FOUNDATIONS
ADVANCES

Volume 76 (2020)

Supporting information for article:

Effect of radiation damage and illumination variability on signal-to-noise ratio in XFEL single-particle imaging

Timur E. Gureyev, Alexander Kozlov, Andrew J. Morgan, Andrew V. Martin and Harry M. Quiney

S1. Histogram of a truncated 2D Gaussian distribution

Let us start from this definition of the histogram of a function defined in a finite-size domain Ω:

$$\lambda_f(s) = |\,\Omega\,|^{-1} \int_\Omega \delta[s - f(x,y)]dxdy\,, \tag{S1.1}$$

see, for example (Gureyev *et al*., 2020). Consider $G(x,y) = A\exp[-r_\perp^2/(2\sigma^2)]$, $r_\perp \equiv (x^2+y^2)^{1/2}$, and Ω being a circle with radius r_{max} in the 2D plane. Then

$$\lambda_G(s) = \frac{1}{\pi r_{\max}^2}\int_0^{2\pi}\int_0^{r_{\max}} \delta[s - G(r)]r_\perp dr_\perp d\varphi = \frac{2}{r_{\max}^2}\int_0^{r_{\max}} \delta\{s - A\exp[-r_\perp^2/(2\sigma^2)]\}r_\perp dr_\perp\,.$$

Taking $r_\perp^2/(2\sigma^2) = t,\ \ r_\perp dr_\perp = \sigma^2 dt$, we obtain

$$\lambda_G(s) = \frac{2\sigma^2}{r_{\max}^2}\int_0^{r_{\max}^2/(2\sigma^2)} \delta[s - A\exp(-t)]dt\,.$$

Now taking $A\exp(-t) = u,\ \ dt = -du/u$, we arrive at the following expression for the histogram of a 2D Gaussian restricted to a circle with radius r_{max}:

$$\lambda_G(s) = \frac{2\sigma^2}{r_{\max}^2}\int_{A\exp[-r_{\max}^2/(2\sigma^2)]}^{A} \delta(s-u)u^{-1}du = \frac{2\sigma^2}{r_{\max}^2}\frac{1}{s}, \tag{S1.2}$$

when $A\exp[-r_{\max}^2/(2\sigma^2)] \le s \le A$, and $\lambda_G(s) = 0$ everywhere else.

Let us check that the integral of this histogram function is equal to 1, as expected (Gureyev *et al.*, 2020), page 3).

$$\int_{A\exp[-r_{\max}^2/(2\sigma^2)]}^{A} \lambda_G(s)ds = \frac{2\sigma^2}{r_{\max}^2}\int_{A\exp[-r_{\max}^2/(2\sigma^2)]}^{A} s^{-1}ds = \frac{2\sigma^2}{r_{\max}^2}\ln\exp[r_{\max}^2/(2\sigma^2)] = 1.$$

Let us also check that the histogram $\lambda_f(s) = (2\sigma^2 / r_{\max}^2)s^{-1}$ satisfies equation (1) from (Gureyev *et al.,* 2020) for the expectation values, i.e. let us check that the expectation value of the histogram is equal to $\int s\lambda_G(s)ds = \bar{G} \equiv |\Omega|^{-1} \int_\Omega G(x,y)dxdy$. Calculating

$$\int\limits_{A\exp[-r_{\max}^2/(2\sigma^2)]}^{A} s\lambda_G(s)ds = (2\sigma^2 / r_{\max}^2) \int\limits_{A\exp[-r_{\max}^2/(2\sigma^2)]}^{A} 1ds = \frac{2A\sigma^2}{r_{\max}^2}\{1-\exp[-r_{\max}^2/(2\sigma^2)]\},$$

and

$$\frac{1}{\pi r_{\max}^2}\int_0^{2\pi}\int_0^{r_{\max}} A\exp\{-r_\perp^2/(2\sigma^2)\}]r_\perp dr_\perp d\varphi = \frac{2A\sigma^2}{r_{\max}^2}\int_0^{r_{\max}^2/(2\sigma^2)} \exp(-t)dt = \frac{2A\sigma^2}{r_{\max}^2}\{1-\exp[-r_{\max}^2/(2\sigma^2)]\},$$

we see that the latter equation is also satisfied, as expected.

Note also that the variance can be calculated as following:

$$\overline{(\delta G)^2} \equiv |\Omega|^{-1}\int_\Omega [G(x,y)-\bar{G}]^2 dxdy = \int s^2\lambda_G(s)ds - \left(\bar{G}\right)^2.$$

Since

$$\int\limits_{A\exp[-r_{\max}^2/(2\sigma^2)]}^{A} s^2\lambda_G(s)ds = (2\sigma^2 / r_{\max}^2) \int\limits_{A\exp[-r_{\max}^2/(2\sigma^2)]}^{A} sds = \frac{A^2\sigma^2}{r_{\max}^2}[1-\exp(-r_{\max}^2/\sigma^2)],$$

we obtain that

$$\overline{(\delta G)^2} = \frac{A^2\sigma^2}{r_{\max}^2}[1-\exp(-r_{\max}^2/\sigma^2)] - \frac{4A^2\sigma^4}{r_{\max}^4}\{1-\exp[-r_{\max}^2/(2\sigma^2)]\}^2.$$

When $r_{\max} = \sqrt{2}\sigma_{in}$, the latter expression becomes $\overline{(\delta G)^2} = A^2[-1/2 - 3/(2e^2) + 2/\mathrm{e}] \cong 0.033\,A^2$.

S2. Coherent mode decomposition

The 3D Fourier transform of equation (2) over **r** gives:

$$(\mathbf{F}_3\rho_e)(\mathbf{q},t) = \sum_j a_j(t) f_j(\mathbf{q}), \tag{S2.1}$$

with $f_j(\mathbf{q}) \equiv (\mathbf{F}_3\rho_j)(\mathbf{q}) \exp(-i2\pi\mathbf{q}\bullet\mathbf{R}_j)$. Consequently,

$$\mathcal{F}(\mathbf{q}) \cong \bar{\mathcal{F}}_{in}(r_e/R)^2 \sum_{jj'} D_{jj'} f_j(\mathbf{q}) f_{j'}^*(\mathbf{q}) + \bar{\mathcal{F}}_B(\mathbf{q}), \tag{S2.2}$$

where $D_{jj'} \equiv \bar{\mathcal{F}}_{in}^{-1} \int I_{in}(t) a_j(t) a_{j'}(t)\, dt$ and $\bar{\mathcal{F}}_{in} \equiv \int \bar{I}_{in}(t) dt$. Note that the incident intensity variability and radiation damage effects are contained in the stochastic "damage matrix" $(D_{jj'})$, while the atomic structure factors $f_j(\mathbf{q})$ are considered deterministic. In the case of identical pulses and no damage, we have $D_{jj'} = 1$ and $\sum_{jj'} D_{jj'} f_j(\mathbf{q}) f_{j'}^*(\mathbf{q}) = |\sum_j f_j(\mathbf{q})|^2 = |(\mathbf{F}_3\rho_{e,o})(\mathbf{q})|^2$, where $\rho_{e,0}(\mathbf{r}) = \sum_j \rho_j(\mathbf{r} - \mathbf{R}_j)$ is the undamaged electron density, which represents the quantity that we would like to reconstruct from the experimentally measured diffraction intensities.

The diffracted intensity in equation (S2.2) is a statistical function of repeated measurements. The mean value over the ensemble is equal to

$$\bar{\mathcal{F}}(\mathbf{q}) \cong \bar{\mathcal{F}}_{in}(r_e/R)^2 \sum_{jj'} \bar{D}_{jj'} f_j(\mathbf{q}) f_{j'}^*(\mathbf{q}) + \bar{\mathcal{F}}_B(\mathbf{q}), \tag{S2.3}$$

where $\bar{D}_{jj'} = \bar{\mathcal{F}}_{in}^{-1} \int_0^T \overline{I_{in}(t) a_j(t) a_{j'}(t)}\, dt$. In equation (S2.3), the sum over all atoms in the sample, indexed by *j*, can be split into the sum of atoms with different atomic numbers, indexed by *Z*, with all atoms with the same atomic number *Z* indexed by m_Z. The indexes *j* can be uniquely identified with

pairs of indexes (Z, m_Z) according to $j(Z, m_Z) = \sum_{Z'=1}^{Z-1} M_{Z'} + m_Z$, $Z = 1, 2, ..., Z_{max}$ and $m_Z = 1, 2, ..., M_Z$, where $Z_{\max}$ is the number of different species of atoms and M_Z is the number of atoms of species *Z* present in the sample. We assume that the average radiation damage is the same for all atoms of the same species in the sample, meaning that the values of $\bar{D}_{j(Z,m_Z)j(Z',m_{Z'})}$ are the same for any values of m_Z and $m_{Z'}$, when *Z* and *Z′* are fixed. Therefore,

$\sum_{jj'} \bar{D}_{jj'} f_j(\mathbf{q}) f_{j'}^*(\mathbf{q}) = \sum_{ZZ'} \tilde{D}_{ZZ'} \tilde{f}_Z(\mathbf{q}) \tilde{f}_{Z'}^*(\mathbf{q})$, where $\tilde{f}_Z(\mathbf{q}) \equiv \sum_{m_Z=1}^{M_Z} f_{j(Z,m_Z)}(\mathbf{q})$ and $\tilde{D}_{ZZ'} \equiv \bar{D}_{j(Z,1)j(Z',1)}$, and we can re-write equation (S2.3) as

$$\bar{\mathcal{F}}(\mathbf{q}) \cong \bar{\mathcal{F}}_{in} (r_e / R)^2 \sum_{ZZ'} \tilde{D}_{ZZ'} \tilde{f}_Z(\mathbf{q}) \tilde{f}_{Z'}^*(\mathbf{q}) + \bar{\mathcal{F}}_B(\mathbf{q}) . \qquad \text{(S2.4)}$$

Note that in the transition from equation (S2.3) to equation (S2.4) we have managed to drastically reduce the dimensionality of the mean damage matrix: while the size of the matrix $(\bar{D}_{jj'})$ can be of the order of $10^5 \times 10^5$ or larger for protein molecules, the matrix $(\tilde{D}_{ZZ'})$ is typically less than 10×10 in size. This model can be further generalized by allowing different indexes for different electron orbitals, that is replacing each index *Z* by multiple pairs (Z, γ), where *Z* is fixed and γ denotes different orbitals of an atom with charge *Z* (Quiney & Nugent, 2011; Curwood *et al.*, 2013). The main conclusions obtained in this paper would remain valid in such a generalized case, however, the mathematical treatment would be somewhat more complicated. In particular, the electron density distributions corresponding to different orbital configurations (neutral and ionised) are not orthogonal in the L_2 sense, in contrast to the orthogonality of the components $\tilde{f}_Z(\mathbf{q})$ in equation (S2.4):

$\int \tilde{f}_Z(\mathbf{q}) \tilde{f}_{Z'}^*(\mathbf{q}) d\mathbf{q} = \int \sum_{m_Z} \rho_Z(\mathbf{r} - \mathbf{R}_{j(Z,m_Z)}) \sum_{m_{Z'}} \rho_{Z'}(\mathbf{r} - \mathbf{R}_{j(Z',m_{Z'})}) \, d\mathbf{r} = \delta_{ZZ'} b_Z^2$,

$b_Z^2 \equiv \int |\, \tilde{f}_Z(\mathbf{q}) \,|^2 \, d\mathbf{q} = M_Z \int \rho_Z^2(\mathbf{r}) d\mathbf{r}$. Here we have used the Parseval's theorem and the assumption that the electron density distributions of different atoms in the sample do not overlap, with $\rho_Z(\mathbf{r})$ being the electron density in a single atom with atomic number *Z*. Consider the eigenvector decomposition of the symmetric matrix $\mathbf{G} \equiv \left(\bar{\mathcal{F}}_{in} (r_e / R)^2 b_Z b_{Z'} \tilde{D}_{ZZ'} \right)$: $G_{ZZ'} = \sum_k w_k g_{kZ} g_{kZ'}$, where w_k are the eigenvalues and $\mathbf{g}_k = (g_{k1}, g_{k2}, ..., g_{kZ_{\max}})$ are the orthonormal eigenvectors of **G**,

$\sum_Z g_{kZ} g_{k'Z} = \delta_{kk'}$. When matrix **G** is degenerate, some of the eigenvalues w_k will be zero. For example, it is straightforward to verify that in the "uniform damage" case $a_Z(t) = a(t)$ (which includes the "undamaged" case $a_Z(t) = 1$), the only non-zero eigenvalue is $w_1 = (r_e / R)^2 (\sum_Z b_Z^2) \int_0^T \overline{I_{in}(t) a(t) a(t)}\, dt$, which corresponds to the total coherently scattered fluence, and the corresponding eigenvector has components $g_{1z} = b_Z / (\sum_Z b_Z^2)^{-1/2}$. In the general case, substituting the eigenvector decomposition $\bar{\mathcal{F}}_{in} (r_e / R)^2 b_Z b_{Z'} \tilde{D}_{ZZ'} = \sum_k w_k g_{kZ} g_{kZ'}$ into equation (S2.4), we obtain the following equation, which provides a representation of the mean diffracted fluence in terms of coherent modes and their mean "occupational numbers" (Mandel & Wolf, 1995; Quiney, 2010; Quiney & Nugent, 2011; Lorenz *et al.*, 2012):

$$\bar{\mathcal{F}}(\mathbf{q}) = \sum_k \bar{\mathcal{F}}_k(\mathbf{q}) + \bar{\mathcal{F}}_B(\mathbf{q}) = \sum_k w_k |\psi_k(\mathbf{q})|^2 + \bar{\mathcal{F}}_B(\mathbf{q}), \quad \psi_k(\mathbf{q}) \equiv \sum_Z g_{kZ} b_Z^{-1} \tilde{f}_Z(\mathbf{q}). \tag{S2.5}$$

The functions $\psi_k(\mathbf{q})$ in equation (S2.5) are orthonormal in the sense that $\int \psi_k(\mathbf{q}) \psi^*_{k'}(\mathbf{q}) d\mathbf{q} = \sum_Z \sum_{Z'} g_{kZ} g_{k'Z'} \int b_Z^{-1} \tilde{f}_Z(\mathbf{q}) b_{Z'}^{-1} \tilde{f}_{Z'}(\mathbf{q}) d\mathbf{q} = \sum_Z \sum_{Z'} g_{kZ} g_{k'Z'} \delta_{ZZ'} = \delta_{kk'}$, where we have used the orthonormality of functions $b_Z^{-1} \tilde{f}_Z(\mathbf{q})$ and vectors $\mathbf{g}_k$. According to equation (S2.5), if the diffracted fluence distributions are collected for a sufficiently broad range of rotational positions of the sample, then the whole 3D reciprocal space, except for a vicinity of the point (0,0,0), of the sample's electron density can be probed (Sayre *et al.*, 1998; Robinson *et al.*, 2001; Marchesini *et al.*, 2003; Chapman *et al.*, 2006). Furthermore, if the coherent modes can be identified and the generalised phase retrieval problem can be solved, in the sense that unique complex functions $\psi_k(\mathbf{q})$, can be found from the measured diffraction intensities on the basis of appropriate assumptions about the sample and the imaging setup, then the undamaged electron density distribution in the sample can be obtained as (Quiney, 2010; Curwood *et al.*, 2013):

$$\bar{\rho}_{e,0}(\mathbf{r}) \equiv \sum_Z \sum_{m_Z} \rho_Z(\mathbf{r} - R_{Zm_Z}) = \sum_Z (\mathbf{F}_3^{-1} \tilde{f}_Z)(\mathbf{r}) = \sum_k \gamma_k (\mathbf{F}_3^{-1} \psi_k)(\mathbf{r}), \tag{S2.6}$$

where $\gamma_k \equiv \sum_Z b_Z g_{kZ}$ and the last expression in equation (S2.6) was obtained using explicit solution $\tilde{f}_Z(\mathbf{q}) = b_Z \sum_k g_{kZ}\psi_k(\mathbf{q})$ of the system of linear equations $\psi_k(\mathbf{q}) = \sum_Z g_{kZ} b_Z^{-1}\tilde{f}_Z(\mathbf{q})$ with orthonormal vectors $\mathbf{g}_k$. In the case of uniform damage, including the case with no radiation damage and no variability of the incident intensity, equation (S2.6) becomes trivial: $\overline{\rho}_{e,0}(\mathbf{r}) = (\sum_Z b_Z^2)^{1/2}(\mathbf{F}_3^{-1}\psi_1)(\mathbf{r})$, with $\psi_1(\mathbf{q}) = (\sum_Z b_Z^2)^{-1/2}\sum_Z \tilde{f}_Z(\mathbf{q})$.

In the "generalized single-mode case" considered in Section 6, equation (24) represents a generalization of the single-mode version of equation (S2.5). Indeed, in the latter case, we have $(\mathbf{F}_3\overline{\rho}_{e,0})(\mathbf{q}) = (\sum_Z b_Z g_{1Z})\psi_1(\mathbf{q})$ and hence $\overline{\mathcal{F}}(\mathbf{q}) - \overline{\mathcal{F}}_B(\mathbf{q}) = w_1 |\psi_1(\mathbf{q})|^2 = w_1(\sum_Z b_Z g_{1Z})^{-2} |(\mathbf{F}_3\overline{\rho}_{e,0})(\mathbf{q})|^2$. Therefore, the "damage coefficient" $D^{-1}(\mathbf{q})$ from equation (24) is equal here to a constant factor, $\overline{\mathcal{F}}_{in}^{-1}(r_e/R)^{-2}(\sum_Z b_Z g_{1Z})^{-2}w_1$. In general, the factor $D^{-1}(\mathbf{q})$ in equation (24) may depend on **q**. An analogue of equation (S2.6) here is

$$\overline{\rho}_{e,0}(\mathbf{r}) = \overline{\mathcal{F}}_{in}^{-1/2}(R/r_e)\, b\, (\mathbf{F}_3^{-1}\psi_1)(\mathbf{r}), \qquad \text{(S2.7)}$$

with $\psi_1(\mathbf{q}) = b^{-1}D^{1/2}(\mathbf{q})\overline{U}_S(\mathbf{q})$ and $b \equiv (\int |D^{1/2}(\mathbf{q})\overline{U}_S(\mathbf{q})|^2\, d\mathbf{q})^{1/2}$, where $\overline{U}_S(\mathbf{q})$ is the complex amplitude obtained by phase retrieval from the "structural" fluence $\overline{\mathcal{F}}_S(\mathbf{q}) \equiv \overline{\mathcal{F}}(\mathbf{q}) - \overline{\mathcal{F}}_B(\mathbf{q})$.

S3. Stochastic behaviour of the modes obtained from experimental data

Firstly, we describe the stochastic behaviour of the diffracted fluence incident on the detector. For this purpose, we introduce "structural" fluence $\mathcal{F}_S(\mathbf{q}) \equiv \mathcal{F}(\mathbf{q}) - \mathcal{F}_B(\mathbf{q}) = U_S(\mathbf{q})U_S^*(\mathbf{q})$, where $U_S(\mathbf{q})$ is a stochastic complex amplitude function (analytic signal) corresponding to $\mathcal{F}_S(\mathbf{q})$. Karhunen-Loeve expansion (Goodman, 1985; Mandel & Wolf, 1995) of the stochastic distribution $U_S(\mathbf{q})$ can be written as $U_S(\mathbf{q}) = \sum_k \alpha_k \chi_k(\mathbf{q})$, with orthonormal deterministic basis functions $\chi_k(\mathbf{q})$, such that $\int \chi_k(\mathbf{q})\chi_{k'}^*(\mathbf{q})d\mathbf{q} = \delta_{kk'}$, and uncorrelated random coefficients α_k, such that $\overline{\alpha_k\alpha_{k'}^*} = \delta_{kk'}\overline{|\alpha_k|^2}$.

According to the construction of the Karhunen-Loeve expansion, $\chi_k(\mathbf{q})$ are the eigenvectors and $\overline{|\alpha_k|^2}$ are the eigenvalues of the integral equation $\int J_S(\mathbf{q},\mathbf{q}')\chi_k(\mathbf{q}')d\mathbf{q}' = \overline{|\alpha_k|^2}\chi_k(\mathbf{q})$, with the kernel equal to the autocorrelation function $J_S(\mathbf{q},\mathbf{q}') \equiv \overline{U_S(\mathbf{q})U_S^*(\mathbf{q}')}$. However, according to the procedure that led to equation (7) above, we can show also that $\int J_S(\mathbf{q},\mathbf{q}')\psi_k(\mathbf{q}')d\mathbf{q}' = w_k\psi_k(\mathbf{q})$. Therefore, from the uniqueness of eigenvalues and eigenfunctions of a Hilbert-Schmidt operator with kernel $J_S(\mathbf{q},\mathbf{q}')$, we can conclude that $\chi_k(\mathbf{q}) = \psi_k(\mathbf{q})$ and $\overline{|\alpha_k|^2} = w_k$. Therefore, a theoretical stochastic form of equation (3) (and equation (S2.5)) is given by equation (5).

In order to understand the stochastic behaviour of the coherent modes recovered from experimental data, we need to consider the way in which the practical recovery of these modes from the experimentally measured data can proceed. The coherent modes can be recovered in practice from the numbers of diffracted photons measured in the finite number of exposures taken during one CDI experiment. The precise values of the ensemble-average quantities are not accessible under these conditions. In the process of generalised phase retrieval (Curwood *et al.*, 2013), that is in the recovery of the coherent mode intensities and their phases, firstly, the mean background fluence can be evaluated by fitting a suitable smooth distribution into the estimated background. After that, the background fluence can be subtracted from the total measured fluence, producing an estimation of the "structural" photon fluence $n_S(\mathbf{q}) \equiv n(\mathbf{q}) - \overline{n}_B(\mathbf{q}) = \overline{n}(\mathbf{q}) - \overline{n}_B(\mathbf{q}) + \delta n(\mathbf{q})$. The structural fluence obtained in this way can then be used to recover the coherent mode fluences in the form $n_k(\mathbf{q}) = \eta_k(\mathbf{q})n_S(\mathbf{q})$, where the coefficients $\eta_k(\mathbf{q})$ are treated as deterministic and $\sum_k \eta_k(\mathbf{q}) = 1$ (Curwood *et al.*, 2013). As a result of this process, the coherent modes are recovered in a form of stochastic distributions that have the means $\overline{n}_k(\mathbf{q}) = \eta_k(\mathbf{q})\overline{n}_S(\mathbf{q})$, and which are completely correlated with each other, $\overline{\delta n_k(\mathbf{q})\delta n_{k'}(\mathbf{q})} = \eta_k(\mathbf{q})\eta_{k'}(\mathbf{q})\overline{[\delta n_S(\mathbf{q})]^2} = \left(\overline{[\delta n_k(\mathbf{q})]^2}\right)^{1/2}\left(\overline{[\delta n_{k'}(\mathbf{q})]^2}\right)^{1/2}$. The latter fact follows from the reconstruction procedure, where the random variation in each recovered mode is a deterministic function, $\delta n_k(\mathbf{q}) = \eta_k(\mathbf{q})\delta n(\mathbf{q})$, of the single measured random distribution $\delta n_S(\mathbf{q}) = \delta n(\mathbf{q})$. This "idealized" reconstruction process leads to the minimal possible variances in the recovered coherent modes satisfying the constraint $\sum_k n_k(\mathbf{q}) = n_S(\mathbf{q})$:

$$\overline{[\delta n_k(\mathbf{q})]^2} = \eta_k^2(\mathbf{q})\overline{[\delta n(\mathbf{q})]^2} = \eta_k^2(\mathbf{q})\overline{n}(\mathbf{q})[1+\sigma_F^2(\mathbf{q})], \tag{S3.1}$$

where we have used equation (4) for the variance of the measured photon fluence. As a consequence, $\sum_k \overline{[\delta n_k(\mathbf{q})]^2} + \sum_{k\neq k'} \overline{\delta n_k(\mathbf{q})\delta n_{k'}(\mathbf{q})} = \overline{[\delta n_S(\mathbf{q})]^2} = \overline{[\delta n(\mathbf{q})]^2}$. Consider for example the case of two modes having equal occupational numbers, such that $n_S(\mathbf{q}) = n_1(\mathbf{q}) + n_2(\mathbf{q})$ and $\eta_k(\mathbf{q}) = \overline{n}_k(\mathbf{q}) / \overline{n}_S(\mathbf{q}) = 1/2$, $k = 1, 2$. Then the variance of each of the two modes recovered with a deterministic process described above will be equal to $\overline{[\delta n_S(\mathbf{q})]^2} / 4$. This variance is twice as small as in the case of two independent Poisson distributions $n_1(\mathbf{q})$ and $n_2(\mathbf{q})$, such that $n_1(\mathbf{q}) + n_2(\mathbf{q}) = n_S(\mathbf{q})$, each containing half of the total number of photons on average, $\overline{n}_1(\mathbf{q}) = \overline{n}_2(\mathbf{q}) = (1/2)\overline{n}_S(\mathbf{q})$. Indeed, in the latter case one has $\overline{[\delta n_1(\mathbf{q})]^2} = \overline{[\delta n_2(\mathbf{q})]^2} = (1/2)\overline{n}_S(\mathbf{q}) = (1/2)\overline{[\delta n_S(\mathbf{q})]^2}$. Note that the minimization of the variances of the recovered modes is a consequence of our simplifying assumption that the generalized phase retrieval is a deterministic process. In reality, both the absolute values and the phases of the recovered coherent modes will contain some additional variances due to the uncertainty of the generalised phase retrieval process. Therefore, the SNR in the reconstructed electron density obtained in section 4 provides only an upper bound for the SNR achievable in practice. Note also that the incoherent scattering both reduces the "structural" photon fluence by means of subtraction of the smooth background, $n_S(\mathbf{q}) = n(\mathbf{q}) - \overline{n}_B(\mathbf{q})$, and contributes to the noise variance in the "structural" fluence via the variability in the diffracted fluence, $\overline{[\delta n_S(\mathbf{q})]^2} = \overline{n}(\mathbf{q}) + \overline{(W - \overline{W})^2}(\mathbf{q})$. The influence of these effects on the SNR in the reconstructed electron density is considered in the next section.

S4. Estimation of variance of the reconstructed electron density

Using the Parseval's theorem, equation (19) and the orthogonality of amplitudes $\tilde{U}_k(\mathbf{q})$, we can write

$$\int < \rho_{e,0}^2(\mathbf{r}) > d\mathbf{r} = < \int |(\mathbf{F}_3 \rho_{e,0})(\mathbf{q})|^2 \, d\mathbf{q} > = < \int \sum_k d_k^2 |\tilde{U}_k(\mathbf{q})|^2 d\mathbf{q} > = \int \sum_k d_k^2 \eta_k(\mathbf{q}) \overline{n}_S(\mathbf{q}) \, d\mathbf{q}$$
$$\cong \sum_{m=1}^{M} \sum_k d_k^2 \eta_k(\mathbf{q}_m) \, \overline{n}_S(\mathbf{q}_m) \, |\Delta\mathbf{q}_m| = \sum_{m=1}^{M} d(\mathbf{q}_m) \overline{n}_S(\mathbf{q}_m) \, |\Delta\mathbf{q}_m|, \quad \text{(S4.1)}$$

where $d(\mathbf{q}_m) \equiv \sum_k d_k^2 \eta_k(\mathbf{q}_m)$, the integral over the scanned area of reciprocal space has been approximated by a sum over discrete voxels with volume $|\Delta\mathbf{q}_m|$ centred at the measurement points

$\mathbf{q}_m$ and where angular brackets denote the ensemble average (we will use such notation interchangeably with the overhead bar, whichever is more convenient in a particular case).

Estimation of the quantity $\int < \rho_{e,0}(\mathbf{r}) >^2 d\mathbf{r} = \int |< (\mathbf{F}_3\rho_{e,0})(\mathbf{q}) >|^2 d\mathbf{q} = \int \sum_k d_k^2 |< \tilde{U}_k(\mathbf{q}) >|^2 d\mathbf{q}$ appearing in the numerator and denominator of equation (20), is more complicated technically than the evaluation of the integral $\int < \rho_{e,0}^2(\mathbf{r}) > d\mathbf{r}$ above, because the former requires an evaluation of statistical properties of the random variable $n_k^{1/2}(\mathbf{q}_m)$. The corresponding analysis is presented below.

Consider the square modulus of the ensemble average of a complex random variable $U = |U| \exp(i\varphi)$, $|< U >|^2 = |J^{-1}\sum_{j=1}^{J} |U_j| \exp(i\varphi_j)|^2$. Applying the Cauchy-Schwarz inequality with the factors $|U_j|^{1/2}$ and $|U_j|^{1/2} \exp(i\varphi_j)$ to the last expression, we obtain that $|< U >|^2 \le J^{-2}\sum_{j=1}^{J} |U_j| \sum_{j=1}^{J} |U_j| \, |\exp(i\varphi_j)|^2 = <|U|>^2$. Using the last inequality for $\tilde{U}_k(\mathbf{q}) \equiv N_k^{1/2}\psi_k(\mathbf{q})$ and equation (19), we arrive at the following result:

$$\int |< (\mathbf{F}_3\rho_{e,0})(\mathbf{q}) >|^2 d\mathbf{q} = \int \sum_k d_k^2 |< \tilde{U}_k(\mathbf{q}) >|^2 d\mathbf{q} \le \sum_m \sum_k d_k^2 < n_k^{1/2}(\mathbf{q}_m) >^2 |\Delta\mathbf{q}_m| . \qquad \text{(S4.2)}$$

It is easy to see that if the phases φ_j are the same for all j, then $|< U >|^2 = <|U|>^2$; in this case the last inequality for the integral of the square of the mean electron density also becomes an equality.

In view of equation (S4.2), we need to estimate the square of the mean of the random variable $n_k^{1/2}(\mathbf{q}_m)$. The following approximation, which is valid when $\overline{n}_k(\mathbf{q}_m) \ge 1/2$, can be obtained as in (Gureyev *et al*., 2018, Mathematics StackExchange, 2018): $\overline{n_k^{1/2}} \cong (\overline{n}_k)^{1/2} - (1/8)(\overline{n}_k)^{-3/2}\overline{(\delta n_k)^2}$, where we omitted the argument $\mathbf{q}_m$ for brevity. According to equation (S3.1), $\overline{(\delta n_k)^2} = \eta_k^2 \overline{n}(1 + \sigma_F^2)$. Therefore,

$$\overline{n_k^{1/2}} \cong \eta_h^{1/2}[(\overline{n}_S)^{1/2} - (1/8)(\overline{n}_S)^{-3/2}\overline{n}(1+\sigma_F^2)]. \qquad \text{(S4.3)}$$

Assuming that $\overline{n}_S > (\overline{n}/\overline{n}_S)^2(1+\sigma_F^2)^2$, we can approximate the square of equation (S4.3) as

$$\left(\overline{n_k^{1/2}}\right)^2 \cong \eta_k[\overline{n}_S - (1/4)(\overline{n}/\overline{n}_S)(1+\sigma_F^2)]. \qquad \text{(S4.4)}$$

Substituting this into equation (S4.2) we obtain

$$\int < \rho_{e,0}(\mathbf{r}) >^2 d\mathbf{r} \le \sum_{m=1}^{M} d(\mathbf{q}_m)\{\overline{n}_S(\mathbf{q}_m) - (1/4)[1+p_B(\mathbf{q}_m)][1+\sigma_F^2(\mathbf{q}_m)]\} \mid \Delta\mathbf{q}_m \mid, \qquad \text{(S4.5)}$$

with $p_B(\mathbf{q}_m) \equiv \overline{n}_B(\mathbf{q}_m)/\overline{n}_S(\mathbf{q}_m) = [\overline{n}(\mathbf{q}_m)/\overline{n}_S(\mathbf{q}_m)] - 1$. The following inequality for the variance of the reconstructed electron density follows from equations (S4.1) and (S4.5):

$$\int [< \rho_{e,0}^2(\mathbf{r}) > - < \rho_{e,0}(\mathbf{r}) >^2] d\mathbf{r} \ge (1/4)\sum_{m=1}^{M} d(\mathbf{q}_m)[1+p_B(\mathbf{q}_m)][1+\sigma_F^2(\mathbf{q}_m)]\} \mid \Delta\mathbf{q}_m \mid, \qquad \text{(S4.6)}$$

provided that $\overline{n}_S(\mathbf{q}_m) > [1+p_B(\mathbf{q}_m)]^2[1+\sigma_F^2(\mathbf{q}_m)]^2$ for almost all data points, where $\sigma_F^2(\mathbf{q})$ is the relative variance of the exposure-time-integrated intensity incident on the detector, as introduced in equation (4), and $p_B(\mathbf{q}_m) = \overline{n}_B(\mathbf{q}_m)/\overline{n}_S(\mathbf{q}_m)$ is the ratio of the mean background and the structural photon fluences.

S5. Atomic form factor model and variance of diffracted intensity in the single-mode model case

We model the electron density in an atom, averaged over an XFEL pulse, by the simplest 1s type Slater orbital (Slater, 1930), $\rho_a(\mathbf{r}) \equiv \alpha \exp(-2 \mid \mathbf{r} \mid / r_0)$, where r_0 plays the role of the "radius" of the atom and α is a dimensionless normalization constant which is determined below. We assume all of the electrons to reside in this single orbital. This model, while being rather inaccurate for chemical or

spectroscopic description of an atom beyond hydrogen and helium, captures the essential physics of the electron density dynamics.

The form factor by definition is the 3D Fourier transform of the electron density distribution, $f(\mathbf{q}) \equiv (\mathbf{F}_3\rho_a)(\mathbf{q})$. It is convenient to calculate the Fourier transform in spherical coordinates: $\mathbf{q} = (q\cos\theta'\sin\phi', q\cos\theta'\cos\phi', q\sin\theta')$, $\mathbf{r} = (r\cos\theta\sin\phi, r\cos\theta\cos\phi, r\sin\theta)$, so that

$$f(\mathbf{q}) = \alpha \int_{-\pi/2}^{\pi/2} d\theta \int_0^{2\pi} d\phi \int_0^{\infty} dr\, r^2 \cos\theta \times \exp\{-i2\pi qr[\cos\theta\cos\theta'\cos(\phi-\phi') + \sin\theta\sin\theta'] - 2r/r_0\}. \quad \text{(S5.1)}$$

Because the Fourier transform of a spherically symmetric function is also a spherically-symmetric function, $f(\mathbf{q})$ is independent of θ' and ϕ'. In particular, we can select $\theta' = \pi/2$ in equation (S5.1), resulting in

$$f(\mathbf{q}) = \alpha \int_{-\pi/2}^{\pi/2} d\theta \int_0^{2\pi} d\phi \int_0^{\infty} dr\, r^2 \cos\theta \exp[-i2\pi qr\sin\theta - 2r/r_0]. \quad \text{(S5.2)}$$

Evaluating the integral over ϕ and making the substitution $s = \sin\theta$, we obtain

$$f(\mathbf{q}) = 2\alpha\pi \int_{-1}^{1} ds \int_0^{\infty} dr\, r^2 \exp[-i2\pi qrs - 2r/r_0]. \quad \text{(S5.3)}$$

The integral over *r* can now be calculated by using the integration by parts formula twice:

$$f(\mathbf{q}) = (\alpha\pi/2) \int_{-1}^{1} ds\, (i\pi qs + 1/r_0)^{-3}. \quad \text{(S5.4)}$$

Calculating the remaining integral in equation (S5.4) is straightforward:

$$f(\mathbf{q}) = \alpha\pi r_0^3 / (1+\pi^2 r_0^2 q^2)^2 . \qquad \text{(S5.5)}$$

The normalization of $f(\mathbf{q})$ requires that $f(0) = \iiint \rho_a(\mathbf{r})d\mathbf{r} = Z$, where Z is the number of electrons in the atom. Therefore, $\alpha = Z/(\pi r_0^3)$ and

$$f(\mathbf{q}) = Z / (1+\pi^2 r_0^2 q^2)^2 . \qquad \text{(S5.6)}$$

Now we can calculate the variance of the diffracted intensity appearing in the denominator of equation (26) in the presence of radiation damage described by equation (25), i.e. we evaluate the expression $\beta^{-2}\sum_{m=1}^{M}\frac{(1+\pi^2 r_D^2 q_m^2)^4}{(1+\pi^2 r_0^2 q_m^2)^4}[1+p_B(\mathbf{q}_m)][1+\sigma_F^2(\mathbf{q}_m)]\,|\,\Delta\mathbf{q}_m\,|$. In order to simplify this expression, we assume that the changes in the atomic radius are small, i.e. $\pi^2(r_0^2 - r_D^2)q_m^2 << 1$. In this case we can approximate $\left(\frac{1+\pi^2 r_D^2 q_m^2}{1+\pi^2 r_0^2 q_m^2}\right)^4 = \left(1-\frac{\pi^2(r_0^2-r_D^2)q_m^2}{1+\pi^2 r_0^2 q_m^2}\right)^4 \cong 1-4\pi^2\Delta_D^2 q_m^2$, where $\Delta_D^2 \equiv r_0^2 - r_D^2$. Assuming further that the relative variance of the diffracted intensity (describing the deviation from the Poisson statistics due to the intensity variation of the incident beam) and the background intensity are both slowly varying as functions of the reciprocal space coordinates, we can approximate them by the corresponding mean values $\overline{\sigma}_F^2 \equiv \int_\Omega \sigma_F^2(\mathbf{q})d\mathbf{q} / \int_\Omega d\mathbf{q}$ and $\overline{p}_B \equiv \int_\Omega p_B(\mathbf{q})d\mathbf{q} / \int_\Omega d\mathbf{q}$, where Ω is the area of the reciprocal space covered by the experiment, and write $\sum_{m=1}^{M}(1-4\pi^2\Delta_D^2 q_m^2)[1+p_B(\mathbf{q}_m)[1+\sigma_F^2(\mathbf{q}_m)]\,|\,\Delta\mathbf{q}_m\,| \cong (1+\overline{p}_B)(1+\overline{\sigma}_F^2)C_r$, where $C_r \equiv \sum_{m=1}^{M}(1-4\pi^2\Delta_D^2 q_m^2)\,|\,\Delta\mathbf{q}_m\,|$. In the case of cylindrical sampling of the reciprocal space we have $|\,\Delta\mathbf{q}_m\,| = s_l \Delta s\,\Delta\phi\,\Delta\eta = lh^3(\pi / L_\phi)$, where $s_l = (l-1)h$, $\Delta s = \Delta\eta = h$, $\Delta\phi = \pi / L_\phi$, $m = (l, m_\eta, m_\phi)$, $l = 1,2,...,L/2$, $m_\phi = 1,2,...,L_\phi$, $m_\eta = 1,2,...,L$, h is the detector pixel size, L_ϕ is the number of rotational positions (i.e. the number of projections) in the experiment, L^2 is the number of pixels in the detector, so that $L_\phi L^2/2 = M$ is the total number of data points in the experiment. It is known (Natterer, 1986) that the optimal (Shannon) sampling is achieved when $L_\phi = \pi L/2$, in

which case $M = L_\phi L^2 / 2 = (\pi / 4)L^3 \cong L^3$. The corresponding number of diffraction patterns, $(\pi / 2)L$, also agrees well with the estimate given for the XFEL-type CDI case in (Ekeberg *et al.*, 2015). Assuming the optimal sampling conditions, we have $\Delta\phi = 2 / L$ and $|\Delta\mathbf{q}_m| = 2(l / L)h^3$. As a consequence, $C_r = (2 / L)h^3 \sum_{l=1}^{L/2} \sum_{m_\eta=1}^{L} \sum_{m_\phi=1}^{L_\phi} l[1 - 4\pi^2 h^2 \Delta_D^2 (l^2 + m_\eta^2)]$. Using the Faulhaber's formulae (Wikipedia, 2018): $\sum_{l=1}^{n} l = n(n+1) / 2 \cong n^2 / 2$, $\sum_{l=1}^{n} l^2 = (1 / 6)n(n+1)(2n+1) \cong n^3 / 3$ and $\sum_{l=1}^{n} l^3 = (1 / 4)n^2 (n+1)^2 \cong n^4 / 4$, when $n >> 1$, we obtain

$$C_r = \pi h^3 [L^3 / 8 - \pi^2 \Delta_D^2 h^2 L^5 (1/16 + 1/6)] \cong \pi h^3 (L^3 / 8)[1 - 2\pi^2 \Delta_D^2 h^2 L^2] = (1 - 8\pi^2 \Delta_D^2 q_{\max}^2) \sum_{m=1}^{M} |\Delta\mathbf{q}_m|,$$

where $q_{\max} = hL / 2$. Combining the above results, we arrive at equation

$$\sum_{m=1}^{M} D(\mathbf{q}_m)[1 + p_B(\mathbf{q}_m)][1 + \sigma_F^2(\mathbf{q}_m)]|\Delta\mathbf{q}_m| \cong \beta^{-2}(1 + \overline{p}_B)(1 + \overline{\sigma}_F^2)(1 - 8\pi^2 \Delta_D^2 q_{\max}^2) \sum_{m=1}^{M} |\Delta\mathbf{q}_m| . \quad \text{(S5.7)}$$

Substituting this back into equation (26), we obtain equation (27).