \quad (34)$$

Taking the Fourier transform of equation (34) gives

$$K_\infty(u) = a \int_{-\infty}^{\infty} f^2(x)\,\mathrm{d}x, \quad (35)$$

i.e. for $B \to \infty$, $K_B(u)$ tends to a constant, which we denote K. Therefore, the average $\langle |F(u-u_\mathrm{b})|^2\rangle_\mathrm{b}$ tends to the constant K/B for a large number of reflections. Since B is large but finite in practice, this shows that, referring to equation (30), the average over the Bragg regions converges to a function proportional to the averaged shape transform, that is

$$Q^2(u) \propto \left\langle I_n(u-u_\mathrm{b})\right\rangle_{\mathrm{b},n}, \quad (36)$$

as required. The above analysis extends straightforwardly to three dimensions. Thus, the average over all Bragg reflections from all patterns generates an estimate of the averaged shape transform which can be inserted into equation (28) to estimate the molecular transform from the measured intensity.

5. Noise characteristics

In practice, the small values of the averaged shape transform between the Bragg reflections make estimates of the molecular transform in these regions by the procedure described in §4 noise sensitive. Heuristically, if we additively decompose the noisy measured average intensity into its noiseless component, $\langle I_n(u)\rangle_n$, and the noise contribution, noise(u), then the Fourier magnitude data used for phasing, denoted $|F(u)|_\mathrm{P}$, are related to the measured data by

$$\left\langle I_n(u)\right\rangle_n + \mathrm{noise}(u) = \left|F(u)\right|_\mathrm{P}^2 Q^2(u). \quad (37)$$

Rearranging and substituting for $\langle I_n(u)\rangle_n$ from equation (8) shows that

$$\left|F(u)\right|_\mathrm{P}^2 = \left|F(u)\right|^2 + \frac{\mathrm{noise}(u)}{Q^2(u)}, \quad (38)$$

so that the measurement noise is amplified in the phasing magnitude by the inverse of the averaged shape transform, with the amplification being smallest at the Bragg peaks and largest halfway between, as illustrated in Fig. 8. The statistics of this noise amplification process are now treated more rigorously.

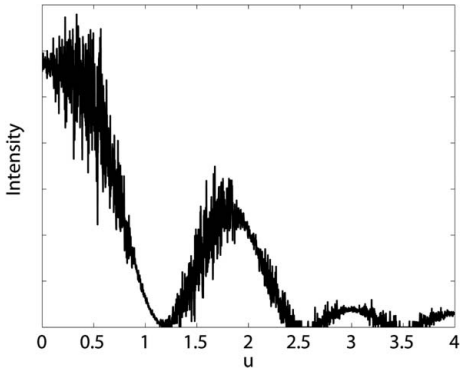

Figure 8
Illustration of the intensity of the molecular transform recovered by dividing a noisy averaged intensity by the averaged shape transform. Note that the resulting noise is a minimum at the Bragg positions and a maximum halfway between.

5.1. Noise statistics for a large number of patterns

In practice, the detector pixel positions, for a particular nanocrystal orientation, are mapped into reciprocal space and then resampled onto a grid for processing. Let the mapped detector location in reciprocal space for the ith detector pixel be u_i. For ease of description, we consider the intensity measurements at the detector. We denote by I_i the value of the squared amplitude of the molecular transform at the ith detector pixel, and by $\mathcal{I}_{in}$ the value of I_i modulated by the shape transform for the nth nanocrystal at that point, *i.e.* $\mathcal{I}_{in} = |S_{in}|^2 I_i$ where $S_{in} = S_n(u_i)$ is the shape transform of the nth nanocrystal at the ith detector pixel. What is recorded at the detector is a noisy measurement of $\mathcal{I}_{in}$, which we denote $\hat{\mathcal{I}}_{in}$. The averaged measured intensity at the ith pixel, denoted $I_{\mathrm{M}i}$, is calculated by averaging over the available snapshot diffraction patterns so that

$$I_{\mathrm{M}i} = \frac{1}{m}\sum_{n=1}^{m} \hat{\mathcal{I}}_{in}, \tag{39}$$

where m is the number of available diffraction patterns. Note that, in general, m depends on i, but here we use the average value m over the data set. We denote by $I_{\mathrm{P}i}$ the estimate of I_i that is calculated from the data $I_{\mathrm{M}i}$ and is used for phasing, so that $I_{\mathrm{P}i} = I_{\mathrm{M}i}/\hat{Q}_i^2$, where $\hat{Q}_i^2$ is the estimate of the averaged shape transform at pixel i, *i.e.* $Q_i^2 = Q^2(u_i)$, obtained as described in §4. Our goal is to find the statistical distribution of $I_{\mathrm{P}i}$ induced by random fluctuations in the detector measurements.

Since the recorded diffraction patterns are weak, we assume that the measurements are dominated by photon noise. For a photon-limited system, the noisy detector output is governed by a Poisson distribution with mean and variance $\left|S_{in}\right|^2 I_i$, and we write this as

$$\hat{\mathcal{I}}_{in} \sim \mathrm{Po}\left(\left|S_{in}\right|^2 I_i\right), \tag{40}$$

where $\sim$ denotes 'is distributed as' and Po(λ) denotes the Poisson distribution with parameter λ. Averaging $\hat{\mathcal{I}}_{in}$ over m patterns, the resulting $I_{\mathrm{M}i}$ is approximately normally distributed for m large. This result is guaranteed by the central limit theorem under the conditions of Lyapunov (Karr, 1993) which applies to situations where the random variables summed are independent but not identically distributed.

The $\hat{Q}_i^2$ are estimated by averaging the intensities over the Bragg reflections as well as over the diffraction patterns so that the errors in $\hat{Q}_i^2$ are expected to be smaller than those in $I_{\mathrm{M}i}$. Thus, we replace $\hat{Q}_i^2$ by Q_i^2, *i.e.* we take $\hat{Q}_i^2$ to be exact. It is then readily shown using equation (39) that for large m, the averaged measured intensity at the ith pixel is normally distributed with mean $Q_i^2 I_i$ and variance $Q_i^2 I_i/m$, or

$$I_{\mathrm{M}i} \sim \mathcal{N}\left(Q_i^2 I_i, \frac{1}{m} Q_i^2 I_i\right), \tag{41}$$

where $\mathcal{N}(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2. In an ideal noiseless system, $I_{\mathrm{M}i}$ is of course equal to $Q_i^2 I_i$, which means that the noise on the measured averaged intensity is normally distributed with a mean of zero and a variance of $Q_i^2 I_i/m$ by equation (41). Since $I_{\mathrm{P}i}$ is obtained by dividing $I_{\mathrm{M}i}$ by Q_i^2, it is also approximately normally distributed and

$$I_{\mathrm{P}i} \sim \mathcal{N}\left(I_i, \frac{I_i}{mQ_i^2}\right), \tag{42}$$

so that the noise on the phasing intensity is governed by a zero mean normal distribution with variance $I_i/(mQ_i^2)$.

5.2. Quantifying the noise amplification

Having obtained distributions for $I_{\mathrm{M}i}$ and $I_{\mathrm{P}i}$ along with the distributions for their respective noise terms, the signal-to-noise ratio can be computed. For a collection of strictly positive measurements, the SNR can be defined as the mean of the signal divided by the standard deviation of the noise (or, equivalently, of the signal) (Kundu, 2010). Let $\mathrm{SNR}_{\mathrm{M}i}$ and $\mathrm{SNR}_{\mathrm{P}i}$ denote the SNR for the measured intensity $I_{\mathrm{M}i}$ and the phasing intensity $I_{\mathrm{P}i}$, respectively. Using the expressions for the mean and standard deviation of the noise signal derived above, we have that

$$\mathrm{SNR}_{\mathrm{M}i} = \mathrm{SNR}_{\mathrm{P}i} = (mQ_i^2 I_i)^{1/2}, \tag{43}$$

i.e. that the SNRs for individual pixels after the averaging and division steps are the same. The overall SNR of the whole data set (all p pixels) is of interest and is calculated as the mean of the signal over all detector pixels, divided by the square root of the mean of the variances of the noise over all detector pixels, and is denoted $\mathrm{SNR}_{\mathrm{M}}$ and $\mathrm{SNR}_{\mathrm{P}}$, for the measured and phasing intensities, respectively. Using equations (41) and (42) gives

$$\mathrm{SNR}_{\mathrm{M}} = \left(\sum_{i=1}^{p} Q_i^2 I_i\right)^{1/2} \left(\frac{m}{p}\right)^{1/2} \tag{44}$$

and

$$\mathrm{SNR}_{\mathrm{P}} = \frac{\sum_{i=1}^{p} I_i}{\left(\sum_{i=1}^{p} I_i/Q_i^2\right)^{1/2}} \left(\frac{m}{p}\right)^{1/2}. \tag{45}$$

To simplify these expressions, let there be L pixels in each Bragg region and B Bragg regions altogether, then $p = LB$. The fact that all Bragg regions have identical averaged shape transforms means that

$$\sum_{i=1}^{p} Q_i^2 = B \sum_{l=1}^{L} Q_l^2, \tag{46}$$

where the sum over l is over the pixels in one Bragg region. As shown in the previous section, summing the molecular transform intensity over many Bragg reflections converges to a constant, so we can write

$$\sum_{i=1}^{p} I_i = L \sum_{b=1}^{B} I_b, \tag{47}$$

where the sum over b is over the equivalent pixel positions in all Bragg regions. Equations (46) and (47) imply that

$$\sum_{i=1}^{p} Q_i^2 I_i = \frac{1}{p}\left(\sum_{i=1}^{p} Q_i^2\right)\left(\sum_{i=1}^{p} I_i\right). \tag{48}$$

Substituting this result into equation (44) for $\mathrm{SNR_M}$ and using an identical argument to factor the sum in the denominator of equation (45) involving the inverse averaged shape transform for $\mathrm{SNR_P}$ gives

$$\mathrm{SNR_M} = \frac{1}{p}\left[m\left(\sum_{i=1}^{p} Q_i^2\right)\left(\sum_{i=1}^{p} I_i\right)\right]^{1/2} \tag{49}$$

and

$$\mathrm{SNR_P} = \left[m\left(\frac{1}{\sum_{i=1}^{p} Q_i^{-2}}\right)\left(\sum_{i=1}^{p} I_i\right)\right]^{1/2}. \tag{50}$$

The overall measured SNR, $\mathrm{SNR_M}$, thus increases as the square root of the total number of patterns m and decreases in proportion to the inverse of the number of detector pixels. This makes sense as the former determines the number of different measurements at the same location while the latter determines the number of different measurements at different locations.

The two SNRs were calculated for a one-dimensional diffraction experiment as a function of mean crystallite size. A Gaussian crystal size distribution with standard deviation equal to one third of the mean as in §2.1 was used. The sum $\sum_{i=1}^{p} I_i$ in equations (49) and (50) was set to unity, as was m, as these do not affect the dependence on mean crystal size. A total of 11 Bragg peaks were simulated and the number of detector pixels $p = 1101$. The results are shown *versus* mean crystallite size in Fig. 9(*a*).

The behaviors of $\mathrm{SNR_M}$ and $\mathrm{SNR_P}$ in Fig. 9(*a*) can be explained as follows. $\mathrm{SNR_M}$ and $\mathrm{SNR_P}$ are proportional to the square root of $\sum_i Q_i^2$ and $1/\sum_i Q_i^{-2}$, respectively. As shown in §2.1, the maximum value of Q_i^2 is proportional to the second moment of the crystal size distribution, which increases without bound as the mean crystallite size increases. This explains the monotonic increase of $\mathrm{SNR_M}$ with mean crystallite size. For $\mathrm{SNR_P}$, the increase in SNR for larger crystals, due to stronger scattering, is balanced by the division by a narrower averaged shape transform for the larger crystals, and $\mathrm{SNR_P}$ approaches a constant as the mean crystallite size increases. An approximation to $\mathrm{SNR_P}$ can be obtained as follows. Using equation (15), the sum $\sum_{i=1}^{p} Q_i^{-2}$ over a large number of detector pixels is given approximately by $p\langle 2\sin^2(\pi a u)\rangle_u = p$, and substituting into equation (50) gives

$$\mathrm{SNR_P} \simeq \left(\frac{m}{p}\sum_{i=1}^{p} I_i\right)^{1/2}. \tag{51}$$

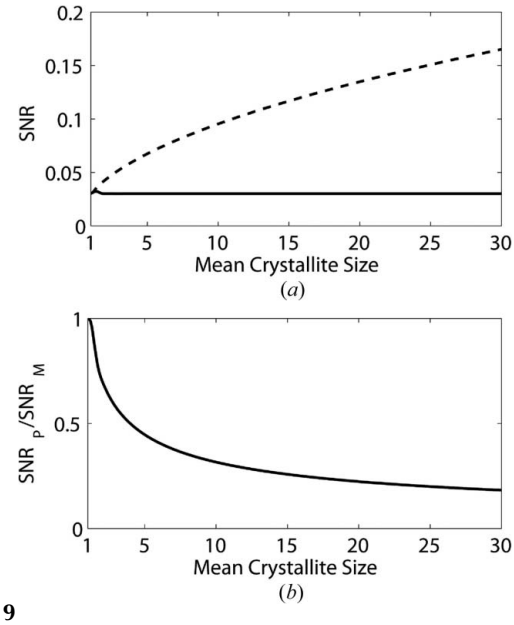

Figure 9
(*a*) Overall SNR of the diffraction pattern measured at the detector, $\mathrm{SNR_M}$ (dashed line) and the phasing SNR, $\mathrm{SNR_P}$ (solid line), and (*b*) the ratio $\mathrm{SNR_P}/\mathrm{SNR_M}$ as a function of mean crystal size.

A key factor in using this methodology to phase the diffraction data is the effect on the SNR that results from dividing by the averaged shape transform, and how this is influenced by the mean crystallite size. This effect is quantified by calculating the ratio of the SNR at the phasing stage compared to that of the data measured at the detector, *i.e.* $\mathrm{SNR_P}/\mathrm{SNR_M}$, which, using equations (49) and (50), is given by

$$\frac{\mathrm{SNR_P}}{\mathrm{SNR_M}} = \frac{1}{p}\left(\sum_{i=1}^{p} Q_i^2 \sum_{i=1}^{p} \frac{1}{Q_i^2}\right)^{-1/2}. \tag{52}$$

This SNR ratio was calculated as a function of the mean crystal size with the same parameters as above and is shown in Fig. 9(*b*). The ratio is unity for crystals with a single unit cell (single molecule) and decreases as the mean crystallite size increases. This decrease in $\mathrm{SNR_P}/\mathrm{SNR_M}$ with increasing mean crystallite size gives a quantitative result that may assist in the design of experiments. Note that $\mathrm{SNR_P}$ increases with increasing incident X-ray pulse flux and with an increasing number of patterns m. The results presented in this section extend straightforwardly to the three-dimensional case and the signal-to-noise ratios show the same kind of general

behavior as a function of mean crystallite size, although it will be different in detail, depending on the kinds of crystal shapes present.

6. Summary

Diffraction data recorded from a stream of nanocrystals in an XFEL can be processed to provide an estimate of the continuous molecular transform, thereby offering the possibility of direct phase retrieval and structure determination in the absence of other experimental information. A key difficulty with this proposal is the resulting noise level in the estimate of the molecular transform. We have described in detail the diffraction from a collection of nanocrystals of different sizes and shown that the frequency of occurrence of different crystal shapes (as measured by the correlation between edge lengths) has only a weak effect on the resulting averaged shape transform. We also show that averaging over the diffracted intensity around each Bragg reflection rigorously converges to an estimate of the averaged shape transform. Signal-to-noise properties of the intensity data and the derived molecular transform have been characterized. Noise in the measured diffraction is amplified between the Bragg reflections when the molecular transform for phasing is estimated from the data. Recorded diffraction patterns are photon limited, but after averaging together multiple patterns, the noise is normally distributed with a variance proportional to the measured intensity. Although amplification of the noise in the estimated molecular transform increases with larger crystal sizes, the overall SNR is approximately independent of the mean crystal size due to the increased signal levels from the larger crystals. The decrease in SNR of the phasing amplitude over that of the measured data as a function of mean crystallite size shown in Fig. 9(*b*) may be useful in designing experiments.

APPENDIX *A*
Two descriptions of diffraction by a finite crystal

An alternative expression to equation (3) for diffraction by a finite crystal can be obtained by considering the finite crystal as a truncated infinite crystal. This description is derived here and comparisons are made with equations (3) and (4).

The electron density of a one-dimensional crystal of N unit cells can be written as

$$g(N,x)=\begin{cases} g_\infty(x)\,\mathrm{rect}(x/Na) & N \text{ odd} \\ g_\infty(x-a/2)\,\mathrm{rect}(x/Na) & N \text{ even}, \end{cases} \tag{53}$$

where $g_\infty(x)$ denotes the electron density of an infinite crystal with the central unit cell centered on the origin as in Fig. 2(*a*), and $\mathrm{rect}(x)=1$ for $|x|\leq 1/2$ and zero elsewhere. Taking the Fourier transform of equation (53) shows that the complex diffracted amplitude for a finite crystal of size N is

$$G(N,u)=\begin{cases} G_\infty(u)\otimes[Na\,\mathrm{sinc}(Nau)] & N \text{ odd} \\ [G_\infty(u)\exp(-i\pi au)]\otimes[Na\,\mathrm{sinc}(Nau)] & N \text{ even}, \end{cases} \tag{54}$$

where $G_\infty(u)$ is the complex diffracted amplitude for an infinite crystal and $\mathrm{sinc}(u)=\sin(\pi u)/(\pi u)$. Since $G_\infty(u)$ is the Bragg diffraction pattern, we have that

$$G_\infty(u)=\frac{1}{a}F(u)\sum_{h=-\infty}^{\infty}\delta\left(u-\frac{h}{a}\right). \tag{55}$$

Substituting equation (55) into equation (54) and combining the N odd and even cases gives

$$G(N,u)=N\sum_{h=-\infty}^{\infty}(-1)^{h(N-1)}F\left(\frac{h}{a}\right)\mathrm{sinc}\left(Na\left(u-\frac{h}{a}\right)\right). \tag{56}$$

Equation (56) is therefore an alternative expression to equation (3) for $G(N,u)$ and shows that the diffraction profile around each Bragg reflection is a weighted sum of sinc functions. The equivalence between the right-hand sides of equations (3) and (56) is not obvious, but since their left-hand sides are the same, the following mathematical identity must be true for all $F(u)$:

$$N\sum_{h=-\infty}^{\infty}(-1)^{h(N-1)}F\left(\frac{h}{a}\right)\mathrm{sinc}\left(Na\left(u-\frac{h}{a}\right)\right)\equiv F(u)\frac{\sin(\pi Nau)}{\sin(\pi au)}. \tag{57}$$

The particular case $F(u)=1$ leads to the following identity:

$$N\sum_{h=-\infty}^{\infty}(-1)^{h(N-1)}\mathrm{sinc}\left(Na\left(u-\frac{h}{a}\right)\right)=\frac{\sin(\pi Nau)}{\sin(\pi au)}. \tag{58}$$

The identity [equation (58)] can be shown directly as follows. Denoting the left-hand side of equation (58) by $H(u)$, substituting for the sinc function and considering the cases for N odd and even shows that

$$H(u)=-\frac{1}{\pi}\sin(\pi Nau)\left[\frac{1}{au}+2au\sum_{h=0}^{\infty}(-1)^h\frac{1}{h^2-(au)^2}\right]. \tag{59}$$

Using the identity given by equation (4.103) of Wheelon (1968),

$$\sum_{h=0}^{\infty}(-1)^h\frac{1}{h^2-a^2}=-\frac{\pi}{2a}\left[\mathrm{cosec}(\pi a)+\frac{1}{\pi a}\right], \tag{60}$$

shows that equation (59) reduces to

$$H(u)=\frac{\sin(\pi Nau)}{\sin(\pi au)}, \tag{61}$$

verifying equation (58).

Although the reflection profiles implied by equation (56) are a sum of weighted sinc functions, if the crystallites are sufficiently large, then there is minimal overlap between them, and the profile in the vicinity of each Bragg reflection reduces simply to $\mathrm{sinc}(Nau)$. The effect of the overlap can be evaluated approximately by considering the relative contribution from

just the two neighboring peaks, denoted here by ε. At the reciprocal-lattice point $u = h/a$,

$$\varepsilon = \frac{1}{F(h/a)}\left[F\left(\frac{h+1}{a}\right)\mathrm{sinc}\left(Na\left(\frac{h}{a}-\frac{h+1}{a}\right)\right) + F\left(\frac{h-1}{a}\right)\mathrm{sinc}\left(Na\left(\frac{h}{a}-\frac{h-1}{a}\right)\right)\right], \quad (62)$$

giving

$$\varepsilon = \mathrm{sinc}(N)\frac{1}{F(h/a)}\left[F\left(\frac{h+1}{a}\right)+F\left(\frac{h-1}{a}\right)\right]. \quad (63)$$

The relative contribution from neighboring reflections depends on their amplitudes relative to $F(h/a)$. However, if we assume that $F[(h+1)/a] \simeq F[(h-1)/a] \simeq F(h/a)$, then

$$\varepsilon \simeq 2\,\mathrm{sinc}(N). \quad (64)$$

The maximum error results when $\sin(\pi N) = \pm 1$ so that the maximum value of ε is

$$\varepsilon_{\max} \simeq \frac{2}{\pi N}. \quad (65)$$

A relative contribution of less than 5% of the value of the structure factor at $u = h/a$ requires N to be more than approximately 13 unit cells. In three dimensions the contributions come from six neighboring structure factors, resulting in $\varepsilon_{\max} \simeq 6/\pi N$ and N greater than about 40 unit cells in each direction would be required to achieve less than 5% relative error. The contribution will be larger if the neighboring reflections are stronger. This analysis shows that for nanocrystals on the order of ten unit cells across, it is necessary to use the full shape function $S(N, u)$ to describe the diffraction, rather than just a single sinc function.

This work was supported by a James Cook Research Fellowship to RPM, an NSF grant (MCB-1021557) to JCHS, NSF STC award 1231306, and a UC Doctoral Scholarship to JPJC. The authors would like to thank Phil Bones and Peter Smith for helpful discussions.

References

Als-Nielsen, J. & McMorrow, D. (2011). *Elements of Modern X-ray Physics*, 2nd ed. Chichester: John Wiley and Sons.
Barty, A. *et al.* (2012). *Nat. Photon.* **6**, 35–40.
Boutet, S. *et al.* (2012). *Science*, **337**, 362–364.
Chapman, H. N. (2009). *Nat. Mater.* **8**, 299–301.
Chapman, H. N. *et al.* (2011). *Nature (London)*, **470**, 73–77.
Chen, J. P. J. & Millane, R. P. (2013). *J. Opt. Soc. Am. A*, **12**, 2627–2634.
Chen, J. P. J., Spence, J. C. H. & Millane, R. P. (2014). *Acta Cryst.* A**70**, 154–161.
DeLucas, L. (2009). *Membrane Protein Crystallization.* Burlington, MA, USA: Academic Press.
Dilanian, R. A., Streltsov, V. A., Quiney, H. M. & Nugent, K. A. (2013). *Acta Cryst.* A**69**, 108–118.
Elser, V. (2013). *Acta Cryst.* A**69**, 559–569.
Kam, Z. (1977). *Macromolecules*, **10**, 927–934.
Karr, A. (1993). *Probability.* New York: Springer.
Kirian, R. A., Wang, X., Weierstall, U., Schmidt, K. E., Spence, J. C. H., Hunter, M., Fromme, P., White, T. A., Chapman, H. N. & Holton, J. (2010). *Opt. Express*, **18**, 5713–5723.
Kirian, R. A., White, T. A., Holton, J. M., Chapman, H. N., Fromme, P., Barty, A., Lomb, L., Aquila, A., Maia, F. R. N. C., Martin, A. V., Fromme, R., Wang, X., Hunter, M. S., Schmidt, K. E. & Spence, J. C. H. (2011). *Acta Cryst.* A**67**, 131–140.
Kundu, S. (2010). *Analog and Digital Communications.* Noida: Pearson Education.
Liu, H., Zatsepin, N. A. & Spence, J. C. H. (2014). *IUCrJ*, **1**, 19–27.
Miao, J., Charalambous, P., Kirz, J. & Sayre, D. (1999). *Nature (London)*, **400**, 342–344.
Miao, J. & Sayre, D. (2000). *Acta Cryst.* A**56**, 596–605.
Millane, R. P. (1990). *J. Opt. Soc. Am. A*, **7**, 394–411.
Neutze, R., Wouts, R., van der Spoel, D., Weckert, E. & Hajdu, J. (2000). *Nature (London)*, **406**, 752–757.
Redecke, L. *et al.* (2013). *Science*, **339**, 227–230.
Saldin, D. K., Poon, H. C., Bogan, M. J., Marchesini, S., Shapiro, D. A., Kirian, R. A., Weierstall, U. & Spence, J. C. H. (2011). *Phys. Rev. Lett.* **106**, 115501.
Sayre, D. (1952). *Acta Cryst.* **5**, 843.
Seibert, M. M. *et al.* (2011). *Nature (London)*, **470**, 78–81.
Spence, J. C. H., Kirian, R. A., Wang, X., Weierstall, U., Schmidt, K. E., White, T., Barty, A., Chapman, H. N., Marchesini, S. & Holton, J. (2011). *Opt. Express*, **19**, 2866–2873.
Spence, J. C. H., Weierstall, U. & Chapman, H. N. (2012). *Rep. Prog. Phys.* **75**, 102601.
Starodub, D. *et al.* (2012). *Nat. Commun.* **3**, 1276.
Stroud, W. J. & Millane, R. P. (1996). *Proc. R. Soc. London Ser. A*, **452**, 151–173.
Vartanyants, I. A. & Robinson, I. K. (2001). *J. Phys. Condens. Matter*, **13**, 10593–10611.
Weierstall, U., Spence, J. C. H. & Doak, R. B. (2012). *Rev. Sci. Instrum.* **83**, 035108.
Welberry, T. R. (2004). *Diffuse X-ray Scattering and Models of Disorder.* Oxford University Press.
Wheelon, A. (1968). *Tables of Summable Series and Integrals Involving Bessel Functions Part I.* Holden-Day Advanced Physics Monographs. San Francisco: Holden-Day.
White, T. A., Kirian, R. A., Martin, A. V., Aquila, A., Nass, K., Barty, A. & Chapman, H. N. (2012). *J. Appl. Cryst.* **45**, 335–341.