This is the accepted manuscript made available via CHORUS. The article has been published as:

Signal, noise, and resolution in correlated fluctuations from snapshot small-angle x-ray scattering

Richard A. Kirian, Kevin E. Schmidt, Xiaoyu Wang, R. Bruce Doak, and John C. H. Spence

Phys. Rev. E **84**, 011921 — Published 25 July 2011

DOI: 10.1103/PhysRevE.84.011921

Signal, noise and resolution in correlated fluctuations from snapshot SAXS

Richard A. Kirian, Kevin E. Schmidt, Xiaoyu
Wang, R. Bruce Doak, and John C. H. Spence*

Department of Physics, Arizona State University, Tempe, Arizona 85287, USA

Abstract

It has been suggested that the three-dimensional structure of one particle may be reconstructed using the scattering from many, identical, randomly-oriented copies *ab-initio*, without modeling or *a-priori* information. This may be possible if these particles are frozen in either space or time, so that the conventional two-dimensional small-angle X-ray scattering (SAXS) distribution contains fluctuations and is no longer isotropic. We consider the magnitude of the correlated fluctuation SAXS (CFSAXS) signal for typical X-ray free-electron laser (XFEL) beam conditions, and compare this against the errors derived with the inclusion of Poisson photon counting statistics. The resulting signal-to-noise ratio (SNR) is found to rapidly approach a limit independent of the number of particles contributing to each diffraction pattern, so that the addition of more particles to a "single-particle-per-shot" experiment may be of little value, apart from reducing solvent background. When the scattering power is significantly less than one photon per particle per Shannon pixel, the SNR grows in proportion to incident flux. We provide simulations for protein molecules in support of these analytical results, and discuss the effects of solvent background scatter. We consider the SNR dependence on resolution and particle size, and discuss the application of the method to glasses and liquids, and the implications of more powerful XFELs, smaller focussed beams, and higher pulse repetition rates for this approach. We find that an accurate CFSAXS measurement may be acquired to sub-nanometer resolution for protein molecules if a 9 keV beam containing 10^{13} photons is focused to a ∼100 nm spot diameter, provided that the effects of solvent background can be reduced sufficiently.

PACS numbers: 61.05.C-, 61.05.cf

[*] Corresponding author: spence@asu.edu

I. INTRODUCTION

We consider experiments in which X-ray scattering is recorded from groups of identical molecules, lying in random orientations, which are frozen in either space or time. Where a pulsed X-ray source is used, the duration of the pulse should be less than the rotational diffusion time of the molecules. In that case, the conventional small-angle X-ray scattering ("solution scattering" or SAXS) which is obtained with long recording times become a two-dimensional scattering distribution, containing more information than the one-dimensional isotropic SAXS pattern. In 1977, Z. Kam suggested [1] that an analysis of correlations amongst fluctuations in such a "snap-shot SAXS" pattern would allow reconstruction of an image of one molecule, *a-priori*, without the modeling commonly used in SAXS analysis. Kam's approach is based on a summation of the angular correlation functions of many diffraction patterns from groups of identical molecules, which may be shown to converge to the angular correlation function for one molecule, under certain conditions. Iterative solutions of the phase problem may then be used to obtain a real-space image of the molecule. In the development of this approach, simulations have shown the feasibility of reconstructing the image of one membrane protein, using the scattering from many, if sufficient signal is obtained from molecules randomly oriented about a single axis [2]. A recent experiment, using soft X-ray scattering from many identical gold nanorods lying on their side on a transparent membrane, has since demonstrated the practicality of the method, at least for this two-dimensional case of single-axis rotations [3]. While the extension of the theory to the three-dimensional case for solution scattering is formally straightforward, the reconstruction of a three-dimensional real-space density map from scattering correlations is limited by the reduced amount of information available after rotational averaging [4], and may not be possible in practice without additional constraints.

In this paper we consider the limitations on resolution imposed by Poisson noise, with particular emphasis on the scattering patterns now available using a free-electron X-ray laser (XFEL) such as the Linac Coherent Light Source (LCLS) [5]. Here, despite the availability of perhaps 1×10^{13} photons per 70 fs pulse of incident hard X-rays, the number scattered by a single molecule at molecular resolution is much less than one per pixel. Recently, at the LCLS, an experimental single-shot diffraction pattern from a single virus has been phased and inverted to give a two-dimensional image of the virus at 32nm resolution [6]. Three-dimensional reconstruction of inorganic nanoparticle data has also been achieved, using expectation maximization to determine the relative orientations of the many identical particles used [7]. This raises the question as to whether higher resolution might be obtained using scattering from many such particles per shot, while relying on the correlated fluctuation SAXS (CFSAXS) method to resolve the orientation-determination problem. Our aim here is to obtain a simple expression for the variance of a fluctuation measurement in terms of the number of particles N per shot and number of shots M, with the inclusion of Poisson photon counting statistics. This will suggest the incident photon flux required, for a given number of particles per shot, and imaging resolution. In addition, we study the effective signal-to-noise ratio as a function of resolution and particle size.

II. BASIC THEORY

Consider a particle in orientation specified by ω scattering into a pixel of small but finite solid angle $\Delta\Omega$ centered at the scattering vector $\boldsymbol{q}$. Given an incident flux J (photons/area), the mean number of X-ray photons collected in the finite pixel is

$$\bar{n}(\boldsymbol{q}, \omega) = J\Theta(R_\omega \boldsymbol{q})\Delta\Omega \tag{1}$$

where $\Theta(q)$ denotes the differential scattering cross section of the particle, and R_ω is a general rotation matrix relating the particle orientation to the laboratory frame (we make no assumptions about the degree of rotational freedom in this paper; the orientation variable ω may specify three Euler angles with R_ω a 3×3 matrix, for instance). For a particle of maximum length L, we choose to sample in reciprocal space at an oversampling ratio of $s \geq 2$, so that the step size in scattering angle is approximately $\Delta\theta = \lambda/sL$ at small angles, and thus the effective pixel solid angle is

$$\Delta\Omega \approx \left(\frac{\lambda}{sL}\right)^2 . \tag{2}$$

For our purposes, a spatial correlation experiment should be carried out at relatively small scattering angles to avoid incomplete data (due to the maximum allowed angle subtended by two scattering vectors which, in order to conserve momentum for elastic scattering, must run from the origin to the surface of the Ewald sphere).

Now consider the case in which there are N identical particles per snapshot, in random orientations and positions, exposed to the same X-ray pulse simultaneously. We assume that N is fixed, the incident fluence J is spatially uniform, and that there are no significant multiple scattering effects. The coherence volume of the beam must be at least as large as a single particle, but a beam with a coherence volume that spans the entire group may also be used, following the analysis of [3, 8]. Here, we assume for simplicity that the particles are sufficiently dilute such that interparticle interference fringes are unobservable (because a single pixel integrates over a solid angle which spans many interference fringes). In this case, the mean photon counts for the kth diffraction pattern is effectively the summation of counts arising from each particle independently:

$$\bar{n}_k(\boldsymbol{q}) = \sum_{\alpha=1}^{N} \bar{n}(\boldsymbol{q}, \omega_\alpha^k) . \tag{3}$$

Since there is no interparticle interference, this is also the expression for the sum of N single-particle-per-shot diffraction patterns. Given the expectation value $\bar{n}$, the probability $p(n;\bar{n})$ of observing n photons will follow the Poisson distribution

$$p(n;\bar{n}) = \frac{\bar{n}^n}{n!}e^{-\bar{n}} \tag{4}$$

with the first and second moments

$$\sum_{n=0}^{\infty} np(n;\bar{n}) = \bar{n} \tag{5}$$

$$\sum_{n=0}^{\infty} n^2 p(n;\bar{n}) = \bar{n}^2 + \bar{n} \; . \tag{6}$$

The mean photon count $\bar{n}$ may also include the disordered solvent molecules in addition to the target solute particles. We will discuss this important contribution in section VI, but first we establish the basic theory in its absence.

We assume that any CFSAXS experiment consists of a sufficiently large number of snapshots, M, so that the central limit theorem applies to measured quantities. We take the experimental average as our measured value, with the standard error of the mean as the statistical error. The experimental average estimator of any quantity O with values O_k for each particular snapshot k $(1 \leq k \leq M)$, is defined as

$$\langle O_k \rangle_k = \frac{1}{M}\sum_{k=1}^{M} O_k \; . \tag{7}$$

The standard error of the mean will be estimated as

$$\mathcal{E}_O = \sqrt{\frac{\sigma_O^2}{M-1}} \tag{8}$$

where the variance is

$$\sigma_O^2 = \left\langle O_k^2 \right\rangle_k - \langle O_k \rangle_k^2 \; . \tag{9}$$

Finally, we define the signal-to-noise ratio of a measured observable O as the the absolute value of the mean divided by the standard error:

$$\mathcal{S}_O = \sqrt{\frac{(M-1)\left\langle O_k \right\rangle_k^2}{\sigma_O^2}} \; . \tag{10}$$

Our definition of the SNR therefore includes all experimental factors which contribute to the variance in a measured quantity (including, but not limited to, Poisson fluctuations).

III. ERROR ANALYSIS OF SNAPSHOT SAXS

We first consider the error analysis of a simple snapshot small-angle X-ray scattering (SAXS) experiment, in which our measurement is simply the mean photon counts in a pixel. We will see that this measurement is necessary to extract the desired single-particle correlation function, as discussed in the next section. Following the definitions in section II, the mean intensity arising from M N-particle patterns approaches

$$I_N(\boldsymbol{q}) \equiv \left\langle n_k(\boldsymbol{q}) \right\rangle_k \tag{11}$$

$$\to N \left\langle \bar{n}(\boldsymbol{q},\omega) \right\rangle_\omega \tag{12}$$

where the arrow hereafter indicates the mathematical limit $M \to \infty$. Similarly, the variance approaches

$$\sigma_{I_N}^2(\boldsymbol{q}) \equiv \left\langle n_k^2(\boldsymbol{q}) \right\rangle_k - \left\langle n_k(\boldsymbol{q}) \right\rangle_k^2 \tag{13}$$

$$\to N \left\{ \left[\left\langle \bar{n}(\boldsymbol{q},\omega)^2 \right\rangle_\omega - \left\langle \bar{n}(\boldsymbol{q},\omega) \right\rangle_\omega^2 \right] + \left\langle \bar{n}(\boldsymbol{q},\omega) \right\rangle_\omega \right\} \; , \tag{14}$$

where we have used the Poisson moments in equations 5 and 6, as detailed in appendix A.

For a sufficiently low flux, the first bracketed term in equation 14 (proportional to J^2) is negligible compared to the remaining term (proportional to J). Such is the case for a typical synchrotron-based SAXS measurement, in which the particles are free to rotate during exposures and thereby occupy a continuum of orientational states. In this low-flux regime, we find the expected result

$$\mathcal{S}_I(\boldsymbol{q}) \approx \sqrt{MN \left\langle \bar{n}(\boldsymbol{q}, \omega) \right\rangle_\omega} = \sqrt{MNJ \left\langle \Theta(R_\omega \boldsymbol{q}) \right\rangle_\omega \Delta\Omega} \,, \tag{15}$$

so that the SNR in a SAXS measurement is proportional to the square root of the product of flux, particle concentration, and number of snapshots; it depends only on the total number of scattered photons in the experiment.

In the case of an extremely intense pulsed beam, the bracketed term in the variance will instead dominate. We may call this the "self" noise term, to distinguish it from Poisson noise, and we recognize it as being the variance of the differential scattering cross section (with respect to the orientational average). The SNR is then approximately

$$\mathcal{S}_I(\boldsymbol{q}) \approx \sqrt{NM \frac{\left\langle \Theta(R_\omega \boldsymbol{q}) \right\rangle_\omega^2}{\left\langle \Theta(R_\omega \boldsymbol{q})^2 \right\rangle_\omega - \left\langle \Theta(R_\omega \boldsymbol{q}) \right\rangle_\omega^2}} \tag{16}$$

and is independent of incident fluence. Since it is generally true that

$$\left\langle \Theta(R_\omega \boldsymbol{q})^2 \right\rangle_\omega \geq \left\langle \Theta(R_\omega \boldsymbol{q}) \right\rangle_\omega^2 \tag{17}$$

we also have the result

$$\mathcal{S}_I(\boldsymbol{q}) \geq \sqrt{MN} \tag{18}$$

so that, apart from a factor which depends on the shape of the particle, the lower limit of the SNR of a SAXS measurement depends only on the total number of particles exposed to the beam (MN). We show in the following section that this result is in contrast to the SNR for a correlated fluctuation SAXS measurement, in which there is an upper bound on the SNR with respect to the number of particles per shot N.

IV. ERROR ANALYSIS OF CORRELATED FLUCTUATION SAXS

In a CFSAXS experiment, the additional information that we are interested in, beyond the conventional SAXS data, lies in the intensity fluctuations

$$\delta n_k(\boldsymbol{q}) \equiv n_k(\boldsymbol{q}) - \langle n_j(\boldsymbol{q})\rangle_j \tag{19}$$

and may be extracted by measuring the fluctuation correlation function

$$\begin{aligned}
\widetilde{C}_N(\boldsymbol{q}_1, \boldsymbol{q}_2) &\equiv \langle \delta n_k(\boldsymbol{q}_1)\delta n_k(\boldsymbol{q}_2)\rangle_k && (20)\\
&\to N\,\langle \bar{n}(\boldsymbol{q}_1,\omega)\bar{n}(\boldsymbol{q}_2,\omega)\rangle_\omega - N\,\langle \bar{n}(\boldsymbol{q}_1,\omega)\rangle_\omega\,\langle \bar{n}(\boldsymbol{q}_2,\omega)\rangle_\omega && (21)\\
&= N\,\langle \delta\bar{n}(\boldsymbol{q}_1,\omega)\delta\bar{n}(\boldsymbol{q}_2,\omega)\rangle_\omega && (22)
\end{aligned}$$

(see appendix B for details). We have assumed that the two pixels are distinct, so that the Poisson noise in each pixel is statistically independent of the other. The desired single-particle correlation function

$$C_1(\boldsymbol{q}_1, \boldsymbol{q}_2) \equiv \langle \bar{n}(\boldsymbol{q}_1,\omega)\bar{n}(\boldsymbol{q}_2,\omega)\rangle_\omega \tag{23}$$

is the first term in equation 21, which may be expressed in terms of *measurable quantities* as

$$C_1(\boldsymbol{q}_1, \boldsymbol{q}_2) = \frac{1}{N}\,\langle \delta n_k(\boldsymbol{q}_1)\delta n_k(\boldsymbol{q}_2)\rangle_k + \frac{1}{N^2}\,\langle n_k(\boldsymbol{q}_1)\rangle_k\,\langle n_k(\boldsymbol{q}_2)\rangle_k\ . \tag{24}$$

We would now like to determine the effective SNR for the measurement of $C_1(\boldsymbol{q}_1, \boldsymbol{q}_2)$.

Since we have already analyzed the errors in the SAXS terms of equation 24, we look at the first term now. As detailed in appendix B, upon factoring out N and inserting the Poisson moments, the variance approaches

$$\begin{aligned}
\sigma^2_{\widetilde{C}_N}(\boldsymbol{q}_1,\boldsymbol{q}_2) &= \left\langle (\delta n_k(\boldsymbol{q}_1)\delta n_k(\boldsymbol{q}_2))^2 \right\rangle_k - \langle \delta n_k(\boldsymbol{q}_1)\delta n_k(\boldsymbol{q}_2)\rangle_k^2 \qquad (25)\\
&\to N\big[\left\langle \delta\bar{n}^2(\boldsymbol{q}_1,\omega)\delta\bar{n}^2(\boldsymbol{q}_2,\omega)\right\rangle_\omega - \langle \delta\bar{n}(\boldsymbol{q}_1,\omega)\delta\bar{n}(\boldsymbol{q}_2,\omega)\rangle_\omega^2 \qquad (26)\\
&\quad + \left\langle \bar{n}^2(\boldsymbol{q}_1,\omega)\bar{n}(\boldsymbol{q}_2,\omega)\right\rangle_\omega + \left\langle \bar{n}(\boldsymbol{q}_1,\omega)\bar{n}^2(\boldsymbol{q}_2,\omega)\right\rangle_\omega\\
&\quad + \langle \bar{n}(\boldsymbol{q}_1,\omega)\rangle_\omega^2 \langle \bar{n}(\boldsymbol{q}_2,\omega)\rangle_\omega + \langle \bar{n}(\boldsymbol{q}_1,\omega)\rangle_\omega \langle \bar{n}(\boldsymbol{q}_2,\omega)\rangle_\omega^2\\
&\quad -2\,\langle \bar{n}(\boldsymbol{q}_1,\omega)\bar{n}(\boldsymbol{q}_2,\omega)\rangle_\omega \langle \bar{n}(\boldsymbol{q}_1,\omega)\rangle_\omega\\
&\quad -2\,\langle \bar{n}(\boldsymbol{q}_1,\omega)\bar{n}(\boldsymbol{q}_2,\omega)\rangle_\omega \langle \bar{n}(\boldsymbol{q}_2,\omega)\rangle_\omega\\
&\quad + \langle \bar{n}(\boldsymbol{q}_1,\omega)\bar{n}(\boldsymbol{q}_2,\omega)\rangle_\omega \big]\\
&\quad +(N^2-N)\big[\langle \delta\bar{n}(\boldsymbol{q}_1,\omega)\delta\bar{n}(\boldsymbol{q}_2,\omega)\rangle_\omega^2 + \left\langle \delta\bar{n}^2(\boldsymbol{q}_1,\omega)\right\rangle_\omega \left\langle \delta\bar{n}^2(\boldsymbol{q}_2,\omega)\right\rangle_\omega\\
&\quad + \left\langle \bar{n}^2(\boldsymbol{q}_1,\omega)\right\rangle_\omega \langle \bar{n}(\boldsymbol{q}_2,\omega)\rangle_\omega + \langle \bar{n}(\boldsymbol{q}_1,\omega)\rangle_\omega \left\langle \bar{n}^2(\boldsymbol{q}_2,\omega)\right\rangle_\omega\\
&\quad - \langle \bar{n}(\boldsymbol{q}_1,\omega)\rangle_\omega^2 \langle \bar{n}(\boldsymbol{q}_2,\omega)\rangle_\omega - \langle \bar{n}(\boldsymbol{q}_1,\omega)\rangle_\omega \langle \bar{n}(\boldsymbol{q}_2,\omega)\rangle_\omega^2\\
&\quad + \langle \bar{n}(\boldsymbol{q}_1,\omega)\rangle_\omega \langle \bar{n}(\boldsymbol{q}_2,\omega)\rangle_\omega \big]\ .
\end{aligned}$$

The two bracketed terms in the variance, with prefactors N and N^2-N, contain terms which scale as $J\Delta\Omega$ to the second, third, and fourth powers. For sufficiently small J and large N (such that far less than one scattered photon per pixel per particle is observed on average), the variance is approximately

$$\sigma^2_{\widetilde{C}_N} \approx N^2 \langle \bar{n}(\boldsymbol{q}_1,\omega)\rangle_\omega \langle \bar{n}(\boldsymbol{q}_2,\omega)\rangle_\omega\ , \qquad (27)$$

and the SNR is then

$$\mathcal{S}_{\widetilde{C}_N}(\boldsymbol{q}_1,\boldsymbol{q}_2) = \sqrt{\frac{M\widetilde{C}_N^2}{\sigma^2_{\widetilde{C}_N}}} \approx J\left(\frac{\lambda}{sL}\right)^2 \sqrt{M\frac{\langle \delta\Theta(\boldsymbol{q}_1,\omega)\delta\Theta(\boldsymbol{q}_2,\omega)\rangle_\omega^2}{\langle \Theta(\boldsymbol{q}_1,\omega)\rangle_\omega \langle \Theta(\boldsymbol{q}_2,\omega)\rangle_\omega}}\ . \qquad (28)$$

Since the factor N vanishes, we must conclude that, for the low-flux limit, *the SNR in a correlated fluctuation SAXS measurement is essentially independent of the number of particles per snapshot.*

Now consider the case of large N, but high flux so that terms with J^4 dominate. (We ignore the limit of $N = 1$ with large J because other imaging methods are likely to be superior to CFSAXS in this regime). The variance then becomes

$$\sigma^2_{\widetilde{C}_N} \approx N^2 \left[\langle \delta\bar{n}(\boldsymbol{q}_1,\omega)\delta\bar{n}(\boldsymbol{q}_2,\omega)\rangle^2_\omega + \left\langle \delta\bar{n}^2(\boldsymbol{q}_1,\omega)\right\rangle_\omega \left\langle \delta\bar{n}^2(\boldsymbol{q}_2,\omega)\right\rangle_\omega \right] \tag{29}$$

and the SNR is

$$\mathcal{S}_{\widetilde{C}_N}(\boldsymbol{q}_1,\boldsymbol{q}_2) \approx \sqrt{M\left(1 + \frac{\langle \delta\Theta(\boldsymbol{q}_1,\omega)\delta\Theta(\boldsymbol{q}_2,\omega)\rangle^2_\omega}{\langle \delta\Theta^2(\boldsymbol{q}_1,\omega)\rangle_\omega \langle \delta\Theta^2(\boldsymbol{q}_2,\omega)\rangle_\omega}\right)}\ . \tag{30}$$

Since, by the Schwartz inequality, it is generally true that [9]

$$\langle \delta\Theta(\boldsymbol{q}_1,\omega)\delta\Theta(\boldsymbol{q}_2,\omega)\rangle^2_\omega \leq \left\langle \delta\Theta^2(\boldsymbol{q}_1,\omega)\right\rangle_\omega \left\langle \delta\Theta^2(\boldsymbol{q}_2,\omega)\right\rangle_\omega\ , \tag{31}$$

it follows that the SNR lies in the range

$$\sqrt{M} \leq \mathcal{S}_{\widetilde{C}_N}(\boldsymbol{q}_1,\boldsymbol{q}_2) \leq \sqrt{2M}\ . \tag{32}$$

Just as we found in the case of the snapshot SAXS error, in the high flux limit we can only improve the measurement through collecting more patterns to provide an orientational average. However, a correlations measurement must also average out the self noise terms associated with products of uncorrelated intensities arising from particles in differing orientations, and as a result *we cannot improve the SNR by increasing the number of particles in each snapshot.*

Finally, we consider how the errors in the measured SAXS terms affect the error in the resulting measurement of $C_1(\boldsymbol{q}_1,\boldsymbol{q}_2)$ (equation 24). If we include these terms in the variance of $C_1(\boldsymbol{q}_1,\boldsymbol{q}_2)$, using the error propagation formula [10]

$$\sigma^2_x \approx \sigma^2_u \left(\frac{\partial x}{\partial u}\right)^2 + \sigma^2_v \left(\frac{\partial x}{\partial v}\right)^2 + \cdots\ , \tag{33}$$

we have from equation 24

$$\sigma^2_{C_1} \approx \frac{1}{N}\left[\sigma^2_{\widetilde{C}_N}(\boldsymbol{q}_1,\boldsymbol{q}_2) + \frac{1}{N}\sigma^2_{I_N}(\boldsymbol{q}_1)I_N(\boldsymbol{q}_2) + \frac{1}{N}\sigma^2_{I_N}(\boldsymbol{q}_2)I_N(\boldsymbol{q}_1)\right]\ . \tag{34}$$

Since the second and third terms on the right-hand side are proportional to N (from equations 12 and 14), these terms are similar to the first bracketed term in $\sigma^2_{\widetilde{C}_N}$. They are therefore insignificant when N is large, and moreover, these SAXS terms may perhaps be measured more accurately using a continuous X-ray source as discussed in section III. We therefore ignore this contribution to the SNR, and take $\mathcal{S}_{\widetilde{C}_N}(\boldsymbol{q}_1, \boldsymbol{q}_2)$ as the figure of merit for our purposes here.

V. INTENSITY STATISTICS: RESOLUTION AND PARTICLE SIZE

In addition to particle counts and incident flux, we would also like to understand how particle size and resolution effect the CFSAXS SNR. We therefore wish to determine typical values for the terms which appear in equations 28 and 30 for a typical protein molecule. Following the Wilson statistical model, we assume the protein contains m atoms that are in essentially random positions $\boldsymbol{r}_j$ (we assume no symmetry), and is of characteristic size L. The scattered intensity is proportional to the scattering cross section

$$\Theta(\boldsymbol{q}) = \left| \sum_{j=1}^{m} f_j(q) e^{i\boldsymbol{q}\cdot\boldsymbol{r}_j} \right|^2 \tag{35}$$

where $f_j(q)$ is an atomic scattering factor (with units of area– the classical electron radius is inclusive). Since, for a large protein with many atoms, the phase $\boldsymbol{q} \cdot \boldsymbol{r_j}$ may be assumed to be a random number, we arrive at the mean values (see appendix C)

$$\langle \Theta(\boldsymbol{q}, \omega) \rangle_\omega = m \left\langle f_j^2(q) \right\rangle_j \tag{36}$$

$$\left\langle \Theta^2(\boldsymbol{q}, \omega) \right\rangle_\omega = 2 \left\langle \Theta(\boldsymbol{q}, \omega) \right\rangle_\omega^2 \tag{37}$$

$$\left\langle \delta\Theta^2(\boldsymbol{q}, \omega) \right\rangle_\omega = \left\langle \Theta(\boldsymbol{q}, \omega) \right\rangle_\omega^2 \tag{38}$$

Next we would like to determine a typical magnitude for the term

$$\langle \delta\Theta(\boldsymbol{q}_1, \omega)\delta\Theta(\boldsymbol{q}_2, \omega) \rangle_\omega^2 = \left[\langle \Theta(\boldsymbol{q}_1, \omega)\Theta(\boldsymbol{q}_2, \omega) \rangle_\omega - \langle \Theta(\boldsymbol{q}_1, \omega) \rangle \langle \Theta(\boldsymbol{q}_2, \omega) \rangle \right]^2 \ .$$

To do this, we calculate the mean value over all pairs of scattering vectors $\boldsymbol{q}_1, \boldsymbol{q}_2$ which, for simplicity, lie on the same resolution shell ($q_1 = q_2 = q$). We expand $\delta\Theta(\boldsymbol{q}, \omega)$ in spherical harmonics as

$$\delta\Theta(\boldsymbol{q}, \omega) = \sum_{l=1}^{l_{\max}} \sum_{m=-l}^{l} A_{lm}(q) \sum_{m'=-l}^{l} Y_{lm'}(\hat{\boldsymbol{q}}) D^{(\omega)}_{lmm'} \tag{39}$$

where $D^{(\omega)}_{lmm'}$ is a Wigner rotation matrix, and $l_{\max} \approx Lq$. Since the scattered intensity is real, we must have $A_{lm}(q) = A^*_{l-m}(q)$, and by Friedel's law, we must have $A_{lm} = 0$ for odd l. Then we have

$$\begin{aligned}\langle \delta\Theta(\boldsymbol{q}_1, \omega)\delta\Theta(\boldsymbol{q}_2, \omega)\rangle_\omega &= \sum_{l,l'=1}^{l_{\max}} \sum_{m=-l}^{l} \sum_{m'=-l'}^{l'} Y_{lm}(\hat{\boldsymbol{q}}_1) Y^*_{l'm'}(\hat{\boldsymbol{q}}_2) \times & (40)\\ &\quad \sum_{m''=-l}^{l} \sum_{m'''=-l'}^{l'} A_{lm''}(q) A^*_{l'm'''}(q) \left\langle D^{(\omega)}_{lmm''} D^{(\omega)*}_{l'm'm'''} \right\rangle_\omega \\ &= \sum_{l=1}^{l_{\max}} \sum_{m'=-l}^{l} \frac{Y_{lm'}(\hat{\boldsymbol{q}}_1) Y^*_{lm'}(\hat{\boldsymbol{q}}_2)}{2l+1} \sum_{m=-l}^{l} |A_{lm}(q)|^2 & (41)\\ &= \frac{1}{4\pi} \sum_{l=1}^{l_{\max}} P_l(\hat{\boldsymbol{q}}_1 \cdot \hat{\boldsymbol{q}}_2) \sum_{m=-l}^{l} |A_{lm}(q)|^2 & (42)\end{aligned}$$

upon using the orthogonality of the Wigner matrices and the spherical harmonic addition theorem [11]. The $P_l(x)$ are Legendre polynomials. If we square this quantity and average over $\hat{\boldsymbol{q}}_1 \cdot \hat{\boldsymbol{q}}_2$ we get

$$\begin{aligned}\left\langle \langle \delta\Theta(\boldsymbol{q}_1, \omega)\delta\Theta(\boldsymbol{q}_2, \omega)\rangle^2_\omega \right\rangle_{\hat{\boldsymbol{q}}_1 \cdot \hat{\boldsymbol{q}}_2} &= \sum_{m=-l}^{l} |A_{lm}(q)|^2 \sum_{m'=-l'}^{l'} |A_{l'm'}(q)|^2 \times & (43)\\ &\quad \frac{1}{16\pi^2} \sum_{l,l'=1}^{l_{\max}} \frac{1}{4\pi} \int_{-1}^{1} d\cos\theta \int_0^{2\pi} d\phi P_l(\cos\theta) P_{l'}(\cos\theta) \\ &= \frac{1}{16\pi^2} \sum_{l=1}^{l_{\max}} \frac{1}{2l+1} \left[\sum_{m=-l}^{l} |A_{lm}(q)|^2 \right]^2 & (44)\end{aligned}$$

after using the orthogonality of the Legendre polynomials [12]. Similarly, we may write

$$\left\langle \delta\Theta^2(\boldsymbol{q},\omega)\right\rangle_\omega = \left\langle \left[\sum_{lmm'} A_{lm'}(q) Y_{lm'}(\hat{\boldsymbol{q}}) D^{(\omega)}_{lmm'}\right]^2 \right\rangle_\omega \tag{45}$$

$$= \frac{1}{4\pi}\sum_{l=1}^{l_{\max}} P_l(\hat{\boldsymbol{q}}\cdot\hat{\boldsymbol{q}}) \sum_{m=-l}^{l} |A_{lm}(q)|^2 \tag{46}$$

$$= \frac{1}{4\pi}\sum_{l=1}^{l_{\max}} \sum_{m=-l}^{l} |A_{lm}(q)|^2 \,. \tag{47}$$

For our statistical model, the SNR depends only on the *magnitudes* of the complex numbers $A_{lm}(q)$, which will vary considerably depending on the shape of the molecule. Let us consider a hypothetical case in which the values of $|A_{lm}(q)|^2$ are equal to a constant A^2. Then, upon taking $|A_{lm}(q)|^2$ outside of the summations, we have

$$\left\langle \langle \delta\Theta(\boldsymbol{q}_1,\omega)\delta\Theta(\boldsymbol{q}_2,\omega)\rangle^2_\omega \right\rangle_{\hat{\boldsymbol{q}}_1\cdot\hat{\boldsymbol{q}}_2} = \frac{A^4}{16\pi^2}(l^2_{\max} + 2l_{\max}) \tag{48}$$

and similarly

$$\left\langle \delta\Theta^2(\boldsymbol{q},\omega)\right\rangle_\omega = \langle \Theta(\boldsymbol{q},\omega)\rangle^2_\omega = \frac{A^2}{4\pi}(l^2_{\max} + 2l_{\max}) \tag{49}$$

where we have made use of equation 38. Inserting these results into equation 30, the resulting high-flux, large-N SNR is

$$\mathcal{S}_{\widetilde{C}_N}(\boldsymbol{q}_1,\boldsymbol{q}_2) \approx \sqrt{M\left(1 + \frac{1}{(Lq/2)^2 + Lq}\right)} \,. \tag{50}$$

where we've used the fact that only even l are permitted. Similarly, for the low-flux limit we have

$$\mathcal{S}_{\widetilde{C}_N}(\boldsymbol{q}_1,\boldsymbol{q}_2) \approx J\left(\frac{\lambda}{sL}\right)^2 \sqrt{M\frac{A^2}{4\pi}} \,. \tag{51}$$

Since the magnitude of A^2 and the applicability of our simplifying assumptions will depend strongly on the shape and size of the particle, we caution against drawing strong conclusions from this model. It is, however, interesting that equation 50 suggests that larger molecules result in a *lower* SNR than smaller molecules. This is, however, a factor of $\sqrt{2}$ at most, according to the general result of equation 32. From the simulations in section VII we find that the SNR is indeed reduced for the larger of the two molecules we consider.

VI. SOLVENT SCATTER AND BACKGROUND

Diffraction from solvent molecules is an important factor in a correlations measurement, will likely be the dominant noise contribution at low photon counts. At low resolution, we may model the affects of the solvent by defining the effective particle electron density as its Babinet contrast against the surrounding solvent. Here, we are concerned with sub-nanometer resolutions, so we must take into consideration the correlated scattering from closely-packed solvent molecules (for a detailed theoretical description, see [8]). The differential scattering cross section for the kth pattern may be written as

$$\Theta_k(\boldsymbol{q}) = \left|\Psi_k^s(\boldsymbol{q}) + \Psi_k^p(\boldsymbol{q})\right|^2 \tag{52}$$

where $\Psi_k^s(\boldsymbol{q})$ is the scattering amplitude for the kth ensemble of disordered solvent molecules (e.g. H_2O), and $\Psi_k^p(\boldsymbol{q})$ is the scattering amplitude from the kth ensemble of solute particles (e.g. proteins). We assume that the solvent molecule positions are uncorrelated with respect to the solute particle positions, otherwise they should be considered part of the structure of the dissolved particles. Upon taking the average over many patterns, we have

$$\langle\Theta_k(\boldsymbol{q})\rangle_k = \left\langle|\Psi_k^s(\boldsymbol{q})|^2\right\rangle_k + \left\langle|\Psi_k^p(\boldsymbol{q})|^2\right\rangle_k \tag{53}$$

since the mean product of uncorrelated amplitudes tends toward zero. In terms of photon counts, we can write

$$\langle n_k(\boldsymbol{q})\rangle_k = \langle n_k^s(\boldsymbol{q})\rangle_k + \langle n_k^p(\boldsymbol{q})\rangle_k \tag{54}$$

where $\langle n_k^s(\boldsymbol{q})\rangle_k$ are the average "background" counts from the solvent, and $\langle n_k^p(\boldsymbol{q})\rangle_k$ are the counts contributed by the particles. Similarly, we may write a correlated product as

$$\begin{aligned}\langle n_k(\boldsymbol{q}_1) n_k(\boldsymbol{q}_2)\rangle_k = &\langle n_k^s(\boldsymbol{q}_1) n_k^s(\boldsymbol{q}_2)\rangle_k + \langle n_k^p(\boldsymbol{q}_1) n_k^p(\boldsymbol{q}_2)\rangle_k \\ &+ \langle n_k^s(\boldsymbol{q}_1)\rangle_k \langle n_k^p(\boldsymbol{q}_2)\rangle_k + \langle n_k^p(\boldsymbol{q}_2)\rangle_k \langle n_k^s(\boldsymbol{q}_2)\rangle_k \ .\end{aligned} \tag{55}$$

The first term is an undesirable background contribution that must be subtracted. Since we cannot assume that background scatter is uncorrelated with itself, we must carefully measure the correlations in the solvent alone in order to properly remove the first background term in equation 55. The last two mixed terms may be measured with conventional SAXS methods, without the need for snapshot diffraction patterns, and as discussed in section IV, we assume that these SAXS terms have a negligible effect on experimental errors. Taking the limiting case of low flux, where background Poisson fluctuations are important, and applying the error propagation formula (equation 33), we find that the variance is simply the summation over the variance in the first two terms. Assuming that the variance in the correlations of closely-packed solvent molecules is roughly similar to that of the solute particles, we may use the result of equation 27 to express the variance as

$$\sigma^2_{\widetilde{C}_N} \approx N_s^2 \langle \bar{n}^s(\boldsymbol{q}_1,\omega)\rangle \langle \bar{n}^s(\boldsymbol{q}_2,\omega)\rangle + N^2 \langle \bar{n}^p(\boldsymbol{q}_1,\omega)\rangle \langle \bar{n}^p(\boldsymbol{q}_2,\omega)\rangle \tag{56}$$

where N_s is the number of solvent molecules. Since the volume fraction of solvent is likely much greater than the solute particles, the solvent term will be the dominant contribution to Poisson fluctuations, which results in an approximate SNR of

$$\mathcal{S}_{\widetilde{C}_N}(\boldsymbol{q}_1, \boldsymbol{q}_2) \approx \sqrt{M \frac{N^2 \left\langle \delta\bar{n}^p(\boldsymbol{q}_1, \omega)\delta\bar{n}^p(\boldsymbol{q}_2, \omega)\right\rangle_\omega^2}{N_s^2 \left\langle \bar{n}^s(\boldsymbol{q}_1, \omega)\right\rangle_\omega \left\langle \bar{n}^s(\boldsymbol{q}_2, \omega)\right\rangle_\omega}} \,. \tag{57}$$

We note that the treatment of other sources of scatter, such as parasitic scatter from the instrument, or solution contaminants, may be treated similarly to the scatter from solvent molecules, and may be grouped into the term $\Psi_k^s(\boldsymbol{q})$. When multiple species of (uncorrelated) solute particles are present, the resulting correlation function is simply the number-weighted average over all species.

VII. SIMULATIONS

We first verified the correctness of the variance expressed in equation 27 and the result that the CFSAXS SNR is practically independent of the number of particles per snapshot N. To do this, we chose a simple analytical expression for the scattering cross section in which the particles are aligned and confined to rotations about an axis parallel to the incident beam. Where the particle is in the orientation specified by the angle ϕ, we define scattered photon counts as the "sawtooth" function

$$\bar{n}(\boldsymbol{q}, \phi) = \begin{cases} J\phi/\pi & 0 \leq \phi < \pi \\ J(\phi - \pi)/\pi & \pi \leq \phi < 2\pi \end{cases} \,. \tag{58}$$

Choosing, for simplicity, the pair of scattering vectors $\boldsymbol{q}_1 = -\boldsymbol{q}_2$, we arrive at the analytical expressions

$$\left\langle \bar{n}(\boldsymbol{q}, \phi)\right\rangle_\phi = J/2 \tag{59}$$

$$\widetilde{C}_N(\boldsymbol{q}, \phi) = NJ^2/12 \tag{60}$$

$$\sigma^2_{\widetilde{C}_N}(\boldsymbol{q}, -\boldsymbol{q}) = N\left[\frac{J^2}{12} + \frac{J^3}{12} + \frac{7J^4}{360}\right] + (N^2 - N)\left[\frac{J^2}{4} + \frac{J^3}{12} + \frac{J^4}{72}\right] \,. \tag{61}$$

The form of the resulting SNR is graphed in figure 1 (solid lines), for various values of N, as a function of mean photon counts per particle. We confirmed this analytical result by Monte Carlo calculation of the fluctuation correlation and its variance, according to equations 7, 9 and 10, using a flat distribution of randomly-generated ϕ values, and values of $n(\boldsymbol{q}, \phi)$ drawn randomly from a Poisson distribution with mean $\bar{n}(\boldsymbol{q}, \phi)$. Shown as circles are the Monte Carlo results after averaging 10,000 simulated experiments, each with 1000 snapshots, for values of $J = 0.02$, 0.2, 2, 20, and 200. The simulated results are in remarkable agreement with analytical expressions, and, as predicted, the SNR scales approximately linearly with increasing flux prior to a mean photon count of one per particle, and then reaches its asymptotic value (within $\sim 5\%$) after 1-2 decades of flux increase. The asymptotic value does not depend strongly on the number of particles; the case of $N = 10$ is remarkably close to the predicted value for $N \to \infty$.

Following the verification of equation 27, we carried out a similar computation of SNR for the more realistic scattering cross sections corresponding to two protein molecules: the large monomer unit of the Photosystem I (PSI) complex (Protein Data Bank (PDB) [13] entry 1JB0), and hen egg white lysozyme (entry 2LYZ). The differential scattering cross sections were first calculated on a GPU for all points on a cartesian grid (with oversampling ratio $s = 6$, as determined by the maximum distance between atom pairs) using equation 35, for all non-hydrogen atomic coordinates in the PDB files (24,198 coordinates in total for PSI, 1,102 for lysozyme). For simplicity, all scattering factors were taken to be that of nitrogen, a good approximation to the average non-hydrogen atomic scattering factor [14], using the expression

$$f_N(\theta) = 7 \exp\left(-10.7\text{Å}^2 \left(\sin\theta/\lambda\right)^2\right) \tag{62}$$

where θ here is the Bragg angle (twice the scattering angle). From the cartesian grid of differential scattering cross sections, the terms in equation 25 were calculated in Monte Carlo fashion for randomly-oriented scattering vector pairs $\boldsymbol{q}_1$, $\boldsymbol{q}_2$ lying on

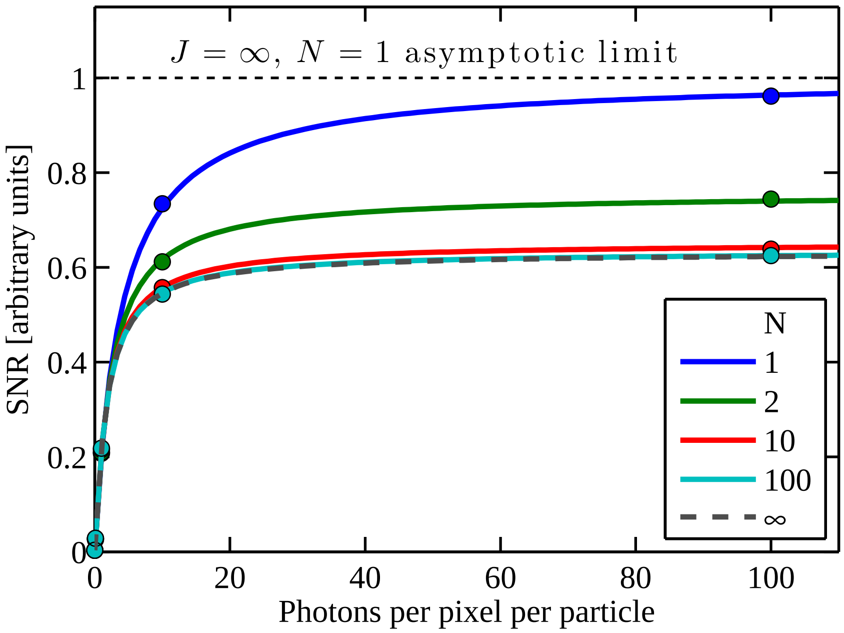

FIG. 1. (Color online) Signal-to-noise for the simple cross section in equation 58. The top dashed line indicates the maximun SNR in the case of infinite incident fluence J and number of particles per shot $N = 1$. Simulated values, using randomized orientations and photon counts, are shown as filled circles.

the same resolution shell $|\boldsymbol{q}_1| = |\boldsymbol{q}_2| = q$, using trilinear interpolation. We therefore assumed a uniform distribution of particle orientations in three-dimensions. The full range of angular separations $\phi = \arccos(\boldsymbol{q}_1 \cdot \boldsymbol{q}_2)$ (Ewald curvature was neglected) were computed, and averages for each separation ϕ were taken over 10^6 randomly-oriented scattering vector pairs for each separation ϕ. The resulting SNR, averaged over the range $0.1\pi < \phi < 0.9\pi$, is graphed against incident flux in figure 2, for various resolutions. (We ignored values near $\phi = 0$ and π since the spikes in intensity contain redundant SAXS information). Values shown represent the SNR for a single snapshot; for the case of M shots, the SNR should be multiplied by $\sqrt{M}$. Similarly, if each snapshot produces multiple observations for a given ϕ, M will be effectively increased by this multiplicity.

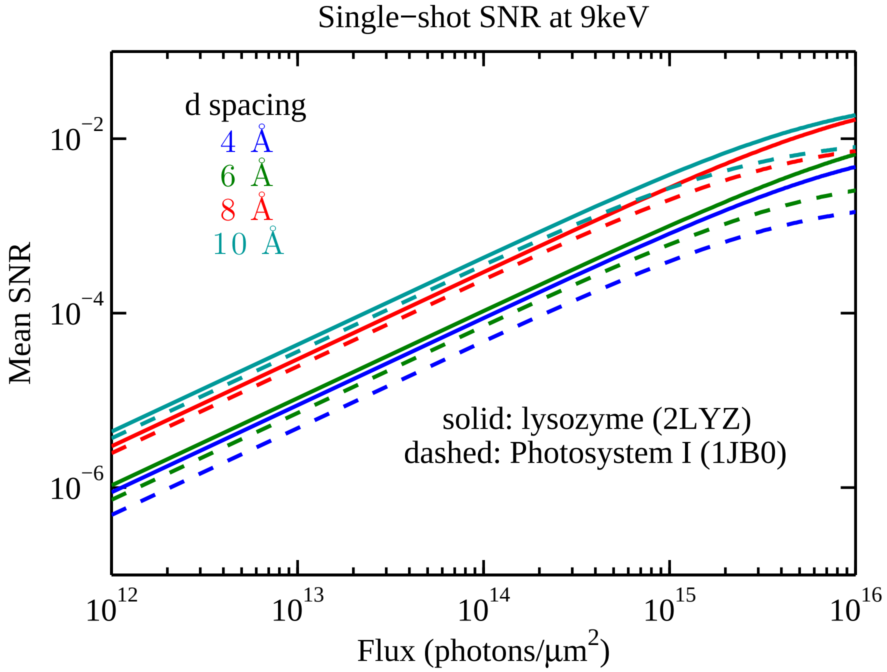

FIG. 2. (Color online) Calculated single-shot SNR for lysozyme and photosystem I (solid and dashed lines, respectively) as a function of incident photon fluence and imaging resolution (d-spacings indicated by colors or grayscale– SNR increases monotonically with d-spacing), at photon energy 9 keV. Solvent background has been neglected. The number of particles per snapshot is $N = 1$, however, other values of N are indistinguishable on this scale.

Shown in figures 3 and 4 are direct simulations of the fluctuation correlation function $C_N(\boldsymbol{q}_1, \boldsymbol{q}_2)$, calculated by averaging the ring autocorrelations (as would be measured in an experiment using an area detector) of 10^6 randomly-oriented rings of scattered counts, each with Poisson noise added. The spacings between ϕ values was taken to be $\Delta\phi = 2L/d$, corresponding to an oversampling ratio of $s = 2$, so that there are $N_\phi = 4\pi L/d$ samples about a full "ring" of scattered intensity (assuming an area detector is used without gaps in the pixel array). Numerical results from the simulations at photon flux $J = 10^{14}$ $\mu\mathrm{m}^{-2}$ are shown in table I, where the $\langle \mathcal{S} \rangle$ in this case is calculated as the RMS fluctuation magnitude divided by the RMS residual magnitude in the range $0.1\pi < \phi < 0.9\pi$. Values of $\langle \mathcal{S} \rangle / \sqrt{MN_\phi}$ represent the equivalent single-snapshot SNR, from which SNR may be estimated for any flux and number of snapshots M using the proportionality SNR $\propto J\sqrt{MN_\phi}$ (for values of $J \lesssim 10^{15}$, where the linear approximation holds).

Finally, we have estimated scattering counts from a water background with 1:1 water-to-protein volume ratio, using pure water scattering factors $|F_{\mathrm{H_2O}}(q)|$ derived from SAXS measurements [15]. Estimations of the decreased SNR due to a water background can be made using equation 57 in conjunction with II, where the background counts should be scaled as appropriate for a given protein concentration.

VIII. DISCUSSION

Zvi Kam long ago appreciated that the SNR in a CFSAXS experiment is essentially independent of the number of particles per snapshot, and suggested that a practical experimental aim is to obtain one scattered photon per particle, per pixel, per snapshot [16]. If experiments are conducted using a synchrotron source, the maximum number of counts per snapshot will be limited by radiation damage. Assuming, for instance, that near-atomic resolution is desired, in which the maximum tolerable dose is approximately 30 MGy [17], and the typical dose ratio of proteins is 2000 photons $\mu\mathrm{m}^{-2}$ Gy^{-1} [18], a single snapshot exposure would necessarily be limited to a flux of 6×10^{10} photons $\mu\mathrm{m}^{-2}$. In our simulations of PSI presented here, this results in approximately 2×10^{-5} photons per pixel per particle at 10 Å resolution, a far cry from Kam's idealized experiment. However, this picture changes

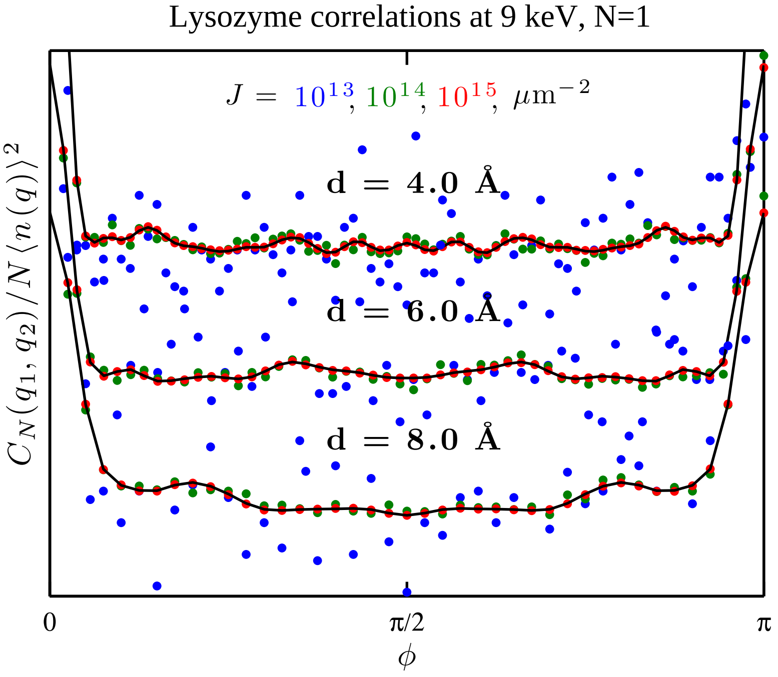

FIG. 3. (Color online) Simulated lysozyme fluctuation correlation function $C_N(\boldsymbol{q}_1, \boldsymbol{q}_2)$, normalized by the factor $1/N\,\langle n(\boldsymbol{q})\rangle^2$ and offset vertically for display purposes. Wavevector pairs lie on the same resolution shell $|\boldsymbol{q}_1| = |\boldsymbol{q}_2| = q$. Results represent an average over of 10^6 patterns with one particle per pattern ($N = 1$), plotted for various values of flux and resolution. Blue, green, and red points correspond to 10^{13}, 10^{14}, and 10^{15} photons/μm^2, respectively (indicated by grayscale in print version). The solid black line indicates infinite flux (i.e. a simulation without poisson noise).

considerably if data are collected using an XFEL in the "diffract-and-destroy" mode, in which the incident pulse terminates prior to significant radiation damage. In this case, doses well beyond the conventionally accepted maximum tolerable dose may be delivered to the target, and since $\mathcal{S} \propto J\sqrt{M}$, halving the number of shots in favor of doubling the single-shot flux improves SNR by a factor of $\sqrt{2}$. At present, the LCLS

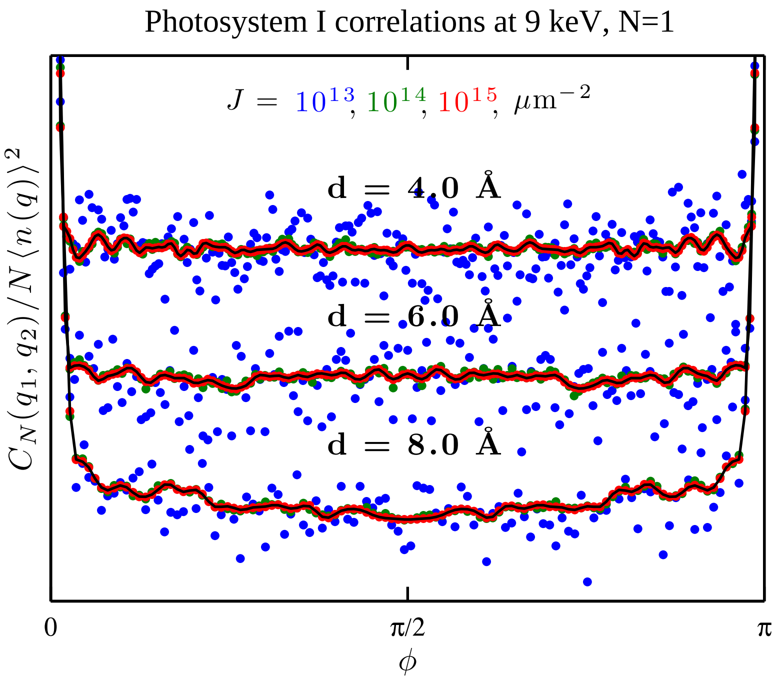

FIG. 4. (Color online) Simulated Photosystem I fluctuation correlation function (see figure 3 caption for details). Note the finer angular sampling due to larger protein size.

can deliver approximately 1×10^{12} photons per pulse at 9 keV, which may be focused to a 0.1 μm beam spot in the near future at the coherent X-ray imaging instrument (CXI) [19]. At the current pulse repetition rate of 120 Hz, the improvement in SNR at the LCLS is approximately 400-fold when compared against a synchrotron that delivers 10^{12} photons per second into the same 0.01 μm^2 beam spot area, but is limited to a single-shot dose of 6×10^{10} photons μm^{-2}. This remains below Kam's ideal.

After signal averaging a large number of shots, our derived SNR is essentially independent of the number of particles per shot, in the absence of solvent molecules and interparticle interference. Therefore a many-particle-per-shot CFSAXS experiment (summed over many shots) is essentially equivalent, from a SNR standpoint,

TABLE I. Results from direct simulations of $M = 10^6$ patterns at flux $J = 10^{14}$ μm^{-2}.

$\frac{2\pi}{d}$ [Å]	N_ϕ	$\langle n(q)\rangle$	$\langle\mathcal{S}\rangle$	$\frac{\langle\mathcal{S}\rangle}{\sqrt{MN_\phi}}$
Lysozyme (2LYZ)				
4	160	8.41×10^{-3}	1.35	1.07×10^{-4}
6	106	6.77×10^{-3}	1.41	1.37×10^{-4}
8	80	7.12×10^{-3}	3.35	3.75×10^{-4}
10	64	15.2×10^{-3}	5.89	7.36×10^{-4}
Photosystem I (1JB0)				
4	450	2.41×10^{-2}	1.19	5.60×10^{-5}
6	300	1.79×10^{-2}	1.61	9.31×10^{-5}
8	226	2.25×10^{-2}	4.86	32.4×10^{-5}
10	180	3.27×10^{-2}	5.66	42.2×10^{-5}

to one with only $N = 1$ per shot (also summed over many shots). It follows that a three-dimensional intensity map derived from a CFSAXS experiment cannot produce a more accurate measurement than one would obtain through direct averaging of intensities from the identical randomly-oriented, single-particle-per-shot experiment (provided that orientations are *accurately* determined). In the presence of solvent molecules, however, there is a clear advantage to increasing the number of particles per shot, since each displaces a small volume of solvent which otherwise contributes to background. High particle concentrations are therefore desirable in this case, provided that the particles do not aggregate or otherwise break our assumptions that interparticle interference is negligible and that positions and orientations are random and uncorrelated. In the case that the particles may form a closely-packed network, or have significant pair correlations, the theory presented here is not applicable. It is

TABLE II. Estimated water background counts for 1:1 volume ratio at flux $J = 10^{14}\ \mu\text{m}^{-2}$.

$\frac{2\pi}{d}$ [Å]	$\lvert F_{\mathrm{H_2O}}(q)\rvert^2$ [e.u.]	$\frac{N_{\mathrm{H_2O}}}{N_{\mathrm{prot}}}\langle n_{\mathrm{H_2O}}(q)\rangle$
Lysozyme (2LYZ)		
4	28.30	26.7×10^{-4}
6	8.30	7.83×10^{-4}
8	6.90	6.51×10^{-4}
10	6.50	6.13×10^{-4}
Photosystem I (1JBO)		
4	28.30	7.34×10^{-3}
6	8.30	2.15×10^{-3}
8	6.90	1.79×10^{-3}
10	6.50	1.69×10^{-3}

possible that, in favorable cases, the SNR will be enhanced by interparticle interference (as, for example, in the case that the "aggregates" are crystalline, as in powder diffraction).

Our approach is also not immediately applicable to glasses or any continuously-bonded random network of atoms. However models of, for example, amorphous silicon, exist in which the structure is represented by a small number of local structural units connected together. For a sample consisting of several different types of molecules, our CFSAXS analysis yields a weighted sum of correlation functions for each. Hence it is possible that this approach could be extended to the case of glasses which can be described in this way (so that a few structural correlations accumulate at the expense of random atomic arrangements) if the interparticle interference terms average to zero.

Where the CFSAXS methodology is applied to single-particle imaging, it must compete favorably with the alternative techniques [20, 21] if it is to be of any practical use. Its greatest strength is perhaps the relative ease with which experiments may be performed. Since the method applies when many particles are exposed in each shot, there is a clear experimental advantage in that the particles may be confined in a droplet using established liquid jet injectors [22–24] in order to ensure that every XFEL pulse meets a target. The CFSAXS method therefore offers a 100 percent hit rate. The CFSAXS data are merged in a straightforward manner, without the need to classify particle orientations, or fit a manifold to hundreds of terabytes of data. The formation of angular correlation functions could conceivably be done at the detector during data acquisition. The final data set is highly compressed, and computations are likely to be easily tractable on a modest computer. The potential merits, however, come at a significant cost. The increased hit rate is likely to be accompanied by increased background and decreased contrast when compared with aerosol injectors (with their much lower hit rate), and the compression of the data may result in an insoluble problem without a further set of constraints provided by additional data, or measurement of triple correlations.

A water background is comparable to the scatter from proteins at resolutions below about 5 Å. A water-to-protein volume ratio of 100:1, for instance, will result in approximately a 100-fold decrease in SNR, according to our approximate expression in equation 57. Factoring in this decrease, and assuming that we can collect 10^7 patterns (a 24-hour beamtime at the current LCLS repetition rate of 120 Hz) snapshots at a flux of 10^{13} photons per 0.01 μm^{-2}, it may be possible to obtain an SNR of $\sim$2 at 10 Å resolution. This analysis, however, has neglected to consider the effects of particle non-homogeneity, and beam effects such as spectral width and divergence.

Despite the apparent difficulty of collecting a high-quality CFSAXS data set at sub-nanometer resolution, we note that the LCLS is a first-of-its-kind instrument, and improvements to XFEL capabilities in general (including the needed increase in X-ray fluence) are inevitable. The possibility of 1 nm resolution would already exceed the resolution attainable through conventional SAXS methods, where resolutions better than 1.5 nm are not possible without *a-priori* high-resolution structural information acquired through other techniques [25]. Provided that inversion algorithms can be developed to properly treat the data, the greatest appeal to the CFSAXS methodology is the possibility to determine structures from proteins *in solution at room temperature*. The extension to ultrafast dynamical studies is natural, given the extremely brief pulse duration produced by XFELs. The potential for opening up this new regime of structural studies, and the experimental feasibility reported here, justifies the extraordinary efforts that may be required to determine the full potential of this method.

IX. CONCLUSIONS

We have determined, through both theory and in simulations, that the signal-to-noise ratio for a CFSAXS experiment is essentially independent of the number of particles per pattern. Therefore, the CFSAXS method cannot improve upon the errors in recovered intensity maps which are derived from methods which can accurately determine particle orientations prior to merging intensities. However, since signal-to-noise scales with the square root of the number of diffraction patterns for any signal averaging method, and since injection of isolated biomolecules to a sub-micron beam remains a significant challenge, the CFSAXS approach may currently hold an important advantage over other methods since well-tested liquid jet injectors may be used for sample injection in order to ensure a 100 percent hit rate. Given the beam parameters anticipated at the LCLS in the near future, our simulations suggest that it may be possible to achieve a signal-to-noise ratio of better than one at sub-nanometer resolution within a 24-hour beamtime, with the effects of a small amount of solvent scatter considered.

ACKNOWLEDGMENTS

Supported by NSF award MCB-1021557. We are grateful for many useful discussions with Profs Sebastian Doniach, Dilano Saldin, Henry Chapman, and Drs Andrew Martin, Daniel Ratner, and Peter Zwart.

APPENDIX

Appendix A: SAXS variance

From equation equation 3 we have the photon counts for the kth N-particle diffraction pattern

$$n_k(\boldsymbol{q}) = \sum_{\alpha=1}^{N} n(\boldsymbol{q}, \omega_\alpha^k) \ . \tag{A1}$$

The mean intensity profile is

$$\begin{aligned} I(\boldsymbol{q}) &= \langle n_k(\boldsymbol{q}) \rangle_k & \text{(A2)} \\ &= \sum_{\alpha=1}^{N} \left\langle n(\boldsymbol{q}, \omega_\alpha^k) \right\rangle_k & \text{(A3)} \end{aligned}$$

which, in the limit of large M, becomes the integral over continuous orientational distribution with probability $\eta(\omega)$ for orientation ω

$$\begin{aligned} I(\boldsymbol{q}) &\to \sum_{\alpha=1}^{N} \int_{\omega_\alpha} d\omega_\alpha \eta(\omega_\alpha) \sum_{n=0}^{\infty} np(n; \bar{n}(\boldsymbol{q}, \omega_\alpha)) & \text{(A4)} \\ &= N \left\langle \bar{n}(\boldsymbol{q}, \omega) \right\rangle_\omega & \text{(A5)} \end{aligned}$$

where we have used the moments in equations 5 and 6. Similarly, the variance will approach

$$
\begin{aligned}
\sigma^2_{I_N}(\boldsymbol{q}) &= \sum_{\alpha,\beta=1}^{N} \left\langle n(\boldsymbol{q},\omega^k_\alpha) n(\boldsymbol{q},\omega^k_\beta)\right\rangle_k - \left(\sum_{\alpha=1}^{N} \left\langle n(\boldsymbol{q},\omega^k_\alpha)\right\rangle_k\right)^2 && \text{(A6)}\\
&\to \sum_{\alpha=1}^{N} \int_{\omega_\alpha} \eta(\omega_\alpha) d\omega_\alpha \left[\sum_{n=1}^{N} n^2 p(n;\bar{n}(\boldsymbol{q},\omega_\alpha))\right] && \text{(A7)}\\
&\quad + \sum_{\alpha\neq\beta=1}^{N} \int_{\omega_\alpha}\int_{\omega_\beta} \eta(\omega_\alpha)\eta(\omega_\beta) d\omega_\alpha d\omega_\beta \left[\sum_{n=1}^{N} np(n;\bar{n}(\boldsymbol{q},\omega_\alpha))\right]\left[\sum_{n=1}^{N} np(n;\bar{n}(\boldsymbol{q},\omega_\beta))\right]\\
&\quad - N^2 \left\langle n(\boldsymbol{q},\omega)\right\rangle^2_\omega\\
&= N\left[\left\langle \bar{n}^2(\boldsymbol{q},\omega)\right\rangle_\omega + \left\langle \bar{n}(\boldsymbol{q},\omega)\right\rangle_\omega\right] + N(N-1)\left\langle \bar{n}(\boldsymbol{q},\omega)\right\rangle^2_\omega && \text{(A8)}\\
&\quad - N^2 \left\langle n(\boldsymbol{q},\omega)\right\rangle^2_\omega\\
&= N\left[\left[\left\langle \bar{n}(\boldsymbol{q},\omega)^2\right\rangle_\omega - \left\langle \bar{n}(\boldsymbol{q},\omega)\right\rangle^2_\omega\right] + \left\langle \bar{n}(\boldsymbol{q},\omega)\right\rangle_\omega\right] && \text{(A9)}
\end{aligned}
$$

The explicit representation of $\int_\omega p(\omega)d\omega$ depends on the degree of rotational freedom given to the particles. For instance, if a uniform distribution of orientations in three-dimensions is allowed, then this integral may be written as

$$\int_\omega \eta(\omega)d\omega = \frac{1}{8\pi^2}\int_{-\pi}^{\pi} d\alpha \int_0^{\pi} \sin\beta d\beta \int_{-\pi}^{\pi} d\gamma \cdot \qquad \text{(A10)}$$

where α, β, γ are the three Euler angles specified by ω.

Appendix B: CFSAXS variance

The fluctuation correlation function is

$$\begin{aligned}
\widetilde{C}_N(\boldsymbol{q}_1, \boldsymbol{q}_2) &\equiv \langle \delta n_k(\boldsymbol{q}_1) \delta n_k(\boldsymbol{q}_2) \rangle_k && \text{(B1)} \\
&= \left\langle \left[n_k(\boldsymbol{q}_1) - \langle n_j(\boldsymbol{q}_1) \rangle_j \right] \left[n_k(\boldsymbol{q}_2) - \langle n_j(\boldsymbol{q}_2) \rangle_j \right] \right\rangle_k && \text{(B2)} \\
&= \langle n_k(\boldsymbol{q}_1) n_k(\boldsymbol{q}_2) \rangle_k - \langle n_k(\boldsymbol{q}_1) \rangle_k \langle n_k(\boldsymbol{q}_2) \rangle_k && \text{(B3)} \\
&\to N \langle \bar{n}(\boldsymbol{q}_1, \omega) \bar{n}(\boldsymbol{q}_2, \omega) \rangle_\omega - N \langle \bar{n}(\boldsymbol{q}_1, \omega) \rangle_\omega \langle \bar{n}(\boldsymbol{q}_2, \omega) \rangle_\omega && \text{(B4)} \\
&= N \langle \delta \bar{n}(\boldsymbol{q}_1, \omega) \delta \bar{n}(\boldsymbol{q}_2, \omega) \rangle_\omega \ . && \text{(B5)}
\end{aligned}$$

The variance is

$$\sigma^2_{\widetilde{C}_N}(\boldsymbol{q}_1, \boldsymbol{q}_2) = \left\langle (\delta n_k(\boldsymbol{q}_1) \delta n_k(\boldsymbol{q}_2))^2 \right\rangle_k - \langle \delta n_k(\boldsymbol{q}_1) \delta n_k(\boldsymbol{q}_2) \rangle_k^2 \ . \qquad \text{(B6)}$$

Upon inspection of the first term on the right-hand side

$$\sum_{\alpha,\beta,\mu,\nu=1}^{N} \left\langle \delta n(\boldsymbol{q}_1, \omega_\alpha^k) \delta n(\boldsymbol{q}_2, \omega_\beta^k) \delta n(\boldsymbol{q}_1, \omega_\mu^k) \delta n(\boldsymbol{q}_2, \omega_\nu^k) \right\rangle_k \qquad \text{(B7)}$$

we see that there is one term where $\alpha = \beta = \mu = \nu$, and then there are three kinds of terms where two pairs of the indices are equal (e.g. $\alpha = \beta \neq \mu = \nu$), but not equal to each other. The remaining terms are products with terms like $\langle \delta n_k(\boldsymbol{q}) \rangle_k$ which approach zero in the large M limit, so that we can write

$$\begin{aligned}
\sigma^2_{\widetilde{C}_N} \to\ & N \left\langle [\delta n(\boldsymbol{q}_1, \omega_k) \delta n(\boldsymbol{q}_2, \omega_k)]^2 \right\rangle_k - N^2 \langle \delta n(\boldsymbol{q}_1, \omega_k) \delta n(\boldsymbol{q}_2, \omega_k) \rangle_k^2 && \text{(B8)} \\
& + N(N-1) \left[2 \langle \delta n(\boldsymbol{q}_1, \omega_k) \delta n(\boldsymbol{q}_2, \omega_k) \rangle_k^2 + \left\langle \delta n(\boldsymbol{q}_1, \omega_k)^2 \right\rangle_k \left\langle \delta n(\boldsymbol{q}_2, \omega_k)^2 \right\rangle_k \right] \\
=\ & N \left[\left\langle [\delta n(\boldsymbol{q}_1, \omega_k) \delta n(\boldsymbol{q}_2, \omega_k)]^2 \right\rangle_k - \langle \delta n(\boldsymbol{q}_1, \omega_k) \delta n(\boldsymbol{q}_2, \omega_k) \rangle_k^2 \right] && \text{(B9)} \\
& + N(N-1) \left[\langle \delta n(\boldsymbol{q}_1, \omega_k) \delta n(\boldsymbol{q}_2, \omega_k) \rangle_k^2 + \left\langle \delta n^2(\boldsymbol{q}_1, \omega_k) \right\rangle_k \left\langle \delta n^2(\boldsymbol{q}_2, \omega_k) \right\rangle_k \right]
\end{aligned}$$

Expanding terms further we have

$$\begin{aligned}
\sigma^2_{\widetilde{C}_N} \to N\big[& \left\langle n^2(\boldsymbol{q}_1,\omega_k) n^2(\boldsymbol{q}_2,\omega_k)\right\rangle_k - \left\langle n(\boldsymbol{q}_1,\omega_k) n(\boldsymbol{q}_2,\omega_k)\right\rangle_k^2 \\
& -2\left\langle n(\boldsymbol{q}_1,\omega_k) n^2(\boldsymbol{q}_2,\omega_k)\right\rangle_k \left\langle n(\boldsymbol{q}_1,\omega_k)\right\rangle_k \\
& -2\left\langle n^2(\boldsymbol{q}_1,\omega_k) n(\boldsymbol{q}_2,\omega_k)\right\rangle_k \left\langle n(\boldsymbol{q}_2,\omega_k)\right\rangle_k \\
& +6\left\langle n(\boldsymbol{q}_1,\omega_k) n(\boldsymbol{q}_2,\omega_k)\right\rangle_k \left\langle n(\boldsymbol{q}_1,\omega_k)\right\rangle_k \left\langle n(\boldsymbol{q}_2,\omega_k)\right\rangle_k \\
& +\left\langle n(\boldsymbol{q}_1,\omega_k)\right\rangle_k^2 \left\langle n^2(\boldsymbol{q}_2,\omega_k)\right\rangle_k \\
& +\left\langle n^2(\boldsymbol{q}_1,\omega_k)\right\rangle_k \left\langle n(\boldsymbol{q}_2,\omega_k)\right\rangle_k^2 - 4\left\langle n(\boldsymbol{q}_1,\omega_k)\right\rangle_k^2 \left\langle n(\boldsymbol{q}_2,\omega_k)\right\rangle_k^2 \big] \\
& +(N^2-N)\big[\left\langle n(\boldsymbol{q}_1,\omega_k) n(\boldsymbol{q}_2,\omega_k)\right\rangle_k^2 \\
& -2\left\langle n(\boldsymbol{q}_1,\omega_k) n(\boldsymbol{q}_2,\omega_k)\right\rangle_k \left\langle n(\boldsymbol{q}_1,\omega_k)\right\rangle_k \left\langle n(\boldsymbol{q}_2,\omega_k)\right\rangle_k \\
& -\left\langle n(\boldsymbol{q}_1,\omega_k)\right\rangle_k^2 \left\langle n^2(\boldsymbol{q}_2,\omega_k)\right\rangle_k - \left\langle n^2(\boldsymbol{q}_1,\omega_k)\right\rangle_k \left\langle n(\boldsymbol{q}_2,\omega_k)\right\rangle_k^2 \\
& +\left\langle n^2(\boldsymbol{q}_1,\omega_k)\right\rangle_k \left\langle n^2(\boldsymbol{q}_2,\omega_k)\right\rangle_k + 2\left\langle n(\boldsymbol{q}_1,\omega_k)\right\rangle_k^2 \left\langle n(\boldsymbol{q}_2,\omega_k)\right\rangle_k^2 \big]\ .
\end{aligned} \tag{B10}$$

Finally, upon inserting the moments in equations 5 and 6, we arrive at the variance expressed in equation 25.

Appendix C: Intensity statistics

We drop the $\boldsymbol{q}$ dependence for brevity. Letting $\phi_{ij} = \boldsymbol{q}\cdot(\boldsymbol{r}_i - \boldsymbol{r}_j)$, and noting that uncorrelated terms $i \neq j$ vanish upon averaging over a large number of random phases, the mean value is

$$\langle \Theta \rangle = \left\langle \sum_{i,j=1}^{m} f_i f_j e^{i\phi_{ij}} \right\rangle \tag{C1}$$

$$= \left\langle \sum_{i=j=1}^{m} f_i^2 \right\rangle + 0 \tag{C2}$$

$$= m\left\langle f^2 \right\rangle \tag{C3}$$

where $\langle f^2 \rangle$ is the number-weighted average of f^2. Similarly, by noting that $\phi_{ii} = 0$ and $\phi_{ij} = -\phi_{ji}$, we arrive at the mean squared value

$$\left\langle \Theta^2 \right\rangle = \left\langle \sum_{i,j,k,l=1}^{m} f_i f_j f_k f_l e^{i(\phi_{ij}+\phi_{kl})} \right\rangle \tag{C4}$$

$$= \sum_{i=j,k,l} f_i^2 f_k f_l e^{i\phi_{kl}} + \sum_{i=l\neq k=j} f_i^2 f_j^2 + 0 \tag{C5}$$

$$= 2m^2 \left\langle f^2 \right\rangle^2 \ . \tag{C6}$$

Finally, we use the previous results to write the variance

$$\left\langle \delta\Theta^2 \right\rangle = \left\langle \Theta^2 \right\rangle - \left\langle \Theta \right\rangle^2 = \left\langle \Theta \right\rangle^2 \ . \tag{C7}$$

[1] Z. Kam, Macromolecules, **10**, 927 (1977).

[2] D. K. Saldin, H. C. Poon, V. L. Shneerson, M. Howells, H. N. Chapman, R. A. Kirian, K. E. Schmidt, and J. C. H. Spence, Physical Review B, **81**, 174105 (2010).

[3] D. K. Saldin, H. C. Poon, M. J. Bogan, S. Marchesini, D. A. Shapiro, R. A. Kirian, U. Weierstall, and J. C. H. Spence, Physical Review Letters, **106**, 115501 (2011).

[4] V. Elser, arXiv:1007.3777v1 (2010).

[5] W. A. Barletta, J. Bisognano, J. N. Corlett, P. Emma, Z. Huang, K. J. Kim, R. Lindberg, J. B. Murphy, G. R. Neil, D. C. Nguyen, C. Pellegrini, R. A. Rimmer, F. Sannibale, G. Stupakov, R. P. Walker, and A. A. Zholents, Nuclea Instruments & Methods in Physics Research Section A–Accelerators Spectrometers Detectors and Associated Equipment, **618**, 69 (2010).

[6] M. M. Seibert, T. Ekeberg, F. R. N. C. Maia, M. Svenda, J. Andreasson, O. Jonsson, D. Odic, B. Iwan, A. Rocker, D. Westphal, M. Hantke, D. P. DePonte, A. Barty, J. Schulz, L. Gumprecht, N. Coppola, A. Aquila, M. N. Liang, T. A. White, A. Martin, C. Caleman, S. Stern, C. Abergel, V. Seltzer, J. M. Claverie, C. Bostedt, J. D.

Bozek, S. Boutet, A. A. Miahnahri, M. Messerschmidt, J. Krzywinski, G. Williams, K. O. Hodgson, M. J. Bogan, C. Y. Hampton, R. G. Sierra, D. Starodub, I. Andersson, S. Bajt, M. Barthelmess, J. C. H. Spence, P. Fromme, U. Weierstall, R. Kirian, M. Hunter, R. B. Doak, S. Marchesini, S. P. Hau-Riege, M. Frank, R. L. Shoeman, L. Lomb, S. W. Epp, R. Hartmann, D. Rolles, A. Rudenko, C. Schmidt, L. Foucar, N. Kimmel, P. Holl, B. Rudek, B. Erk, A. Homke, C. Reich, D. Pietschner, G. Weidenspointner, L. Struder, G. Hauser, H. Gorke, J. Ullrich, I. Schlichting, S. Herrmann, G. Schaller, F. Schopper, H. Soltau, K. U. Kuhnel, R. Andritschke, C. D. Schroter, F. Krasniqi, M. Bott, S. Schorb, D. Rupp, M. Adolph, T. Gorkhover, H. Hirsemann, G. Potdevin, H. Graafsma, B. Nilsson, H. N. Chapman, and J. Hajdu, Nature, **470**, 78 (2011).

[7] N. D. Loh, M. J. Bogan, V. Elser, A. Barty, S. Boutet, S. Bajt, J. Hajdu, T. Ekeberg, F. R. N. C. Maia, J. Schulz, M. M. Seibert, B. Iwan, N. Timneanu, S. Marchesini, I. Schlichting, R. L. Shoeman, L. Lomb, M. Frank, M. Liang, and H. N. Chapman, Physical Review Letters, **104**, 239902 (2010).

[8] M. Altarelli, R. P. Kurta, and I. A. Vartanyants, Physical Review B, **82**, 104207 (2010).

[9] J. W. Goodman, *Statistical Optics* (Wiley-Interscience, 2000).

[10] P. R. Bevington and K. Robinson, *Data Reduction and Error Analysis for the Physical Sciences, 2nd Edition* (McGraw-Hill, 1992).

[11] D. K. Saldin, V. L. Shneerson, R. Fung, and A. Ourmazd, Journal of Physics- Condensed Matter, **21**, 134014 (2009).

[12] J. D. Jackson, *Classical Electrodynamics, Third Edition* (Wiley, New York, 1999).

[13] H. M. Berman, T. Battistuz, T. N. Bhat, W. F. Bluhm, P. E. Bourne, K. Burkhardt, L. Iype, S. Jain, P. Fagan, J. Marvin, D. Padilla, V. Ravichandran, B. Schneider, N. Thanki, H. Weissig, J. D. Westbrook, and C. Zardecki, Acta Crystallographica

Section D-Biological Crystallography, **58**, 899 (2002), ISSN 0907-4449.

[14] J. M. Holton and K. A. Frankel, Acta Crystallographica Section D-Biological Crystallography, **66**, 393 (2010).

[15] G. Hura, J. M. Sorenson, R. M. Glaeser, and T. Head-Gordon, Journal of Chemical Physics, **113**, 9140 (2000).

[16] H. Stuhrmann, ed., *Uses of synchrotron radiation in biology* (Academic Press, 1982).

[17] R. L. Owen, E. Rudino-Pinera, and E. F. Garman, Proceedings of the National National Academy of Sciences of the United States of America, **103**, 4912 (2006).

[18] J. M. Holton, Journal of Synchrotron Radiation, **16**, 133 (2009).

[19] S. Boutet and G. J. Williams, New Journal of Physics, **12**, 035024 (2010).

[20] R. Fung, V. Shneerson, D. K. Saldin, and A. Ourmazd, Nature Physics, **5**, 64 (2009).

[21] N. T. D. Loh and V. Elser, Physical Review E, **80** (2009).

[22] U. Weierstall, R. B. Doak, J. C. H. Spence, D. Starodub, D. Shapiro, P. Kennedy, J. Warner, G. G. Hembree, P. Fromme, and H. N. Chapman, Experiments in Fluids, **44**, 675 (2008).

[23] D. P. DePonte, U. Weierstall, K. Schmidt, J. Warner, D. Starodub, J. C. H. Spence, and R. B. Doak, Journal of Physics D–Applied Physics, **41**, 195505 (2008).

[24] H. N. Chapman, P. Fromme, A. Barty, T. A. White, R. A. Kirian, A. Aquila, M. S. Hunter, J. Schulz, D. P. DePonte, U. Weierstall, R. B. Doak, F. R. N. C. Maia, A. V. Martin, I. Schlichting, L. Lomb, N. Coppola, R. L. Shoeman, S. W. Epp, R. Hartmann, D. Rolles, A. Rudenko, L. Foucar, N. Kimmel, G. Weidenspointner, P. Holl, M. N. Liang, M. Barthelmess, C. Caleman, S. Boutet, M. J. Bogan, J. Krzywinski, C. Bostedt, S. Bajt, L. Gumprecht, B. Rudek, B. Erk, C. Schmidt, A. Homke, C. Reich, D. Pietschner, L. Struder, G. Hauser, H. Gorke, J. Ullrich, S. Herrmann, G. Schaller, F. Schopper, H. Soltau, K. U. Kuhnel, M. Messerschmidt, J. D. Bozek, S. P. Hau-Riege, M. Frank, C. Y. Hampton, R. G. Sierra, D. Starodub, G. J. Williams, J. Ha-

jdu, N. Timneanu, M. M. Seibert, J. Andreasson, A. Rocker, O. Jonsson, M. Svenda, S. Stern, K. Nass, R. Andritschke, C. D. Schroter, F. Krasniqi, M. Bott, K. E. Schmidt, X. Y. Wang, I. Grotjohann, J. M. Holton, T. R. M. Barends, R. Neutze, S. Marchesini, R. Fromme, S. Schorb, D. Rupp, M. Adolph, T. Gorkhover, I. Andersson, H. Hirsemann, G. Potdevin, H. Graafsma, B. Nilsson, and J. C. H. Spence, Nature, **470**, 73 (2011).

[25] M. V. Petoukhov and D. I. Svergun, Current Opinion In Structural Biology, **17**, 562 (2007).