A statistical model for X-ray scattering of a non-periodic sample to high angles is introduced. It is used to calculate analytically the correlation of distinct diffraction measurements of a particle as a continuous function of particle orientation. Diffraction measurements with shot-noise are also considered. This theory provides a general framework for a deeper understanding of single particle imaging techniques used at X-ray free-electron lasers. Many of these techniques use correlations as a measure of diffraction-pattern similarity in order to determine properties of the sample, such as particle orientation.

1. Introduction

X-ray free-electron lasers (XFELs) can produce pulses of sufficient peak brightness to probe individual viruses, nanoparticles and large biological molecules [1]. With XFEL sources, coherent diffractive imaging techniques can be used to extract structural information about the sample from a diffraction pattern. The study of non-periodic samples is known as ‘single particle imaging’ to distinguish it from crystallography, the study of periodic samples.

Single particle imaging experiments at X-ray laser facilities involve taking a large number of diffraction snapshots of individual particles, typically injected into the path of the X-ray pulses in solution or as an aerosol [2]. A large number of measurements is required, because X-rays diffract very weakly from an individual particle, even with the high intensities of XFEL pulses. It is already possible to collect this data, because X-ray lasers typically have high repetition rates (10 Hz at SACLA [3], 120 Hz at LCLS [4] and in the future 27 kHz at the European XFEL [5]). However, the rapid injection of particles makes it very difficult to measure the individual orientation or conformation of a particle at the time of measurement. There are methods that can determine these parameters from the measured X-ray diffraction as part of the data analysis.

Analysis methods for single particle diffraction often search for information that is common to many different noisy diffraction measurements. For example, pattern-to-pattern correlations are used to classify and average similar diffraction patterns to improve signal to noise before determining orientations by the common arcs method [6]. Bayesian methods use likelihood measures to compare noisy data to a three-dimensional intensity model [7] or a manifold [8]. Graph-theoretic methods [9,10] and geodesic methods [11] map networks formed with Euclidean distances. By treating the dataset as a whole, Bayesian methods and graph-based methods are perhaps the most promising for treating the very low signals expected from individual protein molecules ($10^2$–$10^3$ photons). Although notably, combined correlation–classification/common-arc orientation methods have also produced good results in low-signal, high-resolution simulations and continue to be actively pursued [12,13]. Bayesian- and correlation-based methods have also been combined [14].

Although a diverse array of algorithms have been developed, experimental demonstrations are few and at very low resolution [11,15]. It is still unknown how these algorithms will perform with realistic conditions, very low signals...
Table 1. Averages of the structure factor using the statistical model.

<table>
<thead>
<tr>
<th>\langle F \rangle</th>
<th>\langle \vert F \vert^2 \rangle</th>
<th>\langle \vert F^* \rangle \vert F \rangle</th>
<th>\langle \vert F^* \rangle \vert F^* \rangle</th>
<th>\langle \vert F^* \rangle \rangle</th>
<th>\langle \vert F^* \rangle \vert F^* \rangle</th>
<th>\langle \vert F^* \rangle \rangle</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>2\langle F \rangle</td>
<td>0</td>
<td>6\langle F \rangle</td>
<td>0</td>
<td>24\langle F \rangle</td>
</tr>
</tbody>
</table>

2. Theoretical model of single particle diffraction

The intensity scattered by a particle is given by the formula

\[ I(q) = \phi r_e^2 |F(q)|^2 d\Omega, \]  

(2.1)

where \( \phi \) is the incident fluence, \( r_e \) is the classical electron radius, \( d\Omega \) is the solid angle and \( F(q) \) is the structure factor of the particle. To model the structure factor, consider the smallest parallelepiped that encloses the particle, defined by three vectors \( a, b, \) and \( c. \) The Fourier transform of the solid parallelepiped is

\[ S(q) = \text{sinc}(\pi a^* A) \text{sinc}(\pi b^* B) \text{sinc}(\pi c^* C), \]  

(2.2)

where \( a^*, b^*, \) and \( c^* \) are reciprocal vectors and \( A, B, C \) are the lengths of \( a, b, \) and \( c. \) An orthonormal basis for the Fourier transform of any object enclosed by the parallelepiped can be defined by functions \( S(q - q_{ad}), \) where \( q_{ad} = h a^* + k b^* + l c^*. \)

The function \( S(q) \) has the following useful properties

\[ S(q_{000}) = 1, \]  

(2.3)

\[ S(q_{000}) = 0 \quad \text{for values of } |hkl| \neq 0 \]  

(2.4)

and

\[ \sum_{hkl} S^*(q - q_{ad}) = 1, \quad \forall q. \quad \text{for integer } h, k, l. \]  

(2.5)

The first two relations are properties of the sinc functions in equation (2.2). The last property can be proved by first translating all the functions \( S(q - q_{ad}) \) by a constant vector \( q, \) to define a new orthonormal basis

\[ S'(q - q'_{ad}) = S(q - q_{ad} - q). \]  

(2.6)

The new basis can be chosen such that \( q = q'_{ad}. \) In the new basis, there is one term in the sum over \( hkl \) that takes the value 1 and, all other terms contribute 0 (using equations (2.3) and (2.4)). Therefore, equation (2.5) holds for all \( q \) values.

The structure factor of a single non-periodic particle can be written as

\[ F(q) = \sum_{hkl} F_{hkl} S(q - q_{ad}). \]  

(2.7)

The terms \( F_{hkl} \) are samples of the Fourier transform of the particle’s electron density. At high scattering angles, the terms \( F_{hkl} \) can be described statistically by assuming that all atoms are randomly located [16]. The real and imaginary parts of \( F_{hkl} \) are then drawn randomly from the following distribution

\[ p(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \]  

(2.8)

where \( (l) \) is the mean intensity at a pixel. Table 1 summarizes some useful quantities that can be calculated using this distribution. We assume that the variables \( F_{hkl} \) are statistically independent, such that

\[ \langle F_{hkl} F_{h'k'l'} \rangle \approx 0 \quad \text{if } h'k'l' \neq hkl. \]  

(2.9)

This is exact if the object fills the parallelepiped used to construct \( S(q - q_{ad}) \) and the atomic positions are really uncorrelated, otherwise, it is an approximation. By defining the parallelepiped to be the smallest that can enclose the object, the error from making this approximation is minimized.

To account for centrosymmetry, we need to exclude the case where \( h'k'l' = hkl \) from equation (2.9). This term contributes only in special cases, such as a rotation by \( \pi \) around the beam axis at low resolution. In order to present simple and concise derivations in what follows, we ignore the centrosymmetric contributions which are not difficult to include if required.

3. Evaluating correlations

The intensity measured by a pixel is denoted by \( I(q) \), and the intensity measured by the same pixel after the molecule has been rotated is denoted by \( I(q + q_0). \) A general three-dimensional rotation of the molecule can be specified by three Euler
angles. However, it is more convenient to label the displacement vector \( \mathbf{q} \), by the angular distance between the two \( \mathbf{q} \)-space points sampled by the pixel. For a general rotation, \( \mathbf{q} \), is not the same for all pixels and \( \alpha \) is not necessarily equal to any of the Euler angles, though for each pixel it can be calculated from them. The mean value of the correlation between \( I(\mathbf{q}) \) and \( I(\mathbf{q} + \mathbf{q}_a) \) is then given by

$$
\langle I(\mathbf{q}) \rangle = \sum_{n=0}^{\infty} \langle F_0^* F_n F_0 F_n \rangle \times S_n(\mathbf{q}_a) S_n(\mathbf{q} + \mathbf{q}_a).
$$

To simplify the notation, the three subscripts \( hkl \) have been replaced by a single subscript \( n \), such that \( S_n = S(\mathbf{q} - \mathbf{q}_{ahl}) \), and we have defined \( \langle I(\mathbf{q}) \rangle = \langle I(\mathbf{q}) \rangle \). A simplification can be made with a judicious change of basis via equation (2.6), such that \( \mathbf{q} = \mathbf{q}_a \). Then, \( m = n = 0 \) and equation (3.1) simplifies to

$$
\langle I(\mathbf{q}) \rangle = \sum_{m=0}^{\infty} \langle F_0^* F_m F_0 F_m \rangle \times S_m(\mathbf{q}_a) S_m(\mathbf{q}_a).
$$

In order to calculate the mean and standard deviation of the correlation, we can use the fact that many terms in equation (3.2) are statistically independent. If two random variables, \( A \) and \( B \), are statistically independent, then the following relation holds

$$
\langle AB \rangle = \langle A \rangle \langle B \rangle.
$$

Combining this relation with the fact that \( \langle F \rangle = 0 \), we find that there are only two sets of values of \( r \) and \( s \) for which \( \langle F_0^* F_r F_0 F_s \rangle \) is not zero. The first case occurs when \( r = s = 0 \) and contributes the following term

$$
\langle |F|^4 \rangle = \sum_{r=0}^{\infty} S_r^2(\mathbf{q}_a).
$$

The second case arises when \( r = s \neq 0 \) and contributes

$$
\langle |F|^4 \rangle = \sum_{r=0}^{\infty} S_r^2(\mathbf{q}_a) = 1 - S_0^2(\mathbf{q}_a).
$$

Using the results from table 1 and combining the two non-zero terms, we find that

$$
\langle I(\mathbf{q}) \rangle = \langle |F|^4 \rangle S_0^2(\mathbf{q}_a) + \langle |F|^2 \rangle^2 [1 - S_0^2(\mathbf{q}_a)]
$$

$$
= \langle |F|^2 \rangle^2 + \langle |F|^4 \rangle.
$$

The predicted angular-dependence of the correlation is approximately Gaussian for small angles \( \alpha \):

$$
S^2(\mathbf{q}_a) \propto e^{-4\pi^2 q^2 \alpha^2 k^2 / 3},
$$

where a resolution shell at a magnitude \( q \) has been considered, such that \( \mathbf{q}_a = q \alpha \). The width of the angular-dependence is close, but not exactly what is given by simulations reported in [17]. In that paper, the half-width at half maximum value was found to be 1/4R, whereas equation (3.7) predicts a narrower width of \( \sqrt{3/2} \pi R \). The discrepancy is around 20% and is due to the fact that the statistical independence of \( F_{hkl} \) (equation (2.9)) is only approximately true for the icosahedral virus particle used in the simulations for reference [17]. Nevertheless, the theory presented here does account for the most significant contribution to the angular-dependence of the correlation.

### 4. Derivation of the standard deviation

The standard deviation of the correlation \( \sigma_{nl} \) is calculated from

$$
\sigma_{nl}^2 = \langle F^2 \rangle - \langle I(\mathbf{q}) \rangle^2.
$$

We again use equation (2.6) to choose a set of basis functions such that \( \mathbf{q} = \mathbf{q}_a \). Therefore, we can write

$$
\langle F^2 \rangle = \sum_{m=0}^{\infty} \langle F_0^* F_m F_0 F_m \rangle \times S_m(\mathbf{q}_a) S_m(\mathbf{q}_a) S_m(\mathbf{q}_a).
$$

Using the results from table 1, it can be shown that only terms with \( m = n = s \) are not zero. The first case occurs when \( m = n = s = 0 \), whereas equation (3.7) predicts \( \sigma_{nl}^2 = 3 \langle |F|^2 \rangle^2 \). For the case of large \( \alpha \), we have \( \sigma_{nl}^2 = 3 \langle |F|^2 \rangle^2 \). Both these limits agree with those given by Huldt et al. [6].

### 5. Poisson noise

To determine the expected correlation when the measurement contains shot noise, we follow the method of Huldt et al. [6]. We define \( P(K, I) \) to be the probability of measuring a photon count \( K \) and having an expected intensity of \( I \). The probability \( P(K, I) \) can be written as

$$
P(K, I) = P(K|I)P(I),
$$

where \( P(K|I) \) is the conditional probability of measuring \( K \) when the expected intensity is known to be \( I \). The conditional probability is given by the Poisson distribution:

$$
P(K|I) = \frac{I^K e^{-I}}{K!}.
$$

This distribution can be used to show the following relations

$$
\sum_K KP(K|I) = I,
$$

and

$$
\sum_K K^2 P(K|I) = I^2 + 1.
$$
A correlation of the photon count recorded at a detector pixel is given by
\[
\langle KK_a \rangle = \sum_k \sum_{k_1} \int \int KK_a P(K) P(K_a) dI dI_a \\
= \langle I_a \rangle \\
= \langle I \rangle S^2(q_a) + \langle I \rangle.
\]
(5.5)

This shows that the mean correlation is not changed by the introduction of Poisson noise. The variance is calculated by a similar derivation and turns out to be
\[
\sigma^2_{KK_a} = \langle I \rangle^4 [15S^2(q_a) + 2S^2(q_a) + 3]
+ \langle I \rangle^2 [8S^2(q_a) + 4]
+ \langle I \rangle^2 [S^2(q_a) + 1].
\]
(5.6)

In the limit of large \(a\), this agrees with Huldt et al. [6]. However, in that paper, there was a mistake for the \(a = 0\) case. The correct limiting result is \(\sigma^2_{KK_a} = 20\langle I \rangle^4 + 12\langle I \rangle^3 + 2\langle I \rangle^2\).

6. Pattern-to-pattern correlations

The results presented so far refer to the correlation of two diffraction measurements at a point (i.e. a particular pixel on the detector). A correlation of two diffraction patterns involves taking the sum over all the pixels. For an arbitrary rotation of the molecule in three dimensions, the difference of the \(q\) coordinate of a pixel, \(q_{a,m}\), will be different for each pixel. The mean and standard deviation of the pattern-to-pattern correlation involves an integral over the point-to-point correlation (equation (3.6) with a distribution of values for \(q_{a,m}\) generated by the rotation. The intensity values on neighbouring pixels are not independent. As discussed in [6], the pattern-to-pattern correlation is calculated by scaling the mean pixel correlation by the number of independent variables needed to describe the intensity. The number of independent variables needed to describe the intensity is determined by sampling along the \(a^*, b^*, c^*\) directions at twice the rate as that used for the structure factor in equation (2.7).

A simple case occurs when the molecule only rotates around the beam axis, so that \(q_a\) is a function of resolution shell, but not the polar angular coordinate of the pixel. To provide a simple illustration of the results we have obtained, we show an example of this special case rather than show a general three-dimensional rotation. The number of independent sampling points in the resolution shell was set to 200, which corresponds to a resolution 32 times smaller than the particle, e.g., the defraction from a 9.6 nm particle at 0.3 nm resolution. The mean correlation as a function of angle is plotted in figure 1. The range of correlation values within three standard deviations of the mean is also shown, indicating how the distribution narrows as \(\alpha\) increases. As a consequence of this change of distribution, the mean correlation is not a perfect indicator of the most probable angle that gives rise to that correlation value. This can be seen in figure 2, which shows that the mean correlation of perfectly aligned particles, \(\langle I_{a=0} \rangle\), can arise from a range of orientations with close to equal probability. The most likely angle is around \(1/4\alpha R\). Values of the correlation higher than the mean value are needed to ensure selecting the aligned case with the greatest probability.

7. Conclusion

By addressing the issue of continuity in statistical models, we aim to improve their suitability for the analysis of single particle diffraction, which is inherently continuous. We view statistical models as potentially complimentary to numerical simulations in ongoing efforts to address the outstanding analysis issues for single particle imaging, particularly for low-signal, high-resolution applications to individual molecules. Although not all single-particle analysis algorithms are based on correlations, the methods we have presented could potentially be applied in future to Euclidean distances and likelihood functions to broaden the applicability of statistical models to Bayesian methods, graph-theoretic methods and manifold techniques.

Acknowledgements. This research was supported by the Australian Research Council through its Centres of Excellence programme.

Figure 1. The mean correlation of the intensity in a resolution shell as a function of relative molecular orientation for a rotation around the beam axis for the noise-free case. It applies to a resolution shell of \(q = 32/2\alpha\), where \(\alpha\) is the particle's radius. The mean correlation is shown as the solid line and the upper and lower bounds \(3\sigma\) from the mean are indicated by the dashed lines. The correlation has been normalized to 1 at \(\alpha = 0\). Its actual value in an experiment will be dependent on the number of incident photons and the composition of the sample.

Figure 2. The probability of obtaining particular correlation values as a function of relative particle orientation. Note that the height of the \(\langle I_{a>\delta} \rangle\) distribution has been scaled to improve visibility and is illustrative of the shape of the distribution only.
References


