The extraction of single-particle diffraction patterns from a multiple-particle diffraction pattern

A.V. Martin,[1,*] A.J. Morgan,[2] T. Ekeberg,[3] N.D. Loh,[4] F.R.N.C. Maia,[3] F. Wang,[5] J.C.H. Spence,[6] and H.N. Chapman[5,7]

[1] *ARC Centre of Excellence for Coherent X-ray Science, School of Physics, The University of Melbourne, Victoria, 3010, Australia*
[2] *School of Physics, The University of Melbourne, Victoria, 3010, Australia*
[3] *Laboratory of Molecular Biophysics, Department of Cell and Molecular Biology, Uppsala University, Husargatan 3 (Box 596), SE-751 24 Uppsala, Sweden*
[4] *PULSE Institute, SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, CA 94025 U.S.A.*
[5] *Center for Free-Electron Laser Science, DESY, Notkestrasse 85, 22607 Hamburg, Germany*
[6] *Department of Physics, Arizona State University, Tempe, Arizona 85287, U.S.A.*
[7] *University of Hamburg, Laruper Chausee 149, 22761 Hamburg, Germany*

[*] *andrew.martin@unimelb.edu.au*

Abstract: The structures of biological molecules may soon be determined with X-ray free-electron lasers without crystallization by recording the coherent diffraction patterns of many identical copies of a molecule. Most analysis methods require a measurement of each molecule individually. However, current injection methods deliver particles to the X-ray beam stochastically and the maximum yield of single particle measurements is 37% at optimal concentration. The remaining 63% of pulses intercept no particles or multiple particles. We demonstrate that in the latter case single particle diffraction patterns can be extracted provided the particles are sufficiently separated. The technique has the potential to greatly increase the amount of data available for three-dimensional imaging of identical particles with X-ray lasers.

© 2013 Optical Society of America

OCIS codes: (110.7440) X-ray imaging; (110.3010) Image reconstruction techniques.

References and links

1. V. Ayvazyan, N. Baboi, J. Bähr, V. Balandin, B. Beutner, A. Brandt, I. Bohnet, A. Bolzmann, R. Brinkmann, and O. I. Brovko, "First operation of a free-electron laser generating gw power radiation at 32 nm wavelength," Eur. Phys. J. D **37**, 297–303 (2006).
2. P. Emma, R. Akre, J. Arthur, R. M. Bionta, C. Bostedt, J. Bozek, A. Brachmann, P. H. Bucksbaum, R. Coffee, F. J. Decker, Y. Ding, D. Dowell, S. Edstrom, A. Fisher, J. Frisch, S. Gilevich, J. Hastings, G. Hays, P. Hering, Z. Huang, R. Iverson, H. Loos, M. Messerschmidt, A. Miahnahri, S. Moeller, H. -D. Nuhn, D. Pile, D. Ratner, J. Rzepiela, D. Schultz, T. Smith, P. Stefan, H. Tompkins, J. Turner, J. Welch, W. White, J. Wu, G. Yocky, and J. N. Galayda, "First lasing and operation of an angstrom-wavelength free-electron laser," Nat. Photon. **4**, 641 (2010).
3. N. G. A. Abrescia, D. H. Bamford, J. M. Grimes, and D. I. Stuart, "Structure unifies the viral universe," Annu. Rev. Biochem. **81**, 795–822 (2012).
4. E. P. Carpenter, K. Beis, A. D. Cameron, and S. Iwata, "Overcoming the challenges of membrane protein crystallography," Current Opinion in Structural Biology **18**, 581–586 (2008).
5. G. Huldt, A. Szoke, and J. Hajdu, "Diffraction imaging of single particles and biomolecules," J. Struct. Biol. **144**, 219–227 (2003).

6. D. P. DePonte, U. Weierstall, K. Schmidt, J. Warner, D. Starodub, J. C. H. Spence, and R. B. Doak, “Gas dynamic virtual nozzle for generation of microscopic droplet streams,” J. Phys. D: Appl. Phys. **41**, 195505 (2008).
7. M. J. Bogan, S. Boutet, H. N. Chapman, S. Marchesini, A. Barty, W. H. Benner, U. Rohner, M. Frank, S. P. Hau-Riege, S. Bajt, B. W. Woods, M. M. Seibert, B. Iwan, N. Timneanu, J. Hajdu, and J. Schulz, “Aerosol imaging with a soft x-ray free electron laser,” Aerosol Sci. Tech. **44**, i–vi (2010).
8. N. D. Loh and V. Elser, “Reconstruction algorithm for single-particle diffraction imaging experiments,” Phys. Rev. E **80**, 026705 (2009).
9. R. Fung, V. Shneerson, D. K. Saldin, and A. Ourmazd, “Structure from fleeting illumination of faint spinning objects in flight,” Nat. Phys. **5**, 64–67 (2009).
10. G. Bortel and M. Tegze, “Common arc method for diffraction pattern orientation,” Acta Crystallogr. **A67**, 533–543 (2011).
11. D. Starodub, A. Aquila, S. Bajt, M. Barthelmess, A. Barty, C. Bostedt, J. D. Bozek, N. Coppola, R. B. Doak, S. W. Epp, B. Erk, L. Foucar, L. Gumprecht, C. Y. Hampton, A. Hartmann, R. Hartmann, P. Holl, S. Kassemeyer, N. Kimmel, H. Laksmono, M. Liang, N. D. Loh, L. Lomb, A. V. Martin, K. Nass, C. Reich, D. Rolles, B. Rudek, A. Rudenko, J. Schulz, R. L. Shoeman, R. G. Sierra, H. Soltau, J. Steinbrener, F. Stellato, S. Stern, G. Weidenspointner, M. Frank, J. Ullrich, L. Strüder, I. Schlichting, H. N. Chapman, J. C. H. Spence, and M. J. Bogan, “Single-particle structure determination by correlations of snapshot x-ray diffraction patterns,” Nat. Commun. **3**, 1276 (2012).
12. M. J. Buerger, *Vector space and its application in crystal-structure investigation* (Wiley, New York, 1959).
13. H. He, S. Marchesini, M. Howells, U. Weierstall, G. Hembree, and J. C. H. Spence, “Experimental lensless soft-x-ray imaging using iterative algorithms: phasing diffuse scattering,” Acta Crystallogr. **A59**, 143–152 (2003).
14. X. Huang, J. Nelson, J. Steinbrener, J. Kirz, J. Turner, and C. Jacobsen, “Incorrect support and missing center tolerances of phasing algorithms.” Opt. Express **18**, 26441–26449 (2010).
15. P. Thibault, P. Elser, C. Jacobsen, D. Shapiro, and D. Sayre, “Reconstruction of a yeast cell from x-ray diffraction data,” Acta Crystallogr. **A62**, 248–261 (2006).
16. W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipies. The Art of Scientific Computing.* (Cambridge University Press, Cambridge, 1986).
17. A. Martin, N. Loh, C. Hampton, R. Sierra, F. Wang, A. Aquila, S. Bajt, M. Barthelmess, C. Bostedt, J. Bozek, N. Coppola, S. Epp, B. Erk, H. Fleckenstein, L. Foucar, M. Frank, H. Graafsma, L. Gumprecht, A. Hartmann, R. Hartmann, G. Hauser, H. Hirsemann, P. Holl, S. Kassemeyer, N. Kimmel, M. Liang, L. Lomb, F. Maia, S. Marchesini, K. Nass, E. Pedersoli, C. Reich, D. Rolles, B. Rudek, A. Rudenko, J. Schulz, R. Shoeman, H. Soltau, D. Starodub, J. Steinbrener, F. Stellato, L. Strüder, J. Ullrich, G. Weidenspointner, T. White, C. Wunderer, A. Barty, I. Schlichting, M. Bogan, and H. Chapman, “Femtosecond dark-field imaging with an x-ray free electron laser,” Opt. Express **20**, 13501–13512 (2012).

---

1. Introduction

The structures of single biological molecules, like membrane proteins, may soon be determined without the need for crystallization using X-ray free-electron lasers (XFELs) [1, 2]. Many biological samples like viruses and membrane proteins are difficult to crystallize [3, 4] and have remained beyond the reach of X-ray crystallography. In the proposed XFEL experiment, the three-dimensional structure will be extracted from an ensemble of two-dimensional diffraction measurements of a particle in different orientations. Since XFEL imaging is destructive, many identical copies of a particle will be measured [5]. This is achieved by injecting particles into the path of the X-ray pulses by either a liquid jet [6] or as an aerosol [7]. These injection methods deliver the molecules into the path the XFEL pulses in unknown random orientations. Determining the molecule’s 3D structure requires first determining the orientation of each molecule, then assembling the diffraction data into 3D in order to solve the phase problem, before finally recovering the real-space molecular structure. Most algorithms that accomplish the first step of finding the orientations only use measurements of single isolated molecules [8–10], with the exception of the correlation-based approach [11].

The arrival of particles in the path of the XFEL pulses is stochastic and determined by Poisson statistics. We use the term “hit” to refer to the case where a pulse has intercepted the sample. When the concentration is optimized so that on average there is one particle in the X-ray path at a time, then the expected hit rate is 63%. There are 37% single-particle hits and the remaining 26% are multiple-particle hits. A large amount of data from multiple-particle patterns is

therefore wasted. If the data from the multiple-particle patterns were useful, the efficiency of data collection could be greatly increased. This is beneficial for XFEL facilities that can not run multiple experiments in parallel, limiting the access that can be offered to the research community.

Here we show how to extract the diffracted signal of individual particles from a single-pulse measurement of multiple particles. The method requires that the separation between particles perpendicular to the beam axis is larger than their diameter, and that the beam is fully coherent on the length scale of the interparticle separation. These requirements can be easily satisfied if the sample is delivered suspended in a liquid jet or in a droplet. The geometric requirements are achievable with an XFEL because the beam may be a few microns in width, while viruses, proteins or similar samples are only tens or hundreds of nanometers in size. This condition will be more difficult to satisfy if the beam size is reduced close to the particle size (~0.2 micron), which may be desirable if the signal-per-particle is more important than the hit-rate. In that case, there is a significant probability that that multiple particles will lie along the path of the beam. The analysis of these in-line hits is beyond the scope of this work.

It is worth noting that the work presented here contains similarities to the analysis of autocorrelation functions in crystallography (known as Patterson maps) to reveal atomic structures [12]. The goal of the crystallographic analysis is to recover an "image" of the underlying atomic structure of the sample. The comparison can be drawn by replacing the atoms of the crystallographic analysis with the particles (or molecules) in our analysis. The key difference is that we only require the autocorrelation functions of individual particles, whereas the crystallographic analysis seeks an image of the object.

2. Theory

Single-particle diffraction patterns can be extracted from multiple-particle patterns by analyzing the autocorrelation function of the scattered wave $\psi(\boldsymbol{r})$. At the relatively low resolution considered here, we can define a plane just post the specimen, called the exit-surface plane, where the scattered wave is approximately proportional to the projection of the object (equivalent to approximating the Ewald sphere by a plane). This projection approximation holds when the smallest resolved distance is greater than $\sqrt{(\lambda t)/2}$, where λ is the wavelength and t is the thickness of the object. We define the vector $\boldsymbol{r}$ to be in the exit-surface plane of the object. The scattered exit-surface wave $\psi(\boldsymbol{r})$ and the detector-plane wave $\psi(\boldsymbol{q})$ are linked by the Fourier transform $\psi(\boldsymbol{q}) = \mathscr{F}[\psi(\boldsymbol{r})]$, where $\mathscr{F}$ denotes the Fourier transform. Using the convolution theorem, the auto-correlation function $A(r)$ can be obtained from the measured diffraction pattern $I_M(\boldsymbol{q}) = \psi^*(\boldsymbol{q})\psi(\boldsymbol{q})$ as follows,

The autocorrelation function $A(\boldsymbol{r})$ is formed by taking the inverse Fourier transform of the measured diffraction pattern $I_M(\boldsymbol{q})$ as follows,

$$\begin{aligned} A(\boldsymbol{r}) &= \mathscr{F}^{-1}[I_M(\boldsymbol{q})] \\ &= \psi^*(-\boldsymbol{r}) \otimes \psi(\boldsymbol{r}) \\ &= \int \psi^*(\boldsymbol{r'})\psi(\boldsymbol{r}+\boldsymbol{r'})d\boldsymbol{r'} \ , \end{aligned} \quad (1)$$

where $\otimes$ denotes a convolution. It is assumed that the unscattered XFEL pulse is not measured in the diffraction pattern, which has been the case in experiments so far and the practical issues that arise from a central detector hole will be discussed further below. When there are multiple particles in the beam, the scattered wave is the coherent sum of the scattered waves of each of the particles. Defining $\psi_n(\boldsymbol{r})$ to be the scattered wave from the nth particle, such that $\psi(\mathbf{r}) =$

$\sum_n \psi_n(\boldsymbol{r})$, the autocorrelation function can be written as

$$A(\boldsymbol{r}) = \sum_i \psi_i^*(-\boldsymbol{r}) \otimes \psi_i(\boldsymbol{r}) + \sum_i \sum_{j \neq i} \psi_i^*(-\boldsymbol{r}) \otimes \psi_j(\boldsymbol{r}) \; . \tag{2}$$

The first sum on the r.h.s. of Eq. (2) contains the autocorrelation function of the scattered wave from each particle, while the second sum contains cross-correlation terms between the scattered waves from different particles. If there is a sufficient interparticle spacing, then the cross-correlation terms are situated in spatially distinct regions of the autocorrelation function, as shown by the example in Fig. 1, and clearly seen in earlier soft X-ray diffraction patterns from clusters of gold balls, where this separation occurs [13]. Figure 1(a) shows three copies of a "virus" in random orientations, where the distance between virus particles is much larger than the size of a single virus. The corresponding diffraction pattern is shown in Fig. 1(b) scaled to the power of 0.3. The amplitude of the autocorrelation function, Fig. 1(c), shows the cross-correlation terms. Each is displaced from the center and spatially separated from the other cross-correlation terms. The sum of the autocorrelation functions for each individual particle, given by the first sum on the r.h.s. of Eq. (2), is located at the center of the autocorrelation function.

In the case that two particles have a large relative separation along the beam axis, then the scattered wave from the particle closest the X-ray source will propagate and spread before the pulse interacts with a second particle. The cross-correlation term between the two particles can be interpreted as the correlation between the propagated wave of the first particle and the exit-surface wave of the second particle. Since the propagated wave spreads over a larger area, the cross-correlation term between the two particles increases in size as the separation distance along the beam axis increases. This will only prevent the recovery of single-particle diffraction patterns if the cross-correlation term overlaps with a neighboring term.

The regions containing the cross-correlation terms can be identified using a statistical threshold. First the mean and standard deviation of the entire autocorrelation function is calculated. Then pixels greater than 2 standard deviations from the mean are excluded and the standard deviation of the remaining pixels is then calculated. Pixels where the amplitude of the autocorrelation function are greater than 2.8 times the new standard deviation are regarded as signal, other pixels as background. The threshold parameters may be adjusted depending on the dataset. The values quoted here gave adequate results for all applications shown in this paper. After the statistical threshold has been applied, binary morphology operations of opening, closing and dilation were applied in that order with a small circular kernel to ensure that only large connected regions of interest are identified. The opening operation removes small clusters of pixels from the mask, the closing operation removes any small gaps between larger connected regions of the mask, and the final dilation helps to avoid clipping the edges of the cross-correlation terms. For simulation, a kernel of 4 pixel radius was used for opening, a kernel of 8 pixel radius was used for closing, and a kernel of 20 pixel radius was used for dilation. These parameters worked for both our simulations with and without noise, but may need to be adjusted on a case-by-case basis. For the simulation example, the regions that were found by the thresholding procedure are shown in white in Fig. 1(d). The statistical threshold is effective at identifying regions of signal when there is background noise in the autocorrelation function.

Taking the Fourier transform of each cross-correlation term separately, we obtain

$$C_{ij}(\boldsymbol{q}) \equiv \psi_i^*(\boldsymbol{q})\psi_j(\boldsymbol{q}) \; , \tag{3}$$

where the vector $\boldsymbol{q}$ is defined in the plane of the detector. It can be shown with simple algebra that the diffraction pattern for the jth particle is given by

$$I_j(\boldsymbol{q}) \equiv \psi_j^*(\boldsymbol{q})\psi_j(\boldsymbol{q}) = \frac{C_{ij}(\boldsymbol{q})C_{jk}(\boldsymbol{q})}{C_{ik}(\boldsymbol{q})} \; . \tag{4}$$

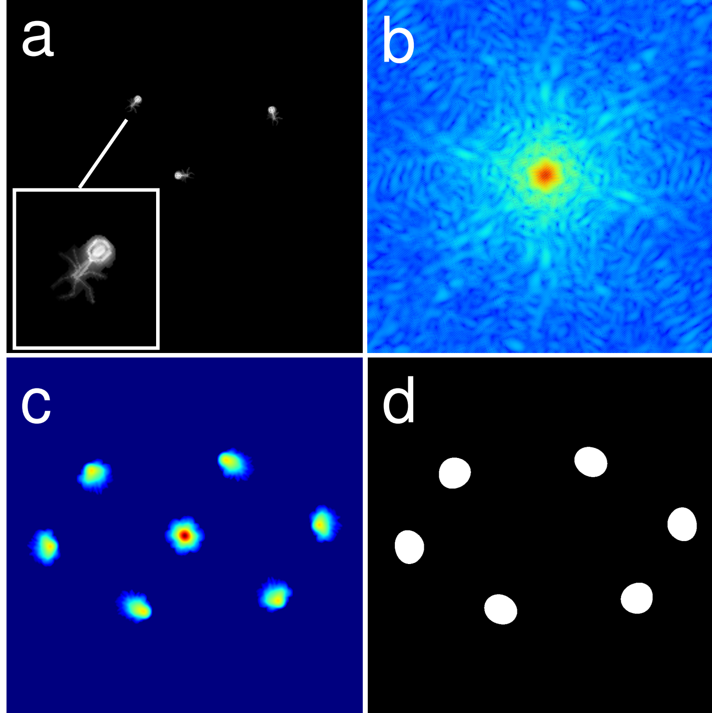

Fig. 1. (a) Three copies of a simulated image of a "virus" in random orientations, with a close up of virus image inset, (b) the corresponding diffraction pattern, (c) the autocorrelation function and (d) the mask showing the regions occupied by the cross-correlation terms.

Equation (4) allows each of the single particle patterns to be calculated, provided the correct triplets of cross-correlation terms can be identified. Since the l.h.s. of Eq. (4) represents an intensity it must be a real and positive function, but the r.h.s. is calculated from complex functions. The degree to which the r.h.s. is real and positive can measured with the following metric:

$$\varepsilon_{\mathrm{rp}} = \sqrt{\frac{\sum_m |[\psi^R(\boldsymbol{q}_m) > 0] - \psi(\boldsymbol{q}_m)|^2}{\sum_m |\psi(\boldsymbol{q}_m)|^2}}\ , \tag{5}$$

where $\boldsymbol{q}_m$ denotes the vector of the mth detector pixel and superscript R denotes the real part of a complex function. It turns out that Eq. (5) provides a very good measure of whether a triplet has been identified and a diffraction pattern can be extracted. In the example simulation without noise, triplets which do not yield a diffraction pattern had a value of $\varepsilon_{\mathrm{rp}}$ greater than 0.85, while the correct triplets had a value less than 0.05.

A second method for identifying triplets is based on distance. We label the centroids of the three cross-correlation terms in a triplet as A, B and C. We denote the displacement vector pointing in a direction from point A to B as $\overrightarrow{AB}$. Then the displacement vectors between the centroids of three cross-correlations that form a triplet must obey the relation:

$$\overrightarrow{AC} = \overrightarrow{AB} + \overrightarrow{BC}\ . \tag{6}$$

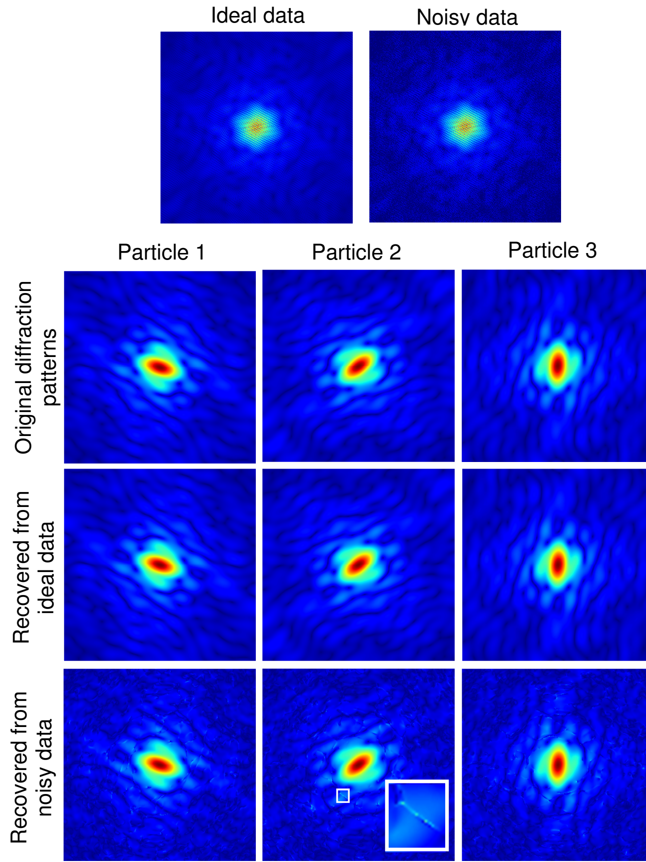

Fig. 2. The simulated multiple-particle diffraction pattern is shown at the top with and without noise. The upper row shows the original diffraction patterns for each individual particle. The second row shows recovered patterns from the ideal diffraction data. The third row shows the recovered single-particle patterns from the noisy data. Shot-noise was simulated assuming a total photon count of 5×10^6 photons in the pattern using the Poisson distribution. An artifact due to noise is shown inset in the bottom central diffraction pattern.

This was applied to our simulation with a tolerance of 5 pixels. The location of each cross-correlation term was set to the centroid of its mask. The three permutations identified by distance were the same as those found using $\varepsilon_{\mathrm{rp}}$. The advantage of checking triplets by distance is computational speed, since Eq. (5) does not need to be evaluated for every permutation. The metric $\varepsilon_{\mathrm{rp}}$ is still extremely useful, however, as it provides a final check that the triplet will yield an accurate single-particle diffraction pattern. In practice, cross-correlation terms may be affected by noise or artifacts. If a term is in fact two overlapping cross-correlation terms, it may still be included in a valid triplet identified by the distance metric, but would be identified as erroneous by $\varepsilon_{\mathrm{rp}}$.

Once the correct triplets have been identified, the single-particle diffraction patterns can be calculated using Eq. (4). For our simulation these are shown in Fig. 2 scaled to the power of 0.3. Visually they are identical to the original diffraction patterns. A sum-squared error metric

was used to compare the recovered diffraction patterns with the correct diffraction patterns:

$$\varepsilon_I = \frac{\sqrt{\sum_m [I_{\mathrm{recovered}}(\boldsymbol{q}_m) - I_{\mathrm{correct}}(\boldsymbol{q}_m)]^2}}{\sqrt{\sum_m I_{\mathrm{correct}}(\boldsymbol{q}_m)^2}} \,. \tag{7}$$

For simulated diffraction patterns without noise, the value of ε_I reflects the numerical precision of the computation.

3. Error analysis

The accuracy of the recovered diffraction patterns will be limited by noise. The biggest errors in this analysis are caused by an ill-posed division in Eq. (4). Where the cross-correlation term used in the denominator approaches zero, the division amplifies errors due to noise. Errors in both the measured numerator or denominator can be amplified. This causes artifacts in the reconstructed intensity in the form of thin lines of high intensity, as shown in the noisy simulation in Fig. 2. It is easy to spot the artifacts by eye because they are smaller than the characteristic speckle size for the object. This is expected because the denominator approaches zero along narrow lines or at points.

Before we can analyse the errors in the recovered diffraction patterns, we need to know the errors in each term $C_{ij}(\boldsymbol{q})$. We assume that noise on a pixel is not correlated with other pixels. Since $C_{ij}(\boldsymbol{q})$ is complex, there is a noise distribution for both the real and imaginary parts. We denote the standard deviation of these distributions by $\sigma[C^R_{ij}(\boldsymbol{q}_m)]$ and $\sigma[C^I_{ij}(\boldsymbol{q}_m)]$, where R and I refer the to real and imaginary parts respectively. We can relate these noise distributions to the standard deviation of the noise at the mth pixel on the measured diffraction pattern $\sigma[I_M(\boldsymbol{q}_m)]$ with

$$\sigma[C^R_{ij}(\boldsymbol{q}_m)] = \sqrt{\sum_n [M^R(\boldsymbol{q}_{n-m})]^2 \sigma^2[I_M(\boldsymbol{q}_n)]} \,, \tag{8}$$

and

$$\sigma[C^I_{ij}(\boldsymbol{q}_m)] = \sqrt{\sum_n [M^I(\boldsymbol{q}_{n-m})]^2 \sigma^2[I_M(\boldsymbol{q}_n)]} \,, \tag{9}$$

where $M(\boldsymbol{q}_m)$ is the Fourier transform of the mask used to select $C_{ij}(\boldsymbol{r})$. If $\sigma[C^R_{ij}(\boldsymbol{q}_m)] \approx \sigma[C^I_{ij}(\boldsymbol{q}_m)]$, then it turns out $\sigma[|C_{ij}(\boldsymbol{q}_m)|] \approx \sigma[C^R_{ij}(\boldsymbol{q}_m)]$ also. If the noise on the diffraction pattern is known, then we can use Eq. (8) directly.

In many experimental scenarios it will not be possible to characterise the noise level accurately enough to use Eq. (8). Fortunately there is a way of estimating $\sigma[|C_{ij}(\boldsymbol{q}_m)|]$ numerically with the measured autocorrelation function. The mask used to isolate the cross-correlation term is translated to a part of the autocorrelation function free of any cross-correlation term. In such a region, the value of the autocorrelation function is solely due to noise. Taking the Fourier transform of this patch provides a noise distribution in the detector plane. By using many noisy patches in the autocorrelation function, we can obtain many estimates of the noise at each detector pixel. Using this ensemble of estimates, the standard deviation of the noise at each pixel can be calculated. This numerical method was tested in our simulations and agreed with noise estimates using Eq. (8). This noise estimation relies on having sufficient sized regions of the autocorrelation function that don't contain any cross-correlation terms. For our example of three viruses this is certainly the case, though it may not be if many more than three particles are illuminated.

Once we have a noise estimate for each $C_{ij}(\boldsymbol{q})$, we can identify the regions of the single-particle diffraction patterns that have been accurately recovered. One useful approach is to analyse the signal-to-noise of the denominator of Eq. (4). A threshold can be set on this signal-to-noise level of, for example, $3\sigma[|C_{ik}(\boldsymbol{q}_m)|]$ in order to construct a mask that identifies the

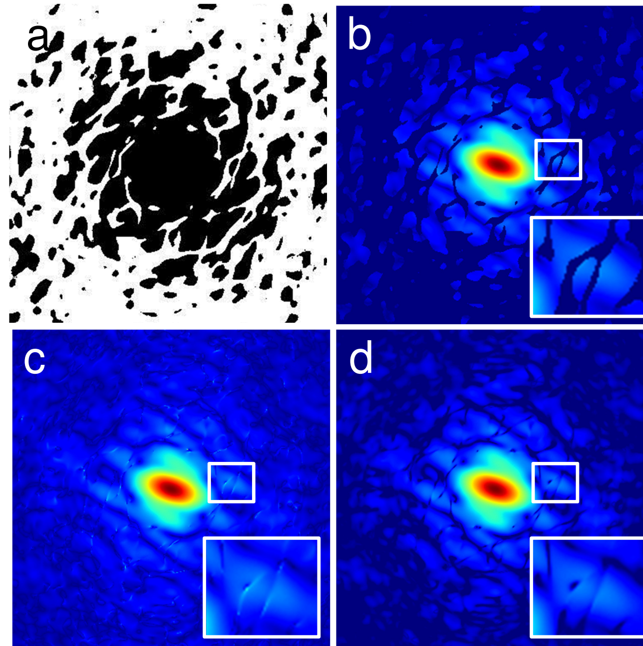

Fig. 3. (a) A mask for regions with a signal-to-noise level greater than three. (b) The data after applying the mask. (c) A recovered diffraction pattern from simulation without filtering and (d) after applying the Wiener filters.

accurately recovered regions. For the virus simulation with noise, this mask is shown in Fig. 3(a) and the masked recovered diffraction pattern is shown in Fig. 3(b). The original reconstruction from noisy data is shown again in Fig. 3(c) for comparison. The mask effectively excludes all the artifacts. If some of the data in the masked region still appears to have errors, this is due to residual errors in the numerator of Eq. (4).

After accurately recovered regions have been identified, the data can be used for 3D XFEL single particle imaging. For 3D XFEL imaging, the relative orientation of each particle is determined directly from the diffraction pattern. Fortunately algorithms for identifying the relative orientation do not need all the pixels on the detector. They can be adapted to function with a subset of pixels. Completeness of the final 3D dataset is achieved by combining a sufficient number of diffraction patterns.

Masking data may also be the preferable method of treating recovered diffraction data for 2D imaging. In this technique, phases that correspond to the measured intensities are found by iterative algorithms. The resulting complex wave function can be transformed to recover the projected structure of the object. Since not all pixels are recovered accurately by this analysis, the number of pixels available per particle has been effectively reduced. Missing data can be problematic for these algorithms. However, if the extent of the missing data region is smaller than a speckle, the data can be recovered during the reconstruction [14]. As shown in Fig. 3(a), in regions that have higher signal levels the regions of missing data are only thin gaps, often narrower than a speckle. It is likely that some of this information could be reconstructed. Whether missing data causes ambiguities for the recovery of the image can be determined in detail on a case-by-case basis [15].

Another method of handling the effects of noise is to use a Wiener filter [16]. There are errors

in our measurement of the numerator and the denominator of Eq. (4) and we will construct a separate filter for each. The filter to treat the noise of the numerator is

$$\phi_N(\mathbf{q}_m) = \frac{S_N^2(\mathbf{q}_m)}{S_N^2(\mathbf{q}_m) + \sigma^2[|C_{ij}(\boldsymbol{q}_m)C_{jk}(\boldsymbol{q}_m)|]} \ . \tag{10}$$

The "signal" $S_N(\mathbf{q}_m)$ is the ideal value of $C_{ij}(\mathbf{q}_m)C_{jk}(\mathbf{q}_m)$ unaffected by measurement errors. Since we only have a noisy measurement of $C_{ij}(\mathbf{q}_m)C_{jk}(\mathbf{q}_m)$, we estimate $S_N(\mathbf{q}_m)$ as [[16]] describes. The estimate is constructed as follows

$$\begin{aligned} S_N(\mathbf{q}_m) &= 0 \qquad\qquad \text{if } |C_{ij}(\boldsymbol{q}_m)C_{jk}(\boldsymbol{q}_m)| < \sigma[|C_{ij}(\boldsymbol{q}_m)C_{jk}(\boldsymbol{q}_m)|] \ , \\ S_N(\mathbf{q}_m) &= |C_{ij}(\boldsymbol{q}_m)C_{jk}(\boldsymbol{q}_m)| \\ &\qquad\qquad\quad \text{if } |C_{ij}(\boldsymbol{q}_m)C_{jk}(\boldsymbol{q}_m)| > \sigma[|C_{ij}(\boldsymbol{q}_m)C_{jk}(\boldsymbol{q}_m)|] \ . \end{aligned} \tag{11}$$

The filter to treat noise on the denominator is defined in a similar way:

$$\phi_D(\mathbf{q}_m) = \frac{S_D^2(\mathbf{q}_m)}{S_D^2(\mathbf{q}_m) + \sigma^2[|C_{ik}(\boldsymbol{q}_m))|]} \ , \tag{12}$$

where we define

$$\begin{aligned} S_D(\mathbf{q}_m) &= 0 \qquad\qquad\quad \text{if } |C_{ik}(\mathbf{q}_m)| < \sigma[|C_{ik}(\boldsymbol{q}_m))|] \ , \\ S_D(\mathbf{q}_m) &= |C_{ik}(\mathbf{q}_m)| \quad \text{if } |C_{ik}(\mathbf{q}_m)| > \sigma[|C_{ik}(\boldsymbol{q}_m))|] \ . \end{aligned} \tag{13}$$

The equation for the filtered diffraction pattern is given by

$$I_{j,\text{filtered}}(\mathbf{q}_m) = \phi_D(\mathbf{q}_m)\phi_N(\mathbf{q}_m)\frac{C_{ij}(\boldsymbol{q}_m)C_{jk}(\boldsymbol{q}_m)}{C_{ik}(\boldsymbol{q}_m)} \ . \tag{14}$$

The result of applying both filters to our simulation case is shown in Fig. 3(d). The noise-affected regions that are suppressed by the filter are very similar to those excluded from the masked data shown in Fig. 3(b). Here the variance of the uncertainty in the numerator (or denominator) was used to construct the Wiener filter, but this could be adjusted to produce a more conservative result. Replacing $\sigma[|C_{ik}(\boldsymbol{q}_m)|] \to 3\sigma[|C_{ik}(\boldsymbol{q}_m)|]$, for example, filters more of the data, which may be preferable in some applications.

In some applications it is necessary to quantify the error on each pixel of the recovered single particle diffraction patterns, e.g. determining particle orientation with Bayesian algorithms. The error in the recovered diffraction pattern is related to the errors in each cross-correlation term by the following equation:

$$\left(\frac{\sigma(I)}{|I|}\right)^2 = \left(\frac{\sigma(|C_{12}|)}{|C_{12}|}\right)^2 + \left(\frac{\sigma(|C_{13}|)}{|C_{13}|}\right)^2 + \left(\frac{\sigma_k(|C_{23}|)}{|C_{23}|}\right)^2 \ . \tag{15}$$

It is easier to see how the errors propagate through the division if we rearrange Eq. (15), so that it becomes

$$\sigma^2(I) = \frac{1}{|C_{13}|^2}\left[|C_{23}|^2\sigma^2(|C_{12}|) + |C_{12}|^2\sigma^2(|C_{23}|) + \frac{|C_{12}|^2|C_{23}|^2}{|C_{13}|^2}\sigma^2(|C_{13}|)\right] \ . \tag{16}$$

This shows that when $|C_{13}|$ is small, our uncertainty at those pixels is also very high. At these pixels, where the artifacts occur, Eq. (16) can not be used accurately because we encounter the same divide-by-zero instability. Equation (16) will still be useful for pixels where C_{13} is well measured, because this quantification of uncertainty is important for determining particle orientation with Bayesian algorithms.

4. Evaluating the accuracy of the recovered diffraction patterns

It is possible to evaluate the accuracy of a reconstruction using the central part of the auto-correlation function. If the diffraction patterns for all the particles exposed by the beam are recovered, then their sum should equal the central region of the autocorrelation function [the first term on the r.h.s. of Eq. (2)]. Defining the central region as C_{cen}, we measure the accuracy of the reconstructed diffraction patterns with

$$\varepsilon_{\text{cen}} = \sqrt{\frac{\sum_m F(\mathbf{q}_m)|C_{\text{cen}}(\mathbf{q}_m) - \sum_j I_j(\mathbf{q}_m)|^2}{\sum_m F(\mathbf{q}_m)|C_{\text{cen}}(\mathbf{q}_m)|^2}} \,, \tag{17}$$

where $F(\mathbf{q})$ is a filter used to select the region to make the comparison. Since the Wiener filter and the artifact mask are slightly different for each recovered diffraction pattern, we set $F(\mathbf{q})$ to be the product of the Wiener filters for each recovered pattern, or the product of the masks. The errors from the simulation with noise shown in Fig. 2 with no masking or filtering was 4.3×10^{-5}, which reduced to 1.2×10^{-5} for the filtered data. The masked data also produced a ε_{cen} of 1.2×10^{-5} when only pixels inside the mask are included in the calculation.

5. More than three particles

The analysis described for three particles can be applied without modification to the case of more than three particles. After the cross-correlation terms have been identified, triplets are identified exactly the same way as the three-particle case. When there are more than three particles, it is possible to calculate the same single particle diffraction pattern from different triplets of cross-correlation terms. This offers the chance to improve accuracy of the recovered pattern. However, there may also be a greater chance that cross-correlation terms overlap and can not be used in this analysis.

If the recovered diffraction patterns are of sufficient quality for whatever further analysis is to be performed, it is possible to maximise the yield of recovered single-particle data by increasing the concentration of the sample. For example, if the mean number of particles intercepted by a pulse was increased to greater than three, then more often three patterns would be extracted from each measurement. It is possible to imagine increasing the concentration until the number of recovered patterns exceeds the number of pulses in the experiment. However this would reduce the number of single-particle patterns measured directly.

6. Discussion

The methods described above do not apply to a two-particle diffraction pattern, because the corresponding auto-correlation function has only a single cross-correlation term. In the case where there is one particle in the interaction region on average, then 18% of the data will be two-particle hits. For these two-diffraction patterns, it may be possible to extract single-particle diffraction patterns by combining the single cross-correlation term $C_{12}(\mathbf{q})$ with the central part of the autocorrelation function $C_{cen}(\mathbf{q})$.

It has been proposed to use XFEL samples on nanometre-sized samples with orders of magnitude less scattering than the simulations shown here. The practical difficultly of applying our methods to very small samples may be the identification of the cross-correlation terms, which may be hard to distinguish from noise fluctuations. However, assuming that the cross-correlation terms can be identified, the error analysis methods of Section 3 could be employed to statistically characterize the recovered single-particle diffraction, so that it is useful for Bayesian methods of determining each particle's orientation, enabling the 3D structure to be determined [8, 9].

In practice, XFEL detectors are often modular and have physical gaps where diffraction data is not measured. Typically there is a central detector hole to allow the unscattered beam to pass. In this case, we can still recover single-particle diffraction data in the regions where the multiple particle diffraction data has been measured. Imagine a mask $G(\mathbf{q})$ that takes the value one where data has been measured and zero where data has not been measured. After applying this mask to the diffraction pattern and taking the inverse Fourier transform, we obtain a calculated autocorrelation function $A_{\mathrm{calc}}(\mathbf{r})$ that is related to the theoretical autocorrelation function $A_{\mathrm{theory}}(\mathbf{r})$ by the relation,

$$A_{\mathrm{calc}}(\mathbf{r}) = G(\mathbf{r}) \otimes A_{\mathrm{theory}}(\mathbf{r}) \, . \tag{18}$$

The potential difficulty is that the convolution with $G(\mathbf{r})$ can spread the cross-correlation terms causing them to overlap. To avoid this we can modulate the mask $G(\mathbf{q})$ with Gaussian functions as is done in dark-field coherent diffractive imaging [17].

7. Conclusion

We have shown that the autocorrelation function of multiple particles can be used to extract the diffraction patterns of each individual particle. Such multi-particle data is inevitably measured in XFEL imaging experiments using a liquid injector, and by the methods we propose, we can make this data useful for 3D XFEL imaging methods. Methods have been developed to handle experimental noise and to characterize the accuracy of the solution, which will help to determine the best use of the recovered data in image-reconstruction algorithms. These noise-handling methods will also be required in future efforts to apply this analysis to large XFEL datasets, where checking the individual results of each measurement may not be practically feasible.

Acknowledgements

This work was supported by the following agencies: the Australian Research Council through its Centres of Excellence programme; the Deutsches Elektronen-Synchrotron, a research center of the Helmholtz Association; the Swedish Research Council; the Knut and Alice Wallenberg Foundation; the European Research Council. N.D.L. was supported through by the Human Frontier Science Program and the AMOS program within the Chemical Sciences, Geo-sciences, and Biosciences Division of the Office of Basic Energy Sciences, Office of Science, U.S. Department of Energy. A.J.M. acknowledges financial assistance by the David Hay Memorial Fund.