Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 28 April 2011
Accepted 4 February 2012

Classifying and assembling two-dimensional X-ray laser diffraction patterns of a single particle to reconstruct the three-dimensional diffraction intensity function: resolution limit due to the quantum noise

Atsushi Tokuhisa,[a] Junichiro Taka,[b] Hidetoshi Kono[b] and Nobuhiro Go[b,c]*

[a]Riken Harima Institute, 1-1-1 Kouto, Sayo-cho, Sayo-gun, Hyogo, 679-5148, Japan, [b]Quantum Beam Science Directorate, Japan Atomic Energy Agency, 8-1-7 Umemidai, Kidugawa-shi, Kyoto, 619-0215, Japan, and [c]XFEL Division, Japan Synchrotron Radiation Research Institute, 1-1-1 Kouto, Sayo-cho, Sayo-gun, Hyogo, 679-5198, Japan. Correspondence e-mail: go.nobuhiro@spring8.or.jp

A new two-step algorithm is developed for reconstructing the three-dimensional diffraction intensity of a globular biological macromolecule from many experimentally measured quantum-noise-limited two-dimensional X-ray laser diffraction patterns, each for an unknown orientation. The first step is classification of the two-dimensional patterns into groups according to the similarity of direction of the incident X-rays with respect to the molecule and an averaging within each group to reduce the noise. The second step is detection of common intersecting circles between the signal-enhanced two-dimensional patterns to identify their mutual location in the three-dimensional wavenumber space. The newly developed algorithm enables one to detect a signal for classification in noisy experimental photon-count data with as low as $\sim$0.1 photons per effective pixel. The wavenumber of such a limiting pixel determines the attainable structural resolution. From this fact, the resolution limit due to the quantum noise attainable by this new method of analysis as well as two important experimental parameters, the number of two-dimensional patterns to be measured (the load for the detector) and the number of pairs of two-dimensional patterns to be analysed (the load for the computer), are derived as a function of the incident X-ray intensity and quantities characterizing the target molecule.

1. Introduction

New, intense X-ray free-electron laser (XFEL) light sources offer a new possibility in imaging single biological macromolecules. The main problems to be solved for realization of this possibility originate from the extreme weakness of the scattered light from a single molecule. One problem is severe damage of a target caused by a single shot of intense X-ray light used to compensate for the weakness. In this respect, a lower intensity of incident X-rays is preferred. Another problem due to the weakness is the quantum noise. Algorithms for structure determination must be developed to process the experimental data immersed in the quantum noise. From this perspective, a higher intensity of incident X-rays is preferred. This paper focuses attention on this latter problem. For this purpose we make a tentative assumption that the damage process can be neglected, and will clarify the mechanism of how the quantum noise sets a limit on the resolution of structure determination. In other words, we are interested in this paper only in the lower bound of the incident X-ray intensity.

The damage problem prevents the possibility of using the same molecule repeatedly as a target. Instead we assume that a target macromolecule assumes a well defined three-dimensional structure and a new molecule from an ensemble of the same molecules is placed repeatedly at the target position, but unfortunately in an unknown random orientation. A macromolecular complex with a definite molecular composition and three-dimensional structure can also be treated. The term 'molecule' is used to mean both a biological macromolecule and its complex. Extension of the results of this paper to cases of large-scale conformational fluctuations with a magnitude larger than the resolution of structure determination will be addressed in a future paper. Because of this limitation, we specifically exclude fibrous macromolecules and assume that molecules are globular with a more-or-less spherical shape. Note also that we develop analyses in this paper under idealizing (as compared with probable experimental realizations) assumptions about the state of the target molecule such as (i) ideally random orientations and (ii) the absence of hydrating water molecules. Adaptation to these realistic problems will also be treated in future papers.

A measurable two-dimensional diffraction intensity pattern depends on an unknown molecular orientation. This missing orientational information is to be recovered computationally during an analysis of a set of many two-dimensional intensity patterns. Also missing in a two-dimensional intensity pattern is the phase information necessary for derivation of a three-dimensional molecular structure. This missing phase information is also to be recovered computationally by the so-called oversampling method (Fienup, 1982; Elser, 2003).

The methods of single-particle imaging by XFEL can be classified into two paths, depending on which of the two types of missing information is recovered first. In the first, path *A*, method, a computational procedure is applied to a set of phase-missing two-dimensional intensity patterns to find their mutual locations in the three-dimensional wavenumber space. When a sufficient number of two-dimensional patterns are properly located, a three-dimensional diffraction intensity function can be constructed, to which the oversampling method is applied to recover the missing phase information. Together with this phase information, a three-dimensional real-space structure can be derived by an inverse Fourier transformation. In the second, path *B*, method, the oversampling method is applied to each of the measured two-dimensional intensity patterns. Together with the recovered phase information, a two-dimensional real-space structure is obtained by an inverse Fourier transformation, which is approximately a projection of a three-dimensional real-space structure along an axis of the incident X-ray beam. Such two-dimensional structures of a minivirus particle as revealed by a single-shot 6.9 Å hard-X-ray free-electron laser have been recently reported (Seibert *et al.*, 2011). From many projected two-dimensional images thus obtained, a three-dimensional real-space structure can be constructed by applying the method of tomography. Such a three-dimensional human chromosome structure as revealed by coherent 2.5 Å X-rays from synchrotron radiation has been reported (Nishino *et al.*, 2009).

Because of the weakness of scattered light from a single molecule, the quantum noise is a serious problem especially at high-angle pixels. The quantum noise appears to limit the resolution of a three-dimensional real-space structure to be obtained at the end, though by different mechanisms in paths *A* and *B*. In path *A*, the quantum noise is expected to set a resolution limit to locating a two-dimensional intensity pattern in the three-dimensional wavenumber space. Because photon-count data at many pixels can be considered integrally in a computational procedure to find a location, effective information seems extractable even from high-noise data at pixels with an expected mean photon count smaller than unity. Even though data at high-angle pixels are very noisy in each of the two-dimensional intensity patterns thus located, data at similar locations can be averaged to reduce the noise. The attainable structural resolution is determined by the wavenumber of a limiting pixel, from the data of which effective information can be extracted. In path *B*, where the oversampling method is applied directly to each two-dimensional intensity pattern, the quantum noise is expected to set a limit on the applicability of this method. When high-noise data from high-angle pixels with an expected photon count smaller than unity are included, the phase recovery procedure is expected to fail to converge and to cease to work. Therefore a pixel with an expected photon count of unity appears to be a limiting pixel, and its wavenumber appears to give the resolution. From such an analysis we expect that the path *A* analysis is better in extracting effective information from noisy data and therefore in deriving higher-resolution real-space structures. For this reason we are interested in this paper in developing a path *A* method. Of course, superior situations of the path *B* method are conceivable depending on problems of developing detecting devices and sample preparation, and also on the biological significance of the results obtained.

Methods hitherto proposed to find the locations of individual two-dimensional intensity patterns in the three-dimensional wavenumber space in path *A* methodology can be classified into two groups. In group 1 (Huldt *et al.*, 2003; Bortel & Faigel, 2007; Shneerson *et al.*, 2008; Bortel *et al.*, 2009; Yang *et al.*, 2010), a method of finding similarity between an arbitrary pair of two-dimensional intensity patterns is prepared. Then, a set of two-dimensional patterns are classified into groups of similar patterns according to this similarity, which are then averaged to reduce the quantum noise. Then, for an arbitrary pair of noise-reduced intensity patterns, their mutual location in the three-dimensional wavenumber space is identified by finding an intersecting circle between them. A three-dimensional diffraction intensity function can be constructed when a sufficient number of two-dimensional patterns are properly located in the three-dimensional wavenumber space. In the methods of group 2 (Fung *et al.*, 2009; Loh & Elser, 2009; Elser, 2009; Loh *et al.*, 2010), a tentative three-dimensional diffraction intensity function (or a function of similar mathematical setting) is assumed. Then, each two-dimensional intensity pattern is located so as to best fit in this three-dimensional intensity function. From a set of two-dimensional patterns thus located, the three-dimensional diffraction intensity function is updated. By repeating this cycle of best fitting and updating, an ultimate three-dimensional diffraction intensity function is obtained. Even though the methods of the second group appear promising, the demonstrated abilities of the individual methods proposed so far are limited.

We focus attention in this paper on developing an algorithm beyond existing methods of single-particle imaging belonging to the path *A*, group 1 methodology. New developments have been attained in two aspects. First, we will develop and improve methods of computational analyses and procedures to arrange a set of many experimentally measurable two-dimensional intensity patterns in the three-dimensional wavenumber space so that a three-dimensional intensity function can be constructed. The newly developed method enables us to attain higher-resolution structures. Second, we will derive explicit theoretical expressions for two main parameters which govern the number $N_{\rm measure}$ of two-dimensional patterns to be measured (the load for the measuring machine) and the number $N_{\rm compare}$ of pairs of two-dimensional patterns to be compared (the main computational load for analyses), as well as for the space resolution attainable from the analysis of the data, in terms of (*a*) the X-ray intensity used for the measurement and two types of quantities characterizing a target; (*b*) the Shannon molecular length (or, simply, molecular length) L (the length of a side of the smallest cubic box that can contain a target globular molecule); and (*c*) the radial diffraction intensity density function $\bar{i}(k)$ (the average of the the squared modulus of the structure-factor function on a sphere $k = |\mathbf{k}|$ in the wavenumber space). The achievement of the second aspect provides basic information for designing new experiments and experimental instruments.

The results in the two aspects are obtained in this paper by taking advantage of simulated diffraction intensity data for a protein, lysozyme, and a protein complex, HslUV complex, for which structural atomic coordinates are available from the Protein Data Bank (PDB) (Berman *et al.*, 2000). The former (Weaver & Matthews, 1987) (PDB code 2lzm, number of residues 164, molecular length 60 Å) is chosen as a typical small globular protein. The latter (Sousa *et al.*, 2002) (PDB code 1kyi, total number of residues 7416, molecular length 200 Å) is chosen from very large protein complexes in the PDB with more-or-less globular shape. But it is in a sense an atypical complex, because it has a big hollow space inside the structure. Simulations are carried out by assuming the wavelength of incident X-rays $\lambda = 1$ Å and for various intensities. The Shannon molecular length L and radial diffraction intensity density function $\bar{i}(k)$ are determined for these two targets from their respective PDB atomic coordinates, and are used to estimate numerical values of the two main experimental parameters, $N_{\rm measure}$ and $N_{\rm compare}$, and also of the attainable resolution as functions of the incident X-ray intensity.

To make the results of this paper more useful for designing new experiments and experimental instruments, the theory must be extended so as to be applicable even for structure-unknown molecules. This objective is studied in a separate paper.

This paper is ordered as follows. In §2, we will discuss a method of finding similarity between an arbitrary pair of two-dimensional diffraction intensity patterns. This method is used to classify two-dimensional patterns into groups of similar patterns. In §3, we will treat the problem of finding relative orientations between groups of similar patterns, which is information to be used for assembling two-dimensional intensity patterns into a three-dimensional diffraction intensity density function. In §4, we will give a somewhat detailed summary. Readers may find it easier to comprehend §§2 and 3 by reading §4 in parallel. Appendices *A*, *B* and *C* describe mathematical derivations of relations used in §§2 and 3.

2. Classification of two-dimensional diffraction intensity patterns

2.1. Two-dimensional diffraction pattern on an Ewald sphere

Here we define notations of pertinent quantities. The experimentally observable diffraction intensity $s(\mathbf{k})$, given in the unit of a number of photons arriving at a pixel of the detector of solid angle ω, is given, except for a phase factor, by

$$s(\mathbf{k}) = C\omega i(\mathbf{k}), \quad C = I_i r_{\rm CE}^2, \quad i(\mathbf{k}) = \left|F(\mathbf{k})\right|^2, \qquad (1)$$

where I_i is the incident X-ray intensity (given, in the following, in the unit of a number of photons per pulse of free-electron laser per mm^2), $r_{\rm CE}$ is the classical electron radius, C is a coefficient given by these quantities, $F(\mathbf{k})$ is the structure factor, $\mathbf{k}$ is the momentum transfer and $i(\mathbf{k})$ is the diffraction intensity density. The magnitude of the momentum transfer is given by

$$k = \frac{2}{\lambda}\sin\frac{\xi}{2}, \qquad (2)$$

where λ is the X-ray wavelength and ξ is the angle of diffraction. Note that, even though this angle is expressed as 2θ in the usual literature, we nevertheless express it as ξ because this quantity, having a meaning as part of a certain polar angle, has an important role in this paper. Even though the structure factor $F(\mathbf{k})$ is a continuous function in the wavenumber space, its squared modulus is measured experimentally by a detector with an array of finite-sized pixels. When $\left|F(\mathbf{k})\right|^2$ is discretely sampled at lattice points of a cubic lattice with a lattice constant of $1/R$, it corresponds to the use of the detector pixel size of $(\lambda/R)^2$ in solid angle, *i.e.*

$$\omega = (\lambda/R)^2 = (\lambda/\sigma L)^2. \qquad (3)$$

In this expression, L is the length of a side of the smallest cubic box (Shannon box) that can contain a target globular molecule. We shall refer to this quantity as the Shannon molecular length or simply the molecular length. From the point of view of the oversampling method for phase retrieval (Fienup, 1982; Elser, 2003), the detector pixel size must be chosen so that $R \geq L$. The ratio $\sigma = R/L$ is called the linear oversampling ratio. A pixel with $\sigma = 1$ will be referred to as a Shannon pixel.

The quantity of equation (1) is an expected number of photons arriving at a detector pixel. However, in real experiments, what is measured is an integral number of photons, given by the quantum-mechanical probability. In this paper we simulate this probability by replacing the function $s(\mathbf{k})$ of equation (1) with a stochastic function $s_Q(\mathbf{k})$ which assumes only integral values according to the Poisson distribution. To distinguish these two functions, let us call the quantity of equation (1) the theoretical diffraction intensity, and the replaced stochastic function the experimental diffraction intensity.

Examples of a simulated two-dimensional experimental and a theoretical diffraction intensity pattern are shown in Figs. 1(*a*) and 1(*b*), respectively, for the case of lysozyme. We see that, when theoretical diffraction intensities are much less than unity, experimental diffraction intensity values at most pixels vanish. From such noisy data, we have to guess the correct mean values. These are patterns for which we develop a method of analysis for structure determination. For this purpose we need some theoretical tools. A mathematical expression of a two-dimensional pattern for a molecule with a given orientation is explored in detail in Appendix *A*. A molecular orientation is described by a Eulerian angle (α, β, γ) with its corresponding 3×3 orthogonal matrix $\mathbf{A}$ given by equation (27). The Eulerian angle is defined so that, out of a set of three angles, $(-\beta, -\gamma)$ are polar angles of the direction of the incident X-ray beam with respect to the molecule, and α is an angle of rotation of the detector plane around the axis of the incident beam. A two-dimensional pattern is given by the quantity of equation (1) on the surface of an Ewald sphere. By introducing a polar coordinate (ξ, φ) on the surface of an Ewald sphere, the two-dimensional pattern is given from equations (40), (39), (35) and (42) by

$$\begin{aligned} s(\alpha, \beta, \gamma; \xi, \varphi) = s_{\mathbf{A}}(\xi, \varphi) &= C\omega \big| F[\mathbf{A}\mathbf{h}(\xi, \varphi)] \big|^2 \\ &= s(0, \beta, \gamma; \xi, \varphi - \alpha). \end{aligned} \tag{4}$$

This equation means that the Ewald spheres corresponding to (α, β, γ) and $(0, \beta, \gamma)$ are essentially the same spheres giving the same surface.

2.2. High correlation line in a correlation pattern

As outlined in §1 we adopt in this paper the basic strategy in which, after measurement of a large number of two-dimensional patterns for a molecule in unknown random orientations, we classify them into groups of similar patterns. In the following we consider a pair of Eulerian angles $(\alpha_i, \beta_i, \gamma_i)$ and $(\alpha_j, \beta_j, \gamma_j)$ with corresponding matrices $\mathbf{A}(i)$ and $\mathbf{A}(j)$, and two-dimensional patterns $s(i; \xi, \varphi)$ and $s(j; \xi, \varphi)$. For this pair we will be interested in an angle β_{ij} between the two beam directions $(-\beta_i, -\gamma_i)$ and $(-\beta_j, -\gamma_j)$, which satisfies

$$\cos\beta_{ij} = \sin\beta_i \sin\beta_j \cos(\gamma_i - \gamma_j) + \cos\beta_i \cos\beta_j. \tag{5}$$

This angle plays the role of a measure of similarity between a pair of two-dimensional patterns. When it is very small, we classify them into one group of similar patterns even for very different α_i and α_j.

Our problem is to judge, for a given pair of experimental two-dimensional patterns such as those shown in Fig. 1(*a*), whether or not they are realizations of a similar theoretical two-dimensional pattern. The starting point is the calculation of a correlation function as was originally proposed by Huldt *et al.* (2003). In this treatment we are interested in pixels on a circle with a fixed value of ξ. Let the number of pixels on a circle be N_ξ. Bortel & Faigel (2007) proposed to pre-normalize the data on a circle to have uniform second moments for the calculation of a correlation function. More recently, they (Bortel *et al.*, 2009) proposed further to pre-normalize so as to have vanishing mean and uniform second moment and also to compare axially rotated patterns. In our method we normalize the data on a circle to have a uniform mean and are interested in the correlation of their deviation from the mean after they are mutually rotated by an angle α. This means that we are concerned with the following correlation function:

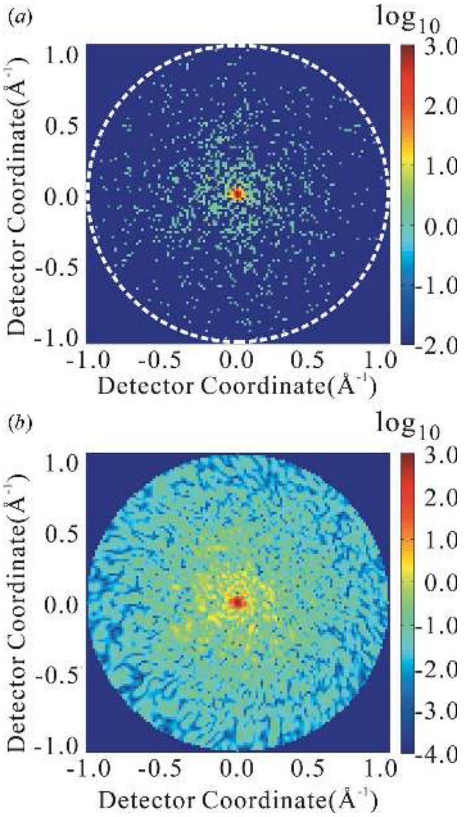

Figure 1
Examples of two-dimensional diffraction intensity patterns simulated for lysozyme by assuming the wavelength of the incident X-rays λ = 1 Å and the intensity $I_i = 5 \times 10^{21}$ photons pulse^{-1} mm^{-2}. In (*b*), the theoretically expected mean number of diffracted photons arriving at a Shannon pixel is shown. In (*a*), the number of photons is an integer, chosen according to the Poisson distribution having the theoretically expected mean value as its mean. For this assumed intensity, the mean count of 0.1 photon is observed at $1/k \cong 2$ Å. In this figure both the abscissa and ordinate show detector coordinates in the sense that they are proportional to coordinates on a flat detector on which the surface of an Ewald sphere is radially projected from its centre (central azimuthal projection).

$$\begin{aligned} c_{ij}(\xi, \alpha) &= \frac{1}{N_\xi} \sum_{l=0}^{N_\xi - 1} \left[\tilde{s}_Q\left(i; \xi, \frac{2\pi l}{N_\xi}\right) - 1 \right] \left[\tilde{s}_Q\left(j; \xi, \frac{2\pi l}{N_\xi} + \alpha\right) - 1 \right], \\ \tilde{s}_Q\left(i; \xi, \frac{2\pi l}{N_\xi}\right) &= s_Q\left(i; \xi, \frac{2\pi l}{N_\xi}\right) \Big/ \overline{s}_Q(i; \xi), \\ \overline{s}_Q(i; \xi) &= \frac{1}{N_\xi} \sum_{l=0}^{N_\xi - 1} s_Q\left(i; \xi, \frac{2\pi l}{N_\xi}\right). \end{aligned} \tag{6}$$

This quantity can also be expressed as follows:

$$\begin{aligned} c_{ij}(\xi, \alpha) &= \frac{\Psi_{ij}(\xi, \alpha)}{\overline{s}_Q(i; \xi) \overline{s}_Q(j; \xi)} - 1, \\ \Psi_{ij}(\xi, \alpha) &= \frac{1}{N_\xi} \sum_{l=0}^{N_\xi - 1} s_Q\left(i; \xi, \frac{2\pi l}{N_\xi}\right) s_Q\left(j; \xi, \frac{2\pi l}{N_\xi} + \alpha\right). \end{aligned} \tag{7}$$

The correlation function of equation (6) is calculated as a two-dimensional function of ξ and α, which will be referred to as a correlation pattern. The merits of using the normalized quantity of equation (6) are to normalize for variations of intensities over different circles and for experimental variations of the pulse intensities of XFELs.

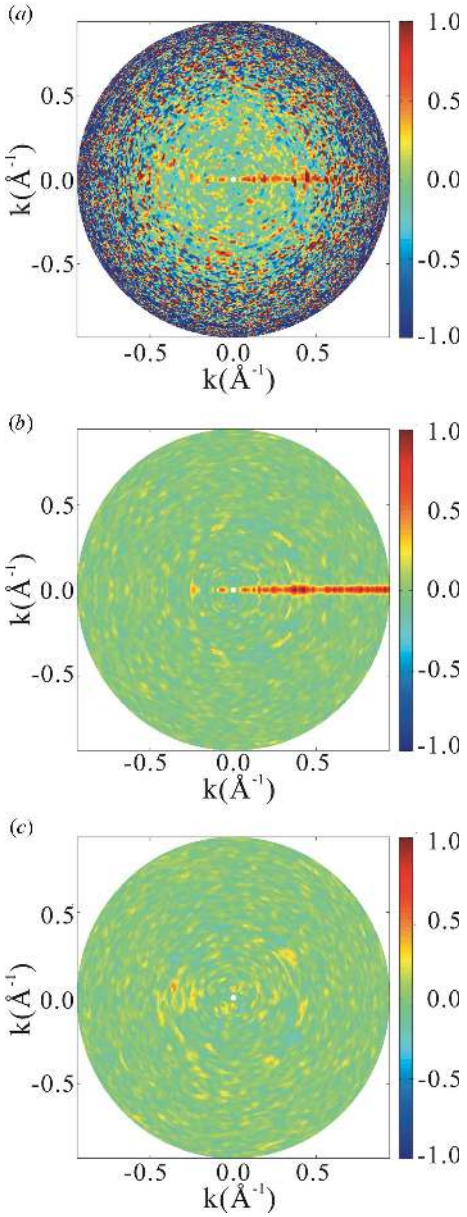

Figure 2
Correlation pattern for lysozyme by assuming the intensity of the incident X-rays $I_i = 5 \times 10^{21}$ photons pulse^{-1} mm^{-2}. (*a*) Correlation pattern of equation (6) for a pair of two-dimensional experimental intensity patterns obtained from an identical two-dimensional theoretical intensity pattern but with different sets of random numbers for the Poisson distribution. A *high correlation line* is observed extending from the origin towards the direction of $\alpha = 0$. The high correlation line fades at a certain value of k because of the quantum noise. (*b*) Correlation pattern of equation (8), which is obtained by averaging the pattern shown in (*a*) over the quantum noise. Because of the suppressed quantum noise, the high correlation line is now observed extending to high k values. (*c*) When the directions of the beam axes are significantly different in a pair of two-dimensional intensity patterns, the high correlation line is absent. In this and subsequent figures (Figs. 2, 4 and 8), unlike in Fig. 1, both the abscissa and ordinate are taken to be proportional to k. In this case, the area in the figures is also proportional to that on the curved Ewald spheres (Lambert's azimuthal equal-area projection).

An example of a correlation pattern is shown in Fig. 2(*a*). In this figure the pattern is shown not as a function of ξ and α but as a function of k and α, where k is given by equation (2). In a general case where the angle β_{ij} between the two beam directions is not very small, this correlation pattern appears to be a random function of k and α. Let us write such $c_{ij}(\xi, \alpha)$ as $c_{\rm BG}(\xi, \alpha)$, where BG stands for background. When β_{ij} is very small, there appears *a line of high correlation* for a certain value of α. In the particular case of Fig. 2(*a*), the two beam directions are identical, and therefore both β_{ij} and α vanish.

The correlation pattern of equation (6) is heavily affected by the quantum noise. When the quantum noise is suppressed, this quantity is given approximately by

$$\langle c_{ij}(\xi, \alpha)\rangle \cong \frac{\langle \Psi_{ij}(\xi, \alpha)\rangle}{\langle \bar{s}_Q(i; \xi)\rangle\langle \bar{s}_Q(j; \xi)\rangle} - 1, \qquad (8)$$

where $\langle \ldots \rangle$ is a mean averaged over the quantum noise. To derive this expression we introduced an approximation of taking the means of three factors independently, an approximation which is good when standard deviations of the three factors are significantly smaller than their respective means. Examples of this quantity are shown in Figs. 2(*b*) and 2(*c*). Fig. 2(*c*) shows a pair of diffraction patterns for which the value of β_{ij} is not small. This is an example of $\langle c_{\rm BG}(\xi, \alpha)\rangle$. Even though Figs. 2(*b*) and 2(*c*) contain no effect of quantum noise, the pattern other than the high correlation line appears to be rather random. Such a behaviour should be a consequence of an appearance of the diffraction intensity density function $i(\mathbf{k})$ that can be captured as a stochastic function.

To characterize the diffraction intensity density function from such a point of view, we studied first how values of $i = i(\mathbf{k})$ are distributed on a sphere of $k = |\mathbf{k}|$ around its mean $\bar{i} = \bar{i}(k)$. For this purpose, $\sim 1.5 \times 10^5$ points are sampled randomly with a uniform probability on each sphere, and the value of $i(\mathbf{k})$ is calculated at each sampled point. We confirmed that, except for small values of k, the distribution is given to a very good accuracy by the exponential distribution as was originally discovered by Wilson (1949). Moreover, as shown in Fig. 3, it is found empirically that, except for small values of k, values of $i(\mathbf{k})$ on the sphere of $k = |\mathbf{k}|$ are correlated in such a way as to satisfy the following simple relation,

$$c_N(\delta) \equiv \frac{\langle i(\mathbf{k}_1) i(\mathbf{k}_2)\rangle_\delta - \langle i\rangle^2}{\langle i^2\rangle - \langle i\rangle^2} = \exp\left[-\left(\frac{k\delta}{k_C}\right)^2\right], \qquad (9)$$

where δ is an angle between the two $\mathbf{k}$ vectors, the average is taken over all pairs of vectors with a given value δ, and the correlation length $k_{\rm C}$ has been found to be independent of the value of k and is given approximately by

$$k_{\rm C} = \frac{1}{L}. \qquad (10)$$

Because, as shown by Wilson (1949), the exponential distribution is a consequence of the irregular three-dimensional structures of biopolymers at the atomic level, we shall refer to the empirically observed distribution as a *structure irregularity distribution*.

To derive the mean behaviour of $c_{\rm BG}(\xi,\alpha)$, we average $\langle c_{\rm BG}(\xi,\alpha)\rangle$ over the structure irregularity distribution. As shown in Appendix *C*, we see that $\langle\langle c_{\rm BG}(\xi,\alpha)\rangle\rangle$ vanishes. This is reasonable because when there is no correlation between two points on two circles appearing in equation (6) an average can be taken on each of $(\tilde{s}_Q-1)$ to yield a vanishing result.

Further examples of a correlation pattern of equation (6) are shown in Fig. 4. Because the mean of the correlation pattern vanishes except for a high correlation line, a mathematical expression of the high correlation line should be obtained as a mean of the correlation pattern over the two distributions, the Poisson distribution and the structure irregularity distribution. As shown in Appendix *C* it is given to a good approximation by the following expression:

$$\langle\langle c_{ij}(\xi,\alpha)\rangle\rangle \cong \exp\left\{-\left(\frac{k}{k_C}\right)^2\left[\left(\alpha-\hat{\alpha}_{ij}\right)^2+\frac{\beta_{ij}^2}{2}\right]\right\} \qquad (11)$$

where k is a function of ξ through equation (2) and the direction of the high correlation line, $\hat{\alpha}_{ij}$, is given to the zeroth order of the small quantity β_{ij} by

$$\hat{\alpha}_{ij}=\left(\alpha_j-\alpha_i\right)+o\left(\beta_{ij}\right). \qquad (12)$$

It should be noted that k appears in equation (11) as a quantity normalized by $k_{\rm C}$, or, because of equation (10), as a product kL.

As mentioned earlier, Bortel *et al.* (2009) proposed to use a quantity pre-normalized to have a vanishing mean and uniform second moment. However, when we apply our analysis to their proposed correlation pattern, an expression similar to equation (11) is obtained but with an additional factor which is approximately $C\omega\bar{i}(k)/[1+C\omega\bar{i}(k)]$ and becomes smaller for larger k. Because of this additional factor, the quantity they employed for detecting similarity between two-dimensional patterns is less sensitive for high k values. This explains why our method is more sensitive for higher k values.

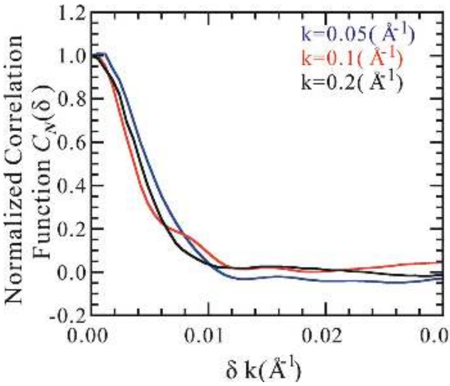

Figure 3
Normalized correlation function $c_N(\delta)$ of equation (9) for the space correlation of the values of $i(\mathbf{k})$ on a sphere $|\mathbf{k}|=k$ of radius k for the HslUV complex. The angle δ is shown in the abscissa as a product with k. For k equal to or larger than 0.2 Å^{-1}, it is given to a very high accuracy by a Gaussian function. Data only up to the case of 0.2 Å^{-1} are shown in the figure. As k becomes smaller, slight deviations from the Gaussian behaviour are observed.

2.3. Identifying the high correlation line against the noisy background and attainable resolution

In Fig. 4 we see that, even though the mean value of the correlation pattern should vanish in the background, its actual values become very noisy for larger values of *k*. This is due to both the quantum noise and the structure irregularity distribution. Identification of the high correlation line would be

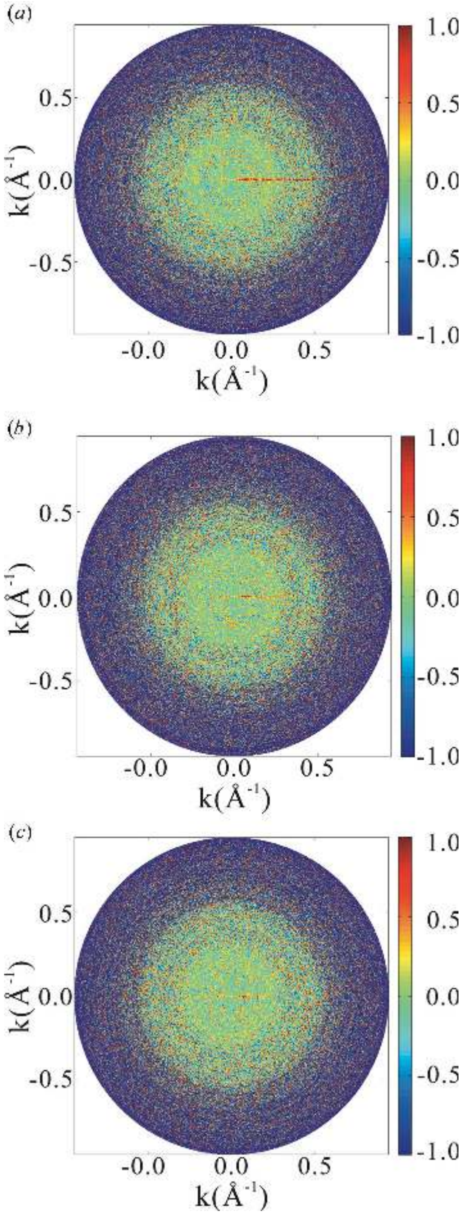

Figure 4
Correlation patterns for the HslUV complex and by assuming the intensity of incident X-rays $I_i=5\times10^{20}$ photons pulse^{-1} mm^{-2}. (*a*), (*b*) and (*c*) are for the angle β_{ij} between the two beam directions being 0, 1 and 3°, respectively. Pixels having a value larger than 1.0 are shown by the colour code of 1.0. We see in this figure that, for β_{ij} = 3°, the high correlation line is barely visible.

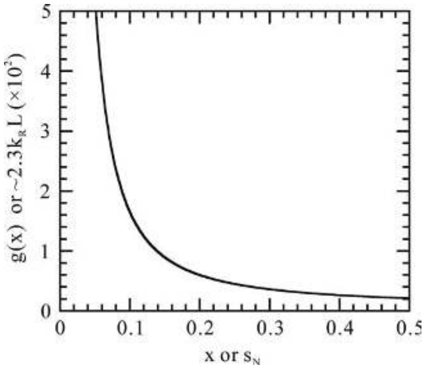

Figure 5
Graph of the function of equation (14). The abscissa x and ordinate y mean physically an expected number s_N of photons arriving at a Shannon pixel at $k = k_N$ and $\sim 2.3 k_R L$, respectively.

affected by the noise. (Note that we are here treating the structure irregularity distribution as a part of the noise.) The level of noise in the quantity of equation (6) can be expressed by its standard deviation. As derived in Appendix *C*, it is given by

$$\sigma_c^2 = \frac{g\big(C\omega\bar{i}\big)}{N_\xi}, \qquad (13)$$

where the function g is defined by

$$y = g(x) = \frac{5x^2 + 6x + 1}{x^2}, \qquad (14)$$

and N_ξ is given by

$$N_\xi = 2\pi\lambda^{-1}\sin\xi/k_C = 2\pi k L\left[1 - \left(\frac{k\lambda}{2}\right)^2\right]^{1/2}, \qquad (15)$$

which means that we are assuming Shannon pixels. Fig. 5 shows the graph of $y = g(x)$. Fig. 6 is a plot of the standard deviation given by equation (13), which is a globally increasing function of k in the high-k region. When averaged over the two distributions, the Poisson distribution and the structure irregularity distribution, the peak value of the high correlation line for a given value of k (or, equivalently ξ) is $\exp[-(k/k_C)^2(\beta_{ij}^2/2)]$ according to equation (11), which is a decreasing function of k. It is expected that the actual high correlation line is observable roughly up to $k = k_{\max}$, where the mean peak value becomes equal to the standard deviation σ_c, *i.e.*

$$\exp\left[-\left(\frac{k_{\max}}{k_C}\right)^2\left(\frac{\beta_{ij}^2}{2}\right)\right] = \sigma_c. \qquad (16)$$

In fact, in Figs. 2(*a*) and 4(*a*) for cases of $\beta_{ij} = 0$ and therefore where the mean peak value should stay unity, the actual high correlation lines are observed for lysozyme up to $k_{\max} \cong 0.7$ Å^{-1} and for the HslUV complex up to $k_{\max} \cong 0.55$ Å^{-1}, where $\sigma_c \cong 1.0$ according to Fig. 6. In Fig. 4(*b*) for the case of the HslUV complex with $\beta_{ij} = 1°$, the actual high correlation line is observable up to $k_{\max} \cong 0.3$ Å^{-1}, where $\sigma_c \cong 0.6$ according to Fig. 6 and equation (16) is roughly satisfied.

From equation (16) and Fig. 6, we see that we can derive the value of β_{ij} from the measured length $k_{\max}$ of the high correlation line. The longer the length $k_{\max}$, the smaller the value of β_{ij}. From the value of β_{ij}, we judge the similarity of a pair of two-dimensional patterns. When judged similar, they are classified into the same group. After the classification, two-dimensional patterns classified into the same group are averaged in order to improve the signal-to-noise ratio. Because this averaging is done for patterns with slightly different directions of the incident beam, the resolution of the resulting three-dimensional structure will be affected. In order to attain the highest possible resolution, we should adopt a strategy in which we classify a pair of two-dimensional patterns into the same group when their high correlation line reaches the highest possible k region. Let us define k_N (subscript N for noise) as the lower bound of such a region. This quantity k_N, *the limiting k value* for correlation recognition, plays a central role in the method of single-particle imaging developed in this paper. In the case of Fig. 4(*a*) we judge that the high correlation line extends up to such a region, where the line can no longer be distinguished from the background. In the case of Fig. 4(*b*) the line appears to have faded away before reaching such a region. The limiting value k_N, to be determined purely operationally in real applications, appears more-or-less well defined. However, we need to interpret the value of k_N in a more theoretical setting. Because it defines the lower bound of the noisy region, it should be characterized by its value of σ_c. From Figs. 4(*a*) and 4(*b*) we see that it should be between 0.6 and 1.0. As a modest estimate, we assume that k_N corresponds to the value of k at which $\sigma_c = 0.6 \cong \exp(-1/2)$. Then, from equation (16) we see that the corresponding value of β_{ij} is estimated to be within

$$\delta_G \equiv k_C/k_N. \qquad (17)$$

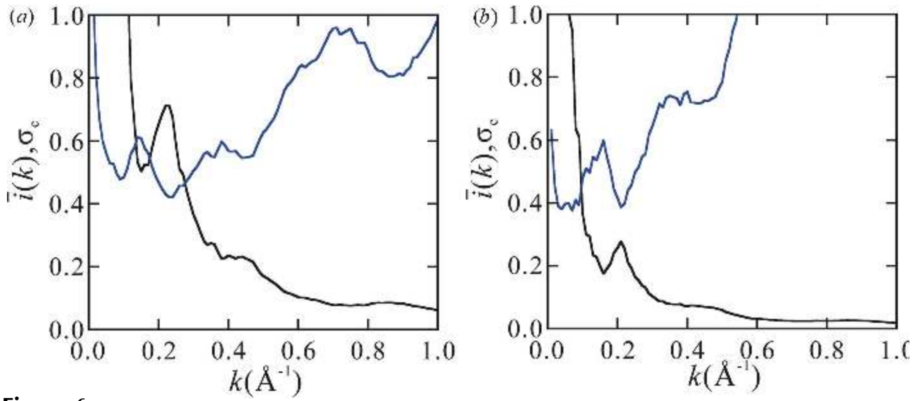

Figure 6
Plot of $\bar{i}(k)$, the average of $i(\mathbf{k})$ of equation (1) on the sphere of $|\mathbf{k}| = k$ (black line), and σ_c of equation (13) (blue line) for lysozyme (*a*) and the HslUV complex (*b*). The intensities assumed are $I_i = 5 \times 10^{21}$ and 5×10^{20} photons pulse^{-1} mm^{-2}, respectively.

Let us now assume that a classification group of similar two-dimensional patterns is constructed by a group of two-dimensional patterns with β_{ij} within this angle from a certain reference two-dimensional pattern. Note that the average distance (root-mean-square distance) between a pair of two-dimensional patterns in this classification group is also given by δ_G. During the procedure of averaging, two-dimensional patterns rotated by β_{ij} around the origin are averaged. The magnitude of displacement in k space by this rotation is given by $\beta_{ij}k$, with its maximum value being $\delta_G k$. When this magnitude is smaller than the correlation length k_C of the diffraction intensity density, the averaging procedure works to attenuate the effect of the noise. When the product becomes larger than k_C, the averaging procedure works to destroy the information in two-dimensional patterns. This means that the structural information is contained in k only up to k_R satisfying $\delta_G k_R = k_C$. Then, from equation (17), we see

$$k_R = k_N. \tag{18}$$

A limiting photon count s_N, which is an expected number of photons arriving at *a limiting pixel*, a Shannon pixel at the limiting k value, $k = k_N$, is given by $s_N = \bar{s}(k_N) = C\omega\bar{i}(k_N)$. In equation (13), $\sigma_c^2 N_\xi$ at $k = k_N$ is given by $\sigma_c^2 N_\xi = 2\pi\sigma_c^2 k_N L[1-(k_N\lambda/2)^2]^{1/2}$. This expression can be approximated as $\sigma_c^2 N_\xi \cong 2.3k_R L$, because $1/k_N$ is in most cases at least a few times larger than λ. In Fig. 5 for a graph of $y = g(x)$, x and y can also be interpreted as s_N and $\sim 2.3k_R L$, respectively. Note that *the normalized resolution*, $k_R L$, is the number of independent structural descriptive elements along the molecular length L. For a method of single-molecule imaging to be useful, this number should be at least 20, hopefully 100. Note that this number is determined mainly by the limiting photon count s_N. Fig. 5 shows this dependence. We see that, to attain $k_R L = 20$–100, s_N should be in the range of 0.25–0.08. We have to measure and analyse such low-photon-number data. Also this number highlights a high sensitivity of the proposed method of analysis to extracting information from noisy data.

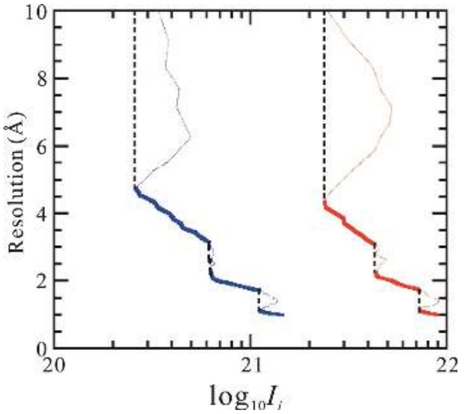

Figure 7
Attainable resolution as a function of the incident X-ray intensity I_i (photons pulse^{-1} mm^{-2}) for lysozyme (dotted and solid right-hand lines) and the HslUV complex (dotted and solid left-hand lines). When there is more than one value of resolution for a given value of I_i, the best value indicated by a solid line can be obtained.

In the above relation between the limiting photon count s_N and the normalized resolution $k_R L$, the incident X-ray intensity I_i is treated as an implicit variable parameter. To identify a particular value of I_i to attain a resolution $k_R = k_N$, we remember the relation $s_N = I_i r_{\rm CE}^2 \omega\bar{i}(k_N)$ [equation (1)]. Then, by defining a function inverse to the function g of equation (14) as $x = g^{-1}(y) = 1/[(4+y)^{1/2} - 3]$, equation (13) can be transformed to

$$I_i = \frac{g^{-1}(2.3k_R L)}{r_{\rm CE}^2 \omega\bar{i}(k_R)}. \tag{19}$$

Fig. 7 shows the resolution k_R as a function of intensity I_i for lysozyme and the HslUV complex obtained by using this equation. When there is more than one value of resolution for a given value of I_i, the best value can be obtained. (The high correlation line may become visible again in a high-k but low-noise region after once becoming invisible in a low-k but high-noise region.)

Since the solid angle of the range of one classification group is given by $\pi\delta_G^2$, and the total solid angle of the direction of the incident beam is 2π because of the centrosymmetric property of the three-dimensional diffraction intensity function, the number of classification groups N_G is given by

$$N_G = \frac{2\pi}{\pi\delta_G^2} = 2\left(\frac{k_N}{k_C}\right)^2 = 2(k_R L)^2. \tag{20}$$

3. Placing two-dimensional patterns in the three-dimensional wavenumber space: a method of finding the relative orientation between two-dimensional patterns

After two-dimensional patterns are classified by the method described in the previous section, patterns classified into the same group are averaged to reduce the noise. When signal-enhanced patterns are obtained, they are to be placed in the three-dimensional wavenumber space by finding their relative orientations. The two-dimensional patterns exist on Ewald spheres. Because all these Ewald spheres have the same radii and their surfaces contain the origin $\mathbf{k} = \mathbf{0}$ of the wavenumber space, any pair of Ewald spheres either contact at the origin or have a circular intersection, which also contains the origin. Shneerson *et al.* (2008) studied the problem of placing two-dimensional patterns in the three-dimensional space by paying attention only to an approximately straight portion of the intersecting circles near the origin $\mathbf{k} = \mathbf{0}$. Yang *et al.* (2010) refined the method of finding the tangential direction of the intersecting circle at the origin $\mathbf{k} = \mathbf{0}$ by explicitly paying attention to the curvature of the intersection. However, for the placement problem they used the method of Singer & Shkolnisky (2011) for cryo-electron microscopy in which only information on tangential directions is used. However, it is obvious that a relative orientation between a pair of Ewald spheres can be determined once their common circle is identified. In this section we first develop a method of identifying

Table 1
Resolution and necessary number of patterns and classification calculations expected for two sample ‘molecules’.

	Lysozyme		HslUV complex	
Correlation length of intensity data $1/k_C = L$	60 Å		200 Å	
Assumed intensity of incident beam I_i (photons pulse^{-1} mm^{-2})	5×10^{21}	10^{22}	5×10^{20}	10^{21}
Noise level becomes high at k_N (Å^{-1})	0.48	0.99	0.28	0.55
Photon number s_N at k_N	0.19	0.12	0.13	0.08
Resolution $1/k_R = 1/k_N$ (Å)	2.08	1.01	3.57	1.82
Range of classification group $\delta_G = k_C/k_N$ (°)	1.99	0.96	1.02	0.52
Number of classification group $N_G = 2/\delta_G^2$	1.7×10^3	7.1×10^3	6.3×10^3	2.5×10^4
Number of two-dimensional patterns to be averaged $N_A = 8/s_N$	41	63	61	100
Necessary number of two-dimensional patterns $N_{\rm measure} = N_A N_G$	6.7×10^4	4.5×10^5	3.8×10^5	2.4×10^6
Number of classification calculations $N_{\rm compare} = N_A N_G^2$	1.1×10^8	3.2×10^9	2.4×10^9	5.8×10^{10}

an intersecting circle for a given pair of signal-enhanced two-dimensional patterns $s_1(\xi, \varphi)$ and $s_2(\xi, \varphi)$, and derive a mathematical expression for the relative orientation. Second, we ask what is the necessary number of patterns to be averaged for possible identification of common circles? The actual construction of a single three-dimensional diffraction intensity function from the data of relative orientations will be treated in a different paper.

We assume that a pair of signal-enhanced two-dimensional patterns $s_1(\xi, \varphi)$ and $s_2(\xi, \varphi)$ exist on Ewald spheres of as yet unknown orientations, $\mathbf{A}(1)$ and $\mathbf{A}(2)$. Because of the centrosymmetric property of the three-dimensional diffraction intensity function, any Ewald sphere $\mathbf{A}$ has its centrosymmetric image characterized by $-\mathbf{A}$. Therefore, Ewald sphere $\mathbf{A}(1)$ should have an intersecting circle with each of the Ewald spheres $\mathbf{A}(2)$ and $-\mathbf{A}(2)$. This means that for any pair of two-dimensional patterns, $s_1(\xi, \varphi)$ and $s_2(\xi, \varphi)$, there exist two common circles. A method of finding them is developed in Appendix *B* and here we describe only the result. Each of the two common circles exists as a circle of vanishing values in each of two groups of plots, (*a*) $s_1(\xi, \varphi) - s_2(\xi, \Psi - \varphi)$ and (*b*) $s_1(\xi, \varphi) - s_2(\xi, \varphi - \Psi')$, where the parameters Ψ and Ψ' generate each of the two groups, respectively. When the plot (*a*) vanishes on a circle with its centre at $\xi = \Xi, \varphi = \Phi$ for a certain parameter value Ψ, the polar coordinates of the centres of intersecting circles are (Ξ, Φ) on $s_{\mathbf{A}(1)}(\xi, \varphi)$ and $(\Xi, \Psi - \Phi)$ on $s_{\mathbf{A}(2)}(\xi, \varphi)$. When the plot (*b*) vanishes on a circle with its centre at $\xi = \Xi', \varphi = \Phi'$ for a certain parameter value Ψ', the polar coordinates of the centres of intersecting circles are (Ξ', Φ') on $s_{\mathbf{A}(1)}(\xi, \varphi)$ and $(\Xi', \Phi' - \Psi')$ on $s_{\mathbf{A}(2)}(\xi, \varphi)$. Then, the Euler angle of the relative orientation $\mathbf{A} = \mathbf{A}(1)^{-1}\mathbf{A}(2)$ is given by

$$\begin{aligned} &\alpha = \Psi - \Phi = \Phi' - \Psi' - \pi, \quad \beta = \pi - 2\Xi = 2\Xi', \\ &\gamma = \pi - \Phi = -\Phi'. \end{aligned} \tag{21}$$

Thus, the same set of Eulerian angles (α, β, γ) is now determined from each of the intersecting circles between Ewald spheres $\mathbf{A}(1)$ and $\mathbf{A}(2)$, and between Ewald spheres $\mathbf{A}(1)$ and $-\mathbf{A}(2)$. Even though the result is redundant, the actual procedures of finding vanishing circles on plots (*a*) and (*b*) are much influenced by experimental noise. In this situation, finding the same quantity simultaneously by the two methods is a desirable numerical procedure.

Fig. 8 shows an example of how circles of vanishing values become visible as the number of averaging patterns is increased. In the case of this example for the HslUV complex, in which the limiting photon count s_N is 0.13, we see by inspection that averaging over about 61 patterns is necessary for the identification. This process has been done for lysozyme and the HslUV complex both for a series of values of the incident X-ray intensity. It has been found that a product of the number N_A of necessary patterns and the limiting photon count s_N (therefore, the intensity I_i of the incident X-ray) is constant in either ‘molecule’. The analysis described in Appendix *C* indicates that the product $N_A s_N$ must be 8 or larger for common circles to be identified. This is exactly the number observed in the case of Fig. 8. Therefore, the necessary number of patterns is given by

$$N_A \cong 8/s_N. \tag{22}$$

Table 1 summarizes the results obtained as applied to the two ‘molecules’.

4. Summary and conclusion

Two aspects of a method of single-particle imaging belonging to the path *A*, group 1 methodology have been developed. First, a new, improved method has been developed for computational analyses and procedures to arrange a set of many experimentally measurable two-dimensional diffraction intensity patterns in the three-dimensional wavenumber space. Second, explicit theoretical expressions have been derived for important experimental parameters in terms of the incident X-ray intensity and two types of quantities characterizing a target.

The number $N_{\rm measure}$ of two-dimensional patterns to be measured is given by the product $N_A N_G$ of the number of classification groups N_G and the average number of two-dimensional patterns N_A to be averaged in each group for noise reduction. The number $N_{\rm compare}$ of pairs of two-dimensional patterns to be analysed is given by $N_A N_G^2$, because the detection of similarity of patterns is to be carried out for each of the pairs, one from patterns representing each group and the other from all measured patterns. We derived theoretical expressions for the two parameters, N_G and N_A.

Concerning the first aspect, we have improved a hitherto proposed method for judging whether or not an arbitrary pair of two-dimensional patterns are similar enough to belong to the same classification group. Also, we developed methods of finding common intersecting circles between an arbitrary pair of noise-reduced two-dimensional patterns, and thereby relatively locating them in the three-dimensional wavenumber space. After locating many two-dimensional diffraction patterns properly in the wavenumber space, we have to construct a single three-dimensional diffraction intensity function. This problem, as well as the problem of application of the phase retrieval procedure to such a three-dimensional function, will be treated in a different paper.

The judgment of similarity is based on a two-dimensional correlation pattern for each pair of two-dimensional diffraction intensity patterns. For the calculation of correlation patterns, a new normalization of measurable two-dimensional intensity patterns is employed, thereby enabling one to enhance the sensitivity of judgment to high-angle k values, and eventually to improve attainable space resolution.

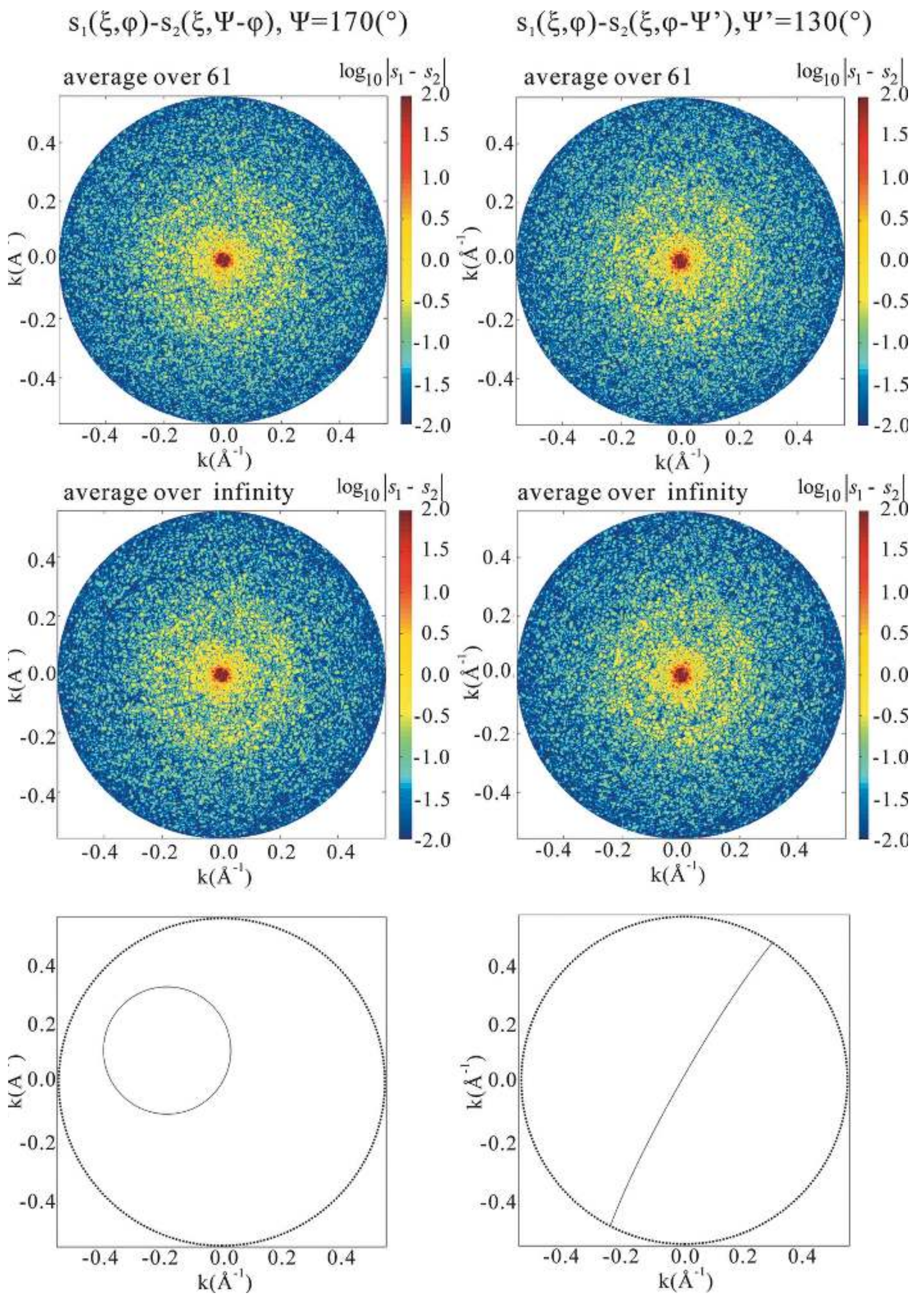

Figure 8

Plots for detecting an intersecting circle between a pair of two-dimensional patterns for the case of the HslUV complex and for the incident X-ray intensity $I_i = 5 \times 10^{20}$ photons pulse^{-1} mm^{-2}. By careful examination of the plot, we see that intersecting circles become visible when averaging of 61 or more patterns is done in this case. In the bottom panels the intersecting circles are highlighted.

A two-dimensional intensity pattern depends on the direction of the incident X-ray beam with respect to the molecule-fixed coordinate system and an angle of rotation of the detector plane placed perpendicularly to the beam axis. When an angle β_{ij} between the directions of the incident beam for a pair of two-dimensional intensity patterns is small, a high correlation line is observed in the two-dimensional correlation pattern as a straight line extending radially from the centre. The angle of the line in the correlation pattern gives a relative angle of rotation of the detector plane. The intensity of a high correlation line is unity near $k = 0$ and becomes weaker at higher angles. The intensity reduces faster for larger values of β_{ij}.

The background of correlation patterns other than the high correlation line is characterized as a pattern of random appearance reflecting the irregular three-dimensional structures of biopolymers at the atomic level superimposed with the quantum noise. Owing to the deliberately adopted, new normalization of measurable two-dimensional intensity patterns, the mean value of the distribution in the background of two-dimensional correlation patterns turns out to vanish. The standard deviation σ_c of the distribution around its vanishing mean becomes globally, but not monotonically, larger as k becomes larger. When the standard deviation σ_c becomes larger than the intensity of a high correlation line, the latter becomes no longer recognizable. The recognizable length of a high correlation line becomes longer as the value of β_{ij} becomes smaller. The latter can be determined from the former.

When the value of β_{ij} is smaller than a certain value, say, δ_G (therefore, when the recognizable length of the corresponding high correlation line becomes longer than a certain value, say, k_N), we classify the pair into the same group. To attain the best resolution, we should employ the largest possible value for k_N. Operationally we determine the value of k_N, *the limiting k value* for correlation recognition, as the lower bound of the noise-dominant k region in two-dimensional correlation patterns. Such a value k_N can be characterized theoretically as the value of k at which the standard deviation σ_c of the background distribution is 0.6. An analytic expression, equation (13), for the standard deviation is derived which is approximately a function of the wavenumber normalized by the Shannon length of the target molecule, *i.e.* kL, and an expected photon count s by a pixel at the position of the wavenumber k. The quantity k_N plays a central role in the method of analysis developed in this paper. It is shown that the structural resolution k_R attainable by this method is given by k_N.

In the method of identification of a high correlation line, upon which judgment of similarity of a pair of two-dimensional intensity patterns is based, effective information is extracted from the very noisy data in the range of wavenumbers up to the limiting value k_N where the value of the standard deviation σ_c is 0.6. From the analytic expression for σ_c, we can derive *the limiting photon count* s_N, an expected photon count at *a limiting pixel*, approximately as a function of *the normalized resolution* $k_RL = k_NL$ (Fig. 5). For a method of structure determination to be useful, the value of the normalized resolution should be in the range of 20–100. The corresponding value of the limiting photon count s_N turns out to be in the range of 0.25–0.08. The proposed method of analysis is sufficiently sensitive to enable one to extract information from such low-photon-count noisy data. This high sensitivity has been attained by employing a new correlation function. When the molecular length L and the radial diffraction intensity density function $\bar{i}(k)$ are known, the above relation between the normalized resolution k_RL and the limiting photon count $s_N = I_i r_{\rm CE}^2 \omega \bar{i}(k_N)$ can be transformed to a relation giving the intensity I_i of the incident beam to be used to attain a resolution k_R.

The angle δ_G to define a range of classification groups is given by an inverse of the normalized limiting wavenumber, $(k_NL)^{-1}$. As a result, the number of classification groups N_G is given by $2(k_RL)^2$.

A method of identifying common circles between an arbitrary pair of noise-reduced two-dimensional patterns is developed. For an arbitrary Ewald sphere, there exists a conjugate Ewald sphere which is centrosymmetric with respect to the origin of the wavenumber space. In the proposed method, when a common circle between two-dimensional patterns A and B is searched, another common circle is searched at the same time between two-dimensional patterns A and B', where B' is a two-dimensional pattern on an Ewald sphere conjugate to the one on which B exists. The average number of two-dimensional patterns N_A to be averaged in each group for identification of common circles has been shown to be given in terms of the limiting photon count s_N by $8/s_N$.

The obtained theoretical expressions are used to evaluate values of important parameters for the two sample 'molecules' by assuming, respectively, two typical intensities of the incident beam. The results are shown in Table 1. We should note the very low limiting photon counts, highlighting the strength of the method developed here. We should also note that the predicted attainable resolutions are remarkably high. This is partly due to the strength of the method of analysis developed here, but also due to the assumed high intensities of the incident beam. The assumed values of intensity in Fig. 7 and Table 1 are in the range of around 10^{21} photons pulse^{-1} mm^{-2}, which is far larger than the peak value of 1.6×10^{16} photons pulse^{-1} mm^{-2} reported in the recent experiment (Seibert *et al.*, 2011) carried out at the Linac Coherent Light Source (LCLS). Because the X-ray beam diameter reported in the experiment at LCLS is about 10 µm and a new technology (Mimura *et al.*, 2010) is now available to focus it down to 10 nm, the values assumed in this paper appear realistic. Since we developed the analysis in this paper under a tentative assumption that damage processes can be neglected, the indicated intensity is the lower bound to attain a targeted resolution. We are now carrying out a study of the damage processes to assess the upper bound of employable intensity. The number of two-dimensional patterns to be measured in Table 1 is not small, but appears tractable for real experiments. At the same time we should note that the number of classification calculations is not small.

APPENDIX *A*
Relations between molecular orientation and the Ewald sphere

We define two right-handed coordinate systems, one fixed to the molecule and the other fixed to the experimental detector. They are defined, respectively, in terms of a set of mutually orthogonal unit vectors $(\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3)$ fixed to the molecule and in terms of another set of mutually orthogonal unit vectors $(\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3)$ fixed to the detector. The position of any point in the molecule can be expressed either in terms of coordinates $\mathbf{x} = (x_1, x_2, x_3)^t$ (here the superscript t means transpose) in the molecule-fixed coordinate system (MFCS) as $\mathbf{a}_1x_1 + \mathbf{a}_2x_2 + \mathbf{a}_3x_3 = (\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3)\mathbf{x}$ or in terms of coordinates $\mathbf{y} = (y_1, y_2, y_3)^t$ in the detector-fixed coordinate system (DFCS) as $\mathbf{b}_1y_1 + \mathbf{b}_2y_2 + \mathbf{b}_3y_3 = (\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3)\mathbf{y}$. Thus

$$(\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3)\mathbf{x} = (\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3)\mathbf{y}. \qquad (23)$$

The relative orientation between the MFCS and DFCS can be described in terms of a 3 × 3 orthogonal matrix $\mathbf{A}$ as

$$(\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3) = (\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3)\mathbf{A}. \qquad (24)$$

This is a shorthand notation for

$$\mathbf{b}_i = \sum_{j=1}^{3} \mathbf{a}_j a_{ji}, \quad i = 1, 2, 3, \qquad (25)$$

where a_{ji} is an element of the matrix $\mathbf{A}$. This is an orthogonal matrix, *i.e.* $\mathbf{A}^{-1} = \mathbf{A}^t$. By introducing equation (24) into equation (23), we have the following relation between coordinates in the two coordinate systems:

$$\mathbf{x} = \mathbf{A}\mathbf{y}. \tag{26}$$

It is convenient to express the orthogonal matrix in terms of a Eulerian angle (α, β, γ) as

$$\mathbf{A} = \mathbf{A}_3(-\gamma)\mathbf{A}_2(-\beta)\mathbf{A}_3(-\alpha), \tag{27}$$

$$\mathbf{A}^{-1} = \mathbf{A}_3(\alpha)\mathbf{A}_2(\beta)\mathbf{A}_3(\gamma), \tag{28}$$

where

$$\mathbf{A}_3(\alpha) = \begin{pmatrix} \cos\alpha & -\sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{29}$$

$$\mathbf{A}_2(\beta) = \begin{pmatrix} \cos\beta & 0 & \sin\beta \\ 0 & 1 & 0 \\ -\sin\beta & 0 & \cos\beta \end{pmatrix}. \tag{30}$$

The range of variation of the Eulerian angle is taken as follows:

$$0 \le \alpha < 2\pi, \quad 0 \le \beta \le \pi, \quad 0 \le \gamma < 2\pi. \tag{31}$$

Molecular structure is described by its electron density:

$$\rho(\mathbf{x}) = \rho(x_1, x_2, x_3). \tag{32}$$

When the molecule is in the orientation described by an orthogonal matrix $\mathbf{A}$ with respect to the detector, the electron density in the DFCS is given by

$$\rho_{\mathbf{A}}(\mathbf{y}) = \rho_{\mathbf{A}}(y_1, y_2, y_3) = \rho(\mathbf{x}) = \rho(\mathbf{A}\mathbf{y}). \tag{33}$$

The structure factor is calculated in the MFCS as

$$F(\mathbf{k}) = \textstyle\int \mathrm{d}\mathbf{x}\, \rho(\mathbf{x}) \exp(-2\pi i \mathbf{k}^t \mathbf{x}). \tag{34}$$

The structure factor in the DFCS is given by

$$\begin{aligned} F_{\mathbf{A}}(\mathbf{h}) &= \textstyle\int \mathrm{d}\mathbf{y}\, \rho_{\mathbf{A}}(\mathbf{y}) \exp(-2\pi i \mathbf{h}^t \mathbf{y}) \\ &= \textstyle\int \mathrm{d}\mathbf{x}\, \rho(\mathbf{x}) \exp\big[-2\pi i (\mathbf{A}\mathbf{h})^t \mathbf{x}\big] = F(\mathbf{A}\mathbf{h}). \end{aligned} \tag{35}$$

This equation has the same structure as equation (33), indicating that the electron density and the structure factor behave in the same way for rotation. It follows from this that the wavenumber vectors $\mathbf{k}$ in the MFCS and $\mathbf{h}$ in the DFCS are related to each other similarly as in equation (26):

$$\mathbf{k} = \mathbf{A}\mathbf{h}. \tag{36}$$

Experimentally, the diffracted X-rays from a single molecule in a certain specific orientation are measured by a two-dimensional detector as a continuous diffraction pattern. This diffraction pattern is given by the squared modulus of the structure factor on the following Ewald sphere in the wavenumber space in the DFCS,

$$\left|\mathbf{h} - \lambda^{-1}\mathbf{b}_3\right|^2 = \lambda^{-2}. \tag{37}$$

Here λ is the wavelength of X-rays used in the experiment. The incident X-rays are assumed to proceed to the negative direction of the third axis $\mathbf{b}_3$ of the DFCS. The first and second axes are taken on the surface of the detector. Equation (37) describes a sphere in the $\mathbf{h}$ space. The surface of the sphere contains the origin of the wavenumber space and its centre is located at $\lambda^{-1}\mathbf{b}_3$.

We now introduce a polar coordinate (ξ, φ) on the surface of this Ewald sphere by

$$\begin{aligned} \mathbf{h}(\xi, \varphi) &= \lambda^{-1}\big[\mathbf{E} - \mathbf{A}_3(\varphi + \pi)\mathbf{A}_2(\xi)\big]\mathbf{e}_3 \\ &= \lambda^{-1}(\sin\xi\cos\varphi, \sin\xi\sin\varphi, 1 - \cos\xi)^t, \quad \mathbf{e}_3^t = (0, 0, 1). \end{aligned} \tag{38}$$

The origin $\xi = 0$, $\varphi = 0$ of the polar coordinate corresponds to the origin $\mathbf{h} = \mathbf{0}$ of the wavenumber space. This polar coordinate system is adopted so that when the detector is viewed from the direction of the incident X-rays, the detector's positive horizontal and vertical axes coincide with the directions of $\varphi = 0$ and $\varphi = \pi/2$, respectively. The observable diffraction pattern is given by equation (1) as

$$s_{\mathbf{A}}(\xi, \varphi) = C\omega\left|F_{\mathbf{A}}[\mathbf{h}(\xi, \varphi)]\right|^2. \tag{39}$$

Because the orthogonal matrix $\mathbf{A}$ can be specified by a Eulerian angle (α, β, γ), we sometimes write

$$s_{\mathbf{A}}(\xi, \varphi) = s(\alpha, \beta, \gamma; \xi, \varphi). \tag{40}$$

We also identify various Ewald spheres by their corresponding orthogonal matrix $\mathbf{A}$.

Let us now locate the Ewald sphere in the MFCS. It is given by

$$\begin{aligned} \mathbf{k} &= \mathbf{A}\mathbf{h}(\xi, \varphi) \\ &= \mathbf{A}_3(-\gamma)\mathbf{A}_2(-\beta)\mathbf{A}_3(-\alpha)\mathbf{h}(\xi, \varphi) \\ &= \mathbf{A}_3(-\gamma)\mathbf{A}_2(-\beta)\mathbf{h}(\xi, \varphi - \alpha). \end{aligned} \tag{41}$$

The last equality means

$$s(\alpha, \beta, \gamma; \xi, \varphi) = s(0, \beta, \gamma; \xi, \varphi - \alpha). \tag{42}$$

This equation means that the Ewald spheres corresponding to (α, β, γ) and $(0, \beta, \gamma)$ are essentially the same spheres giving the same surface. The former is obtained from the latter by rotating the latter anticlockwise around its third axis by angle α.

Here, the different ways in which a Eulerian angle (α, β, γ) appears in the two different coordinate systems may be worth noting. In the DFCS, a set of three angles describes the spatial orientation of the molecule. In the MFCS, the angles $(-\beta, -\gamma)$ are nothing but the polar coordinates of the incident X-ray beam axis $\mathbf{b}_3$ and the angle α describes the angle of rotation of the detector around this beam axis. Because of these clear meanings of the Eulerian angles in the MFCS, we employ them to describe molecular orientations rather than often mathematically better behaved quaternions.

Equation (42) suggests an experimental procedure of classifying various observable diffraction patterns $s(\alpha, \beta, \gamma; \xi, \varphi)$ into groups only with respect to values of β and γ. Those with the same values of β and γ but with different values of α are to be classified into the same group.

Because the electron density $\rho(\mathbf{x})$ is a real function, the structure factor satisfies the relation

$$F(\mathbf{k})^* = F(-\mathbf{k}). \tag{43}$$

Because of this relation, the same diffraction pattern is observed on a pair of different Ewald spheres. From equations (40), (39), (35) and (43) we have

$$\begin{aligned} s(\alpha,\beta,\gamma;\xi,\varphi) &= s_{\mathbf{A}}(\xi,\varphi) = C\omega\left|F_{\mathbf{A}}[\mathbf{h}(\xi,\varphi)]\right|^2 \\ &= C\omega\left|F[\mathbf{A}\mathbf{h}(\xi,\varphi)]\right|^2 = C\omega\left|F[-\mathbf{A}\mathbf{h}(\xi,\varphi)]\right|^2. \end{aligned} \tag{44}$$

The last equation indicates that two Ewald spheres, characterized by rotation matrices $\mathbf{A}$ and $-\mathbf{A}$, respectively, exist always in a pair. They occupy positions centrosymmetric to each other with respect to the origin of the wavenumber space, and they are in touch with each other at the origin. When we introduce a right-handed coordinate system to the Ewald sphere $\mathbf{A}$ with its origin at the centre of the sphere, the corresponding coordinate system of the centrosymmetric Ewald sphere $-\mathbf{A}$ is now left handed. This fact is understood as the two Ewald spheres having different handedness or parity.

APPENDIX *B*
Finding the relative orientation between a pair of unknown orientations from their corresponding diffraction intensity patterns

Real, experimentally observable diffraction patterns are subject to severe quantum noise. To cope with such noise, the same patterns should be measured many times and their means should be calculated to reduce the noise. In the following, the quantity of equation (39) is used under the assumption that such averaging has already been done so that the effect of noise can be neglected.

Let the two unknown orientations be given by

$$\mathbf{A}(i) = \mathbf{A}_3(-\gamma_i)\mathbf{A}_2(-\beta_i)\mathbf{A}_3(-\alpha_i), \quad i = 1, 2. \tag{45}$$

The corresponding Ewald spheres $\mathbf{A}(i)$ and $-\mathbf{A}(i)$ are given, respectively, by

$$\mathbf{k}(i;\xi,\varphi) = \mathbf{A}(i)\mathbf{h}(\xi,\varphi), \quad i = 1, 2 \tag{46}$$

and

$$\mathbf{k}'(i;\xi,\varphi) = -\mathbf{A}(i)\mathbf{h}(\xi,\varphi), \quad i = 1, 2. \tag{47}$$

Because all these Ewald spheres have the same radii and their surfaces contain the origin $\mathbf{k} = \mathbf{0}$ of the wavenumber space, they either contact at the origin or have a circular intersection, which also contains the origin.

Let us now study the intersecting circle between Ewald spheres $\mathbf{A}(1)$ and $\mathbf{A}(2)$. Let (ξ_1,φ_1) on the first Ewald sphere $\mathbf{A}(1)$ and (ξ_2,φ_2) on the second Ewald sphere $\mathbf{A}(2)$ be the same point on the intersecting circle,

$$\mathbf{A}(1)\mathbf{h}(\xi_1,\varphi_1) = \mathbf{A}(2)\mathbf{h}(\xi_2,\varphi_2). \tag{48}$$

Therefore, by setting

$$\mathbf{A} = \mathbf{A}(1)^{-1}\mathbf{A}(2), \tag{49}$$

we have

$$\mathbf{h}(\xi_1,\varphi_1) = \mathbf{A}\mathbf{h}(\xi_2,\varphi_2), \quad \mathbf{h}(\xi_2,\varphi_2) = \mathbf{A}^{-1}\mathbf{h}(\xi_1,\varphi_1). \tag{50}$$

These equations mean that the polar coordinates (ξ_1,φ_1) on Ewald sphere $\mathbf{A}(1)$ for points on the intersecting circle with Ewald sphere $\mathbf{A}(2)$ are also polar coordinates on Ewald sphere $\mathbf{E}$ for points on the intersecting circle with Ewald sphere $\mathbf{A}$. Similarly, the polar coordinates (ξ_2,φ_2) on Ewald sphere $\mathbf{A}(2)$ for points on the intersecting circle with Ewald sphere $\mathbf{A}(1)$ are also polar coordinates on Ewald sphere $\mathbf{E}$ for points on the intersecting circle with Ewald sphere $\mathbf{A}^{-1}$.

We now calculate the intersecting circle between Ewald spheres $\mathbf{E}$ and $\mathbf{A}$, and that between Ewald spheres $\mathbf{E}$ and $\mathbf{A}^{-1}$. We assume that $\mathbf{A}$ is given by equation (27). The centres of the three Ewald spheres, $\mathbf{O}_{\mathbf{E}}$, $\mathbf{O}_{\mathbf{A}}$ and $\mathbf{O}_{\mathbf{A}^{-1}}$, are given in the MFCS by

$$\begin{aligned} &\mathbf{O}_{\mathbf{E}} = \lambda^{-1}\mathbf{e}_3, \\ &\mathbf{O}_{\mathbf{A}} = \mathbf{A}\mathbf{O}_{\mathbf{E}} = \lambda^{-1}\mathbf{A}\mathbf{e}_3 = \lambda^{-1}(-\cos\gamma\sin\beta, \sin\gamma\sin\beta, \cos\beta)^t, \\ &\mathbf{O}_{\mathbf{A}^{-1}} = \mathbf{A}^{-1}\mathbf{O}_{\mathbf{E}} = \lambda^{-1}\mathbf{A}^{-1}\mathbf{e}_3 = \lambda^{-1}(\cos\alpha\sin\beta, \sin\alpha\sin\beta, \cos\beta)^t. \end{aligned} \tag{51}$$

Therefore

$$\left|\mathbf{O}_{\mathbf{A}} - \mathbf{O}_{\mathbf{E}}\right| = \left|\mathbf{O}_{\mathbf{A}^{-1}} - \mathbf{O}_{\mathbf{E}}\right| = 2\lambda^{-1}\sin\frac{\beta}{2}. \tag{52}$$

Let the centre of the intersecting circle between Ewald spheres $\mathbf{E}$ and $\mathbf{A}$ on Ewald sphere $\mathbf{E}$ be $\mathbf{C}_{\mathbf{A}}$. Similarly, let the centre of the intersecting circle between Ewald spheres $\mathbf{E}$ and $\mathbf{A}^{-1}$ on Ewald sphere $\mathbf{E}$ be $\mathbf{C}_{\mathbf{A}^{-1}}$. They are given by

$$\begin{aligned} \mathbf{C}_{\mathbf{A}} - \mathbf{O}_{\mathbf{E}} &= \lambda^{-1}(\mathbf{O}_{\mathbf{A}} - \mathbf{O}_{\mathbf{E}})/\left|\mathbf{O}_{\mathbf{A}} - \mathbf{O}_{\mathbf{E}}\right| \\ &= \lambda^{-1}\left(-\cos\gamma\cos\frac{\beta}{2}, \sin\gamma\cos\frac{\beta}{2}, -\sin\frac{\beta}{2}\right)^t, \\ \mathbf{C}_{\mathbf{A}^{-1}} - \mathbf{O}_{\mathbf{E}} &= \lambda^{-1}(\mathbf{O}_{\mathbf{A}^{-1}} - \mathbf{O}_{\mathbf{E}})/\left|\mathbf{O}_{\mathbf{A}^{-1}} - \mathbf{O}_{\mathbf{E}}\right| \\ &= \lambda^{-1}\left(\cos\alpha\cos\frac{\beta}{2}, \sin\alpha\cos\frac{\beta}{2}, -\sin\frac{\beta}{2}\right)^t. \end{aligned} \tag{53}$$

We now proceed to describe a procedure to find the intersecting circles for a given pair of experimental diffraction patterns $s_{\mathbf{A}(1)}(\xi,\varphi)$ and $s_{\mathbf{A}(2)}(\xi,\varphi)$. When we move clockwise along the intersecting circle on the Ewald sphere $\mathbf{A}(1)$, it appears as an anticlockwise motion along the intersecting circle on the Ewald sphere $\mathbf{A}(2)$. Therefore, we calculate the following quantity for all assumed values of Ψ in the range of $[0, 2\pi]$:

$$s_{\mathbf{A}(1)}(\xi,\varphi) - s_{\mathbf{A}(2)}(\xi, \Psi - \varphi). \tag{54}$$

This quantity should vanish on a certain circle for a certain value of Ψ. Because the intersecting circle contains the polar origin $\xi = 0$, $\varphi = 0$, it can be uniquely specified by the polar coordinate of its centre. When the quantity of equation (54) vanishes on a circle with its centre at $\xi = \Xi$, $\varphi = \Phi$, the polar coordinates of the centres of intersecting circles are (Ξ, Φ) on $s_{\mathbf{A}(1)}(\xi,\varphi)$ and $(\Xi, \Psi - \Phi)$ on $s_{\mathbf{A}(2)}(\xi,\varphi)$. These polar coordinates of the centres are to be compared with equation (53). Therefore,

$$\begin{cases} \sin\Xi\cos\Phi = -\cos\gamma\cos\dfrac{\beta}{2}, & \sin\Xi\sin\Phi = \sin\gamma\cos\dfrac{\beta}{2}, \\ \sin\Xi\cos(\Psi-\Phi) = \cos\alpha\cos\dfrac{\beta}{2}, & \sin\Xi\sin(\Psi-\Phi) = \sin\alpha\cos\dfrac{\beta}{2}. \end{cases} \tag{55}$$

From these relations we have a half of equation (21) in the main text.

Let us now proceed to study the intersecting circle between Ewald spheres $\mathbf{A}(1)$ and $-\mathbf{A}(2)$. By following similar procedures as above we have the following relations from equation (47):

$$\mathbf{h}(\xi_1,\varphi_1) = -\mathbf{A}\mathbf{h}(\xi_2,\varphi_2), \quad \mathbf{h}(\xi_2,\varphi_2) = -\mathbf{A}^{-1}\mathbf{h}(\xi_1,\varphi_1). \tag{56}$$

By introducing similar quantities and following similar procedures, we have the following results:

$$\begin{aligned} \mathbf{C}_{-\mathbf{A}} - \mathbf{O}_{\mathbf{E}} &= \lambda^{-1}(\mathbf{O}_{-\mathbf{A}} - \mathbf{O}_{\mathbf{E}})/|\mathbf{O}_{-\mathbf{A}} - \mathbf{O}_{\mathbf{E}}| \\ &= \lambda^{-1}\left(\cos\gamma\sin\frac{\beta}{2}, -\sin\gamma\sin\frac{\beta}{2}, -\cos\frac{\beta}{2}\right)^t, \\ \mathbf{C}_{-\mathbf{A}^{-1}} - \mathbf{O}_{\mathbf{E}} &= \lambda^{-1}(\mathbf{O}_{-\mathbf{A}^{-1}} - \mathbf{O}_{\mathbf{E}})/|\mathbf{O}_{-\mathbf{A}^{-1}} - \mathbf{O}_{\mathbf{E}}| \\ &= \lambda^{-1}\left(-\cos\alpha\sin\frac{\beta}{2}, -\sin\alpha\sin\frac{\beta}{2}, -\cos\frac{\beta}{2}\right)^t. \end{aligned} \tag{57}$$

To find the intersecting circle between Ewald spheres $\mathbf{A}(1)$ and $-\mathbf{A}(2)$ for a given pair of experimental diffraction patterns $s_{\mathbf{A}(1)}(\xi,\varphi)$ and $s_{\mathbf{A}(2)}(\xi,\varphi)$, we calculate the following quantity for all assumed values of Ψ' in the range of $[0, 2\pi]$:

$$s_{\mathbf{A}(1)}(\xi,\varphi) - s_{\mathbf{A}(2)}(\xi,\varphi-\Psi'). \tag{58}$$

When we move clockwise along the intersecting circle on Ewald sphere $\mathbf{A}(1)$, it appears as an anticlockwise motion along the intersecting circle on Ewald sphere $-\mathbf{A}(2)$. Because Ewald spheres $\mathbf{A}(2)$ and $-\mathbf{A}(2)$ have opposite parities, this motion appears as a clockwise motion along its centrosymmetric circle on Ewald sphere $\mathbf{A}(2)$. This is the reason why the sign in front of φ in $s_{\mathbf{A}(2)}$ in equation (58), unlike in equation (54), is now plus. When the quantity of equation (58) vanishes on a circle with its centre at $\xi = \Xi', \varphi = \Phi'$, the polar coordinates of the centres of intersecting circles are (Ξ', Φ') on $s_{\mathbf{A}(1)}(\xi,\varphi)$ and $(\Xi', \Phi'-\Psi')$ on $s_{\mathbf{A}(2)}(\xi,\varphi)$. These polar coordinates are to be compared with equation (57). Therefore

$$\begin{cases} \sin\Xi'\cos\Phi' = \cos\gamma\sin\dfrac{\beta}{2}, & \sin\Xi'\sin\Phi' = -\sin\gamma\sin\dfrac{\beta}{2}, \\ \sin\Xi'\cos(\Phi'-\Psi') = -\cos\alpha\sin\dfrac{\beta}{2}, & \sin\Xi'\sin(\Phi'-\Psi') = -\sin\alpha\sin\dfrac{\beta}{2}. \end{cases} \tag{59}$$

From these relations we have another half of equation (21) in the main text.

APPENDIX *C*
Derivation of equations (11), (12), (13) and (22)

The mean of the quantities of equations (6) and (7) with respect to the Poisson distribution is given by

$$\begin{aligned} \langle\Psi_{ij}(\xi,\alpha)\rangle &= \frac{1}{N_\xi}\sum_{l=0}^{N_\xi-1} s\left(i;\xi,\frac{2\pi l}{N_\xi}\right) s\left(j;\xi,\frac{2\pi l}{N_\xi}+\alpha\right) \\ &= \frac{1}{2\pi}\int_0^{2\pi} \mathrm{d}\varphi\, s(i;\xi,\varphi)s(j;\xi,\varphi+\alpha) \\ &= \frac{C^2\omega^2}{2\pi}\int_0^{2\pi} \mathrm{d}\varphi\, i(0,\beta_i,\gamma_i;\xi,\varphi) \\ &\quad \times i(0,\beta_j,\gamma_j;\xi,\varphi+\alpha+\alpha_i-\alpha_j). \end{aligned} \tag{60}$$

$$\langle\bar{s}_Q(i;\xi)\rangle = \frac{1}{N_\xi}\sum_{l=0}^{N_\xi-1} s\left(i;\xi,\frac{2\pi l}{N_\xi}\right) = \frac{C\omega}{2\pi}\int_0^{2\pi}\mathrm{d}\varphi\, i(i;\xi,\varphi). \tag{61}$$

When the quantity of equation (60) is averaged over the structure irregularity distribution, we have from equation (9) and the relation $\langle i^2\rangle = 2\langle i\rangle^2$ due to the exponential distribution

$$\langle\langle\Psi_{ij}(\xi,\alpha)\rangle\rangle = \left(C\omega\bar{i}\right)^2\left(1+\frac{1}{2\pi}\int_0^{2\pi}\mathrm{d}\varphi\exp\left\{-\left[\frac{k\delta(\varphi)}{k_C}\right]^2\right\}\right), \tag{62}$$

where $\delta(\varphi)$ is an angle between the two $\mathbf{k}$ vectors given explicitly as follows according to equation (41):

$$\begin{aligned} \mathbf{k}_i &= \mathbf{A}(i)\mathbf{h}(\xi,\varphi) = \mathbf{A}_3(-\gamma_i)\mathbf{A}_2(-\beta_i)\mathbf{A}_3(-\alpha_i)\mathbf{h}(\xi,\varphi), \\ \mathbf{k}_j &= \mathbf{A}(j)\mathbf{h}(\xi,\varphi+\alpha) = \mathbf{A}_3(-\gamma_j)\mathbf{A}_2(-\beta_j)\mathbf{A}_3(-\alpha_j)\mathbf{h}(\xi,\varphi+\alpha). \end{aligned} \tag{63}$$

When the two $\mathbf{k}$ vectors are always significantly different, equation (62) reduces to

$$\langle\langle\Psi_{\mathrm{BG}}(\xi,\alpha)\rangle\rangle = \left(C\omega\bar{i}\right)^2. \tag{64}$$

The average of the quantity of equation (61) over the structure irregularity distribution is given by

$$\langle\langle\bar{s}_Q(i;\xi)\rangle\rangle = C\omega\bar{i}. \tag{65}$$

By introducing an approximation

$$\langle\langle c_{ij}(\xi,\alpha)\rangle\rangle \cong \frac{\langle\langle\Psi_{ij}(\xi,\alpha)\rangle\rangle}{\langle\langle\bar{s}_Q(i;\xi)\rangle\rangle\langle\langle\bar{s}_Q(j;\xi)\rangle\rangle} - 1, \tag{66}$$

we see from equations (64) and (65) that $\langle\langle c_{\mathrm{BG}}(\xi,\alpha)\rangle\rangle$ vanishes as mentioned in the main text. Also, from equations (62), (65) and (66) we have

$$\langle\langle c_{ij}(\xi,\alpha)\rangle\rangle = \frac{1}{2\pi}\int_0^{2\pi}\mathrm{d}\varphi\exp\left\{-\left[\frac{k\delta(\varphi)}{k_C}\right]^2\right\}. \tag{67}$$

This is an expected expression when all background features having random appearance due to the quantum noise and the structure irregularity distribution are erased from such high correlation lines as observed in Figs. 2(*a*) and 2(*b*).

We now proceed to simplify equation (67). At first we calculate $\delta(\varphi)$, an angle between the two vectors of equation

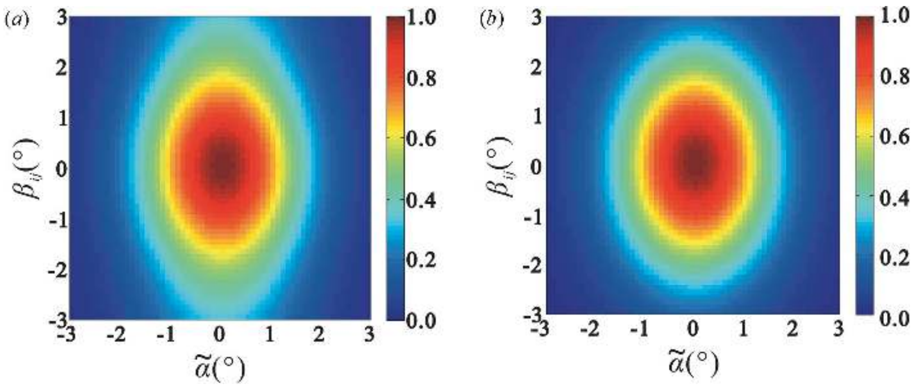

Figure 9
The quantity of equation (67) shown as a function of $\tilde{\alpha}$ and β_{ij} by assuming $k = 0.174$ Å^{-1} [value corresponding to $\xi = 10°$ and $\lambda = 1$ Å in equation (2)] and $1/k_C = 200$ Å (the value for the HslUV complex). (*a*) Exact but numerically calculated value. (*b*) Value according to the approximate but analytic expression of equation (74).

(63). The magnitude of the two vectors is given by equation (2). To calculate the inner product between the two vectors, we introduce the relative Eulerian angle $\left(\alpha_{ij}, \beta_{ij}, \gamma_{ij}\right)$ by

$$\begin{aligned}\mathbf{A}(i)^{-1}\mathbf{A}(j) &= \mathbf{A}_3(\alpha_i)\mathbf{A}_2(\beta_i)\mathbf{A}_3(\gamma_i)\mathbf{A}_3\left(-\gamma_j\right)\mathbf{A}_2\left(-\beta_j\right)\mathbf{A}_3\left(-\alpha_j\right)\\ &= \mathbf{A}_3\left(-\gamma_{ij}\right)\mathbf{A}_2\left(-\beta_{ij}\right)\mathbf{A}_3\left(-\alpha_{ij}\right). \qquad (68)\end{aligned}$$

Here β_{ij} is the angle between beam directions given by equation (5). In terms of the relative Eulerian angle, we have

$$\begin{aligned}\cos\delta(\varphi) = \frac{\mathbf{k}_i^t\mathbf{k}_j}{k^2} &= \sin^2\frac{\xi}{2}\cos\beta_{ij} + \sin\frac{\xi}{2}\cos\frac{\xi}{2}\sin\beta_{ij}\\ &\quad\times[\cos(\tilde{\varphi}+\tilde{\alpha}) - \cos\tilde{\varphi}] + \cos^2\frac{\xi}{2}\\ &\quad\times\left[\cos\beta_{ij}\cos(\tilde{\varphi}+\tilde{\alpha})\cos\tilde{\varphi} + \sin(\tilde{\varphi}+\tilde{\alpha})\sin\tilde{\varphi}\right],\\ \tilde{\varphi} &= \varphi + \gamma_{ij}, \qquad (69)\end{aligned}$$

$$\tilde{\alpha} = \alpha - \hat{\alpha}_{ij}, \quad \hat{\alpha}_{ij} = \alpha_{ij} + \gamma_{ij}. \qquad (70)$$

The quantity of equation (67) is to be obtained by substituting $\delta(\varphi)$ from this equation. Let us write the result of the integration with respect to φ in equation (67) by writing three independent variables explicitly as $\left\langle\!\left\langle c\left(k/k_C,\tilde{\alpha}, \beta_{ij}\right)\right\rangle\!\right\rangle$. Here, remember that ξ and k are related by equation (2). Even though it is difficult to derive an analytic form of this quantity, we can calculate it numerically. Fig. 9(*a*) shows an example of this quantity obtained numerically. For practical applications we want to have an analytic form of this quantity even when it may be somewhat approximate. To derive an approximate but analytic form we at first notice that the integrand of equation (67) has a significant contribution only when $\tilde{\alpha}$ and β_{ij} are small. By using this fact we Taylor-expand both sides of equation (69) in terms of $\delta(\varphi)$, $\tilde{\alpha}$ and β_{ij}, and retain only up to the second-order terms to obtain

$$\begin{aligned}\delta(\varphi)^2 &= \tilde{\alpha}^2\cos^2\frac{\xi}{2} + {\beta_{ij}}^2\left(1 - \cos^2\frac{\xi}{2}\sin^2\tilde{\varphi}\right)\\ &\quad + \tilde{\alpha}\beta_{ij}\sin\xi\sin\tilde{\varphi}. \qquad (71)\end{aligned}$$

The quantity of equation (67), obtained numerically by using this expression, was found to have almost no difference from $\left\langle\!\left\langle c\left(k/k_C,\tilde{\alpha}, \beta_{ij}\right)\right\rangle\!\right\rangle$, meaning that equation (71) is a very good approximation to equation (69). Even for this expression of equation (71), the derivation of an analytic form for the integral of equation (67) is difficult. In this situation we introduce a drastic approximation of replacing the quantity of equation (71) by its mean

$$\left\langle\delta^2\right\rangle = \tilde{\alpha}^2\left(1 - \sin^2\frac{\xi}{2}\right) + \frac{{\beta_{ij}}^2}{2}\left(1 + \sin^2\frac{\xi}{2}\right), \qquad (72)$$

and use the value of δ thus obtained in equation (67). Writing the result thus obtained as $\left\langle\!\left\langle c_1\left(k/k_C,\tilde{\alpha}, \beta_{ij}\right)\right\rangle\!\right\rangle$ we have

$$\begin{aligned}\left\langle\!\left\langle c_1\left(k/k_C, \tilde{\alpha}, \beta_{ij}\right)\right\rangle\!\right\rangle = \exp\Bigg\{-\left(\frac{k}{k_C}\right)^2\Bigg[\tilde{\alpha}^2\left(1 - \sin^2\frac{\xi}{2}\right)\\ + \frac{\beta_{ij}^2}{2}\left(1 + \sin^2\frac{\xi}{2}\right)\Bigg]\Bigg\}. \qquad (73)\end{aligned}$$

For practical applications we can approximate this expression further as

$$\left\langle\!\left\langle c_2\left(k/k_C,\alpha, \beta\right)\right\rangle\!\right\rangle = \exp\left[-\left(\frac{k}{k_C}\right)^2\left(\tilde{\alpha}^2 + \frac{\beta_{ij}^2}{2}\right)\right]. \qquad (74)$$

Fig. 9(*b*) is a plot of this quantity. By comparing it with Fig. 9(*a*), we see that this simple analytic form is a reasonably good approximation.

Equation (74) is equation (11) of the main text. Equation (12) can be obtained from equations (68) and (70) by assuming that β_{ij} and therefore both $\left|\beta_i - \beta_j\right|$ and $\left|\gamma_i - \gamma_j\right|$ are small quantities of the same order.

For the purpose of deriving equation (13), we evaluate the standard deviations of the quantities appearing in equations (6) and (7). After some calculations they are obtained as

$$\begin{aligned}\sigma^2_{\Psi_{ij}} &= \left\langle\!\!\left\langle\left(\Psi_{ij}\right)^2\right\rangle\!\!\right\rangle - \left\langle\!\left\langle\left(\Psi_{ij}\right)\right\rangle\!\right\rangle^2 = \frac{1}{N_\xi}\left[3\left(C\omega\bar{i}\right)^4 + 4\left(C\omega\bar{i}\right)^3 + \left(C\omega\bar{i}\right)^2\right],\\ \sigma^2_{\bar{s}_Q} &= \left\langle\!\!\left\langle\left(\bar{s}_Q\right)^2\right\rangle\!\!\right\rangle - \left\langle\!\left\langle\left(\bar{s}_Q\right)\right\rangle\!\right\rangle^2 = \frac{1}{N_\xi}\left[\left(C\omega\bar{i}\right)^2 + \left(C\omega\bar{i}\right)\right]. \qquad (75)\end{aligned}$$

During the derivation of this equation we assumed that the values of the diffraction intensity density at the neighbouring Shannon pixels on a circle are not correlated. This means that the distance in the wavenumber space between the neighbouring Shannon pixels must be larger than the correlation length k_C of equation (10). Because the correlation fades quickly beyond k_C, we assume that equation (75) still holds for this distance. Therefore we have equation (15). The standard deviation of the correlation pattern is then given by

$$\sigma_c^2 = \frac{2\sigma_{\overline{s}_Q}^2}{\langle\langle \overline{s}_Q \rangle\rangle^2} + \frac{\sigma_{\Psi_{ij}}^2}{\langle\langle \Psi_{ij} \rangle\rangle^2}, \tag{76}$$

from which we have equation (13).

We now proceed to derive equation (22). We study how the value of the product $N_A s_N$ is determined for common circles to be identified. At first we note that a sum of Poisson distributions is a Poisson distribution. Therefore, a sum of N_A two-dimensional patterns is equal to one sample of a two-dimensional pattern whose expected photon number is given by $N_A s(\mathbf{k})$, which we shall write as η in equation (77) below. We shall also write the mean of η on a sphere $k = |\mathbf{k}|$ as $\overline{\eta}$. Now we consider the distribution of values of the difference $D = s_1(\xi, \varphi) - s_2(\xi, \varphi')$ (1) on and (2) off the common circle. (1) On the common circle, D is given by a difference of two integers, both given by a Poisson distribution. Its mean vanishes of course. The mean of its absolute value may be estimated approximately as

$$\langle\langle |D|_{\rm on} \rangle\rangle \cong \left(\langle\langle D^2 \rangle\rangle\right)^{1/2} = (2\overline{\eta})^{1/2}. \tag{77}$$

(2) Off the common circle, the values of $s_1(\xi, \varphi)$ and $s_2(\xi, \varphi')$ are both distributed according to the two distributions, the Poisson distribution and the exponential distribution. The composite distribution is given by

$$P(n) = (1-\zeta)\zeta^n, \quad \zeta = \frac{\overline{\eta}}{1+\overline{\eta}}. \tag{78}$$

For this distribution, the mean of the absolute value is estimated approximately as follows:

$$\langle\langle |D|_{\rm off} \rangle\rangle \cong \left(\langle\langle D^2 \rangle\rangle\right)^{1/2} = \left(2\overline{\eta} + 2\overline{\eta}^2\right)^{1/2}. \tag{79}$$

Let us assume that, for clear recognition of $|D|_{\rm on}$ out of $|D|_{\rm off}$, the following condition must be satisfied up to $k = k_N$: $\langle\langle |D|_{\rm on} \rangle\rangle \leq \langle\langle |D|_{\rm off} \rangle\rangle/3$, which means

$$\overline{\eta} \geq 8. \tag{80}$$

Equation (22) follows directly from this equation.

This study has been supported by the 'X-ray Free Electron Laser Utilization Research Project' of the Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT). We are heavily indebted to Dr Yasumasa Joti for his help during the revision of the manuscript.

References

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.

Bortel, G. & Faigel, G. (2007). *J. Struct. Biol.* **158**, 10–18.

Bortel, G., Faigel, G. & Tegze, M. (2009). *J. Struct. Biol.* **166**, 226–233.

Elser, V. (2003). *J. Opt. Soc. Am. A*, **20**, 40–55.

Elser, V. (2009). *IEEE Trans. Inf. Theory*, **55**, 4715–4722.

Fienup, J. R. (1982). *Appl. Opt.* **21**, 2758–2769.

Fung, R., Shneerson, V., Saldin, D. K. & Ourmazd, A. (2009). *Nat. Phys.* **5**, 64–67.

Huldt, G., Szoeke, A. & Hajdu, J. (2003). *J. Struct. Biol.* **144**, 219–227.

Loh, N. D. *et al.* (2010). *Phys. Rev. Lett.* **104**, 225501.

Loh, N. D. & Elser, V. (2009). *Phys. Rev. E*, **80**, 026705.

Mimura, H. *et al.* (2010). *Nat. Phys.* **6**, 122–125.

Nishino, Y., Takahashi, Y., Imamoto, N., Ishikawa, T. & Maeshima, K. (2009). *Phys. Rev. Lett.* **102**, 018101.

Seibert, M. M. *et al.* (2011). *Nature (London)*, **470**, 78–81.

Shneerson, V. L., Ourmazd, A. & Saldin, D. K. (2008). *Acta Cryst.* A**64**, 303–315.

Singer, A. & Shkolnisky, Y. (2011). *SIAM J. Imag. Sci.* **4**, 543–572.

Sousa, M. C., Kessler, B. M., Overkleeft, H. S. & McKay, D. B. (2002). *J. Mol. Biol.* **318**, 779–785.

Weaver, L. H. & Matthews, B. W. (1987). *J. Mol. Biol.* **193**, 189–199.

Wilson, A. J. C. (1949). *Acta Cryst.* **2**, 318–321.

Yang, C., Wang, Z. & Marchesini, S. (2010). *Proc. SPIE*, **7800**, 78000P1–78000P10.