# Aspects of phase retrieval for diffraction data from very small crystals

Joe Chen[1], John Spence[2] and Rick Millane[1]
[1]Computational Imaging Group, Department of Electrical and Computer Engineering,
University of Canterbury, Christchurch, New Zealand
[2]Department of Physics, Arizona State University, Tempe, AZ 85287, USA
Email: rick.millane@canterbury.ac.nz

***Abstract*—Nanocrystallography is a form of X-ray coherent diffraction imaging that utilises a stream of small crystals of the biological assembly under study. These small crystallites are termed nanocrystals in contrast to the large crystals used in conventional X-ray crystallography. Diffraction snapshots of individual nanocrystals can be obtained using femtosecond pulses from an X-ray free-electron laser. A key advantage of nanocrystallography is that many membrane proteins for example do not easily form large crystals, but do form nanocrystals and thus crystallography techniques can be applied to image these previously inaccessible structures. Here we present results from a study of the performance of phase retrieval algorithms in the context of nanocrystallography. We use a signal-to-noise ratio that is modulated by the interference function averaged over a distribution of crystallite sizes and investigate convergence of the difference map algorithm for phase retrieval as a function of the overall SNR. The results give a picture of the noise levels and crystallite sizes that can be tolerated in nanocrystallography.**

## I. Introduction

X-ray crystallography is a technique for imaging single biological molecules using X-rays diffracted from a cyrstalline specimen. Since the diffracted X-ray amplitude is the Fourier transform of the electron density in the crystal (from which the positions of the atoms in the molecule can be inferred), in principle this electron density can be recovered by the inverse Fourier transformation. There are three difficulties however. First, only the amplitude, but not the phase, of the diffracted X-rays can be measured - this is an example of what is often called a "phase problem" where the Fourier phase information is lost and needs to be retrieved. Second, since the crystal is periodic, its Fourier transform, i.e. the diffracted X-ray amplitudes, are discrete in Fourier space and we observe the so-called Bragg reflections or Bragg peaks. These peaks undersample the amplitude of the diffraction pattern and so the phase problem is underdetermined. The third difficulty of X-ray crystallography is simply that for some molecular assemblies the sample of interest cannot be easily crystallised.

The first obstacle is addressed by various techniques that make use of either additional experimental data or information on a related molecule [1]. The second and third difficulties of coherent X-ray diffraction imaging - the undersampling of the diffraction pattern and the crystallisation problem, have spurred much of the effort to extend traditional crystallography techniques to either single molecules or to crystals with a small number of unit cells, termed "nanocrystals". In contrast to larger crystals with many unit cells, small crystals avoid the need to crystallise the molecular sample in question and provide information on the continuous diffraction pattern from a single molecule.

It is not until very recently that the imaging of nanocrystals has been practically achieved. This is because without the repetition of a large number of unit cells from conventional crystals, the diffracted signal levels drop significantly and the detection process is overwhelmed by noise. The simplest way to overcome this problem is to increase the power of the incident X-ray beam, giving a better signal-to-noise ratio (SNR). However in doing so, the biological specimen can sustain severe radiation damage, resulting in a change in its intrinsic structure; or worse, its destruction. Intense, ultrafast X-ray pulses created in accelerator facilities such as the Linac Coherent Light Source (LCLS) at the Stanford Linear Accelerator Center in the US provide a solution to this conundrum. These so-called X-ray free-electron lasers (XFEL) generate pulses of X-rays lasting only around 100fs but have an average of $10^{12}$ photons per pulse [2]. With these intense femtosecond X-ray pulses, appropriate signal levels can be obtained for nanocrystals before radiation damage occurs [3].

The potential of this approach has been demonstrated by assembling structure factor data from nanocrystals of photosystem I and showing that they are consistent with the known structure [4-6]. Since diffraction from nanocrystals gives information between the Bragg reflections, one approach is to collect such data and perform structure determination ab initio using a phase retrieval algorithm. The feasibility of this method has been demonstrated successfully by recovering the molecular transform using simulated nanocrystal diffraction data from the small protein alpha-conotoxin PnIB [7]. Here we explore the effect of noise on phase retrieval in nanocrystalline coherent X-ray imaging.

## II. Diffraction by a collection of nanocrystals

Consider the 1-dimensional case where the molecule's electron density is denoted by $f(x)$ and is repeated within unit cells of width $a$, $N$ times along the $x$ direction as shown in Figure 1. The result is a 1-dimensional crystal with a density $g(x)$ composed of repeated copies of $f(x)$ which can be

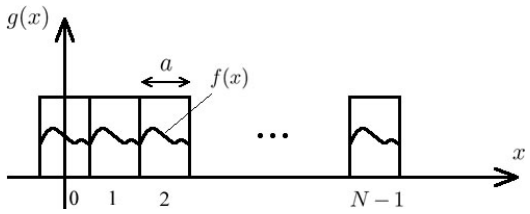



Fig. 1. Schematic view of a 1-dimensional nanocrystal.

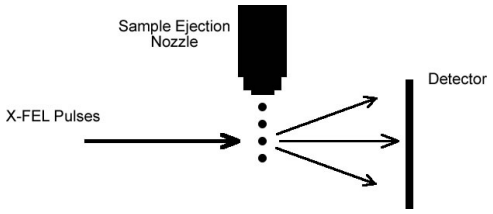



Fig. 2. Experimental set-up of diffraction pattern measurements from a stream of nanocyrstals.

expressed as a convolution with a train of delta functions,

$$g(x) = f(x) \otimes \sum_{h=0}^{N-1} \delta(x - ha). \tag{1}$$

The resulting complex X-ray diffraction amplitude is the Fourier transform of $g(x)$; namely, $G(u)$. It is easily shown that

$$G(u) = F(u) \frac{\sin(\pi a u N)}{\sin(\pi a u)} e^{-j\pi a u(N-1)}, \tag{2}$$

where $F(u)$ is what we are interested in - the transform of the molecular density in question. We write equation (2) in the form

$$G(u) = F(u) S_N(u), \tag{3}$$

where

$$S_N(u) = \frac{\sin(\pi a u N)}{\sin(\pi a u)} e^{-j\pi a u(N-1)} \tag{4}$$

is called the interference function.

In practice, nanocrystals of the biological sample of interest are delivered to the high intensity femtosecond X-ray pulses through an ejector nozzle as shown conceptually in Figure 2. The intensity of a single diffraction pattern is the squared magnitude of the transform of the crystal, $|G(u)|^2$. If we denote $P(N)$ as the probability density function for the sizes of the ejected nanocrystals, then the diffraction intensity averaged over a large collection of individual diffraction patterns is

$$I(u) = \sum_{\text{all N}} P(N) \left|G(u)\right|^2, \tag{5}$$

giving

$$I(u) = \left|F(u)\right|^2 \left|Q(u)\right|^2, \tag{6}$$

where

$$Q(u) = \sqrt{\sum_{\text{all N}} P(N) \left(\frac{\sin(\pi a u N)}{\sin(\pi a u)}\right)^2} \tag{7}$$

is the averaged interference function.

The recorded diffraction patterns are first analysed to select those that derive from a single nanocrystal. The interference function around each Bragg reflection can then be averaged to obtain $Q(u)$ for the particular crystallite size distribution of the particular experiment. The desired molecular transform amplitude $|F(u)|$ is therefore calculated from the diffraction data $I(u)$ as $|F(u)| = \sqrt{I(u)}/Q(u)$.

The measured intensity in practice will always contain some noise $n(u)$ and so assuming this noise is additive, the estimate of the magnitude of the molecular transform is

$$|F(u)|_{\text{meas}} = |F(u)| + \frac{n(u)}{Q(u)}. \tag{8}$$

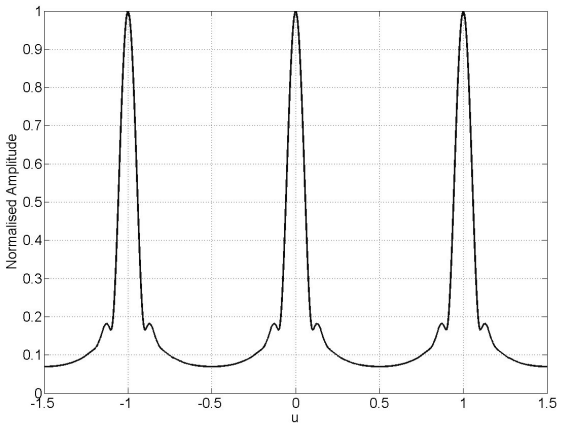
The measurement noise is therefore amplified by the inverse of the averaged interference function $Q(u)$.

Figure 3(a) shows the normalised amplitude of the averaged interference function for a collection of nanocrystals with a Gaussian size distribution. The mean crystallite size is 10 unit cells and the standard deviation is 2 unit cells. It can be seen that the peaks of $Q(u)$ centre around integer multiples of the inverse unit cell width, $1/a$, where $a = 1$ in this case. These points are known as the reciprocal lattice points and this is what gives rise to the Bragg peaks seen in measured diffraction patterns. As the nanocrystals' sizes increase, these peaks become concentrated at the lattice points and the result is a discrete diffraction pattern. The data between the Bragg peaks are thus lost for large crystals. The noise amplification or $1/Q(u)$ is shown in Figure 3(b). Although the above description is in one dimension, it extends straightforwardly to three dimensions as is required in crystallography.

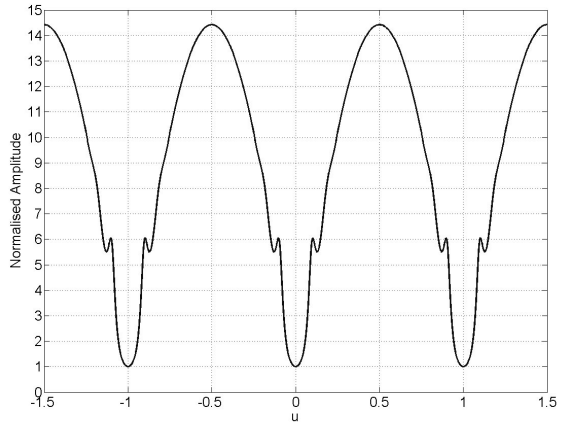
Thus, the sizes of the nanocrystals and their distribution determines the level of noise corruption in the measured data, which in turn poses a significant influence on the process of retrieving the Fourier phases in order to reconstruct the electron density of the molecule.

## III. Phase retrieval

To retrieve the Fourier phases when only the magnitudes are known is by itself not possible in principle as these two quantities are independent of one another. However in practice we can utilise further information specific to the problem at hand. For example the knowledge that all objects in practice are non-negative and of finite extent helps a great deal in constraining the problem to allow us to arrive at a unique solution of the phases [8]. The formulation of the problem then becomes that of trying to find the intersection between two constraint sets: the set of all possible objects that have the measured Fourier modulus and the set of all possible objects that are contained within the given finite-extent region termed the support. The object in our case is the electron density of the biological molecule in question. In this paper, the terms "objects", "images" and "densities" shall be used interchangeably and assumed to have the same meaning.




(a)




(b)

Fig. 3. The averaged interference function and its inverse for a normally distributed collection of nanocrystals with a mean crystallite size of 10 and a standard deviation of 2. (a) The normalised function $Q(u)$ and (b) $1/Q(u)$.

Iterative projection algorithms (IPAs) are specifically designed to tackle these constraint satisfaction problems. They seek out the intersection between two constraint sets by iteratively searching through the multi-dimensional space that the problem resides in using operators called projections and a solution can usually be found. It is convenient to formulate IPAs as operations on vectors in an $n$-dimensional metric space. A vector $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ in this space represents an $n$-pixel image where each of its $n$ components correspond to the value of a particular pixel. These $n$ coordinates form an orthogonal basis for the corresponding $n$-dimensional space.

A projection $P_A$ is then defined as an operation that takes a point in this metric space to the closest point on some set $A$ in this space. For example, the support constraint in phase retrieval requires the density in question to be zero outside a support region $S$. The projection operator, $P_S$, that achieves this is obtained by setting all values of the image outside the support to zero. So for all $i = 1, \cdots, n$ we have,

$$P_S\mathbf{x} = \begin{cases} x_i & \text{if } x_i \in S \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

Similarly, in the Fourier domain, the projection $P_M$ sets the magnitude of a complex number to that of the measured magnitude whilst leaving the phases unchanged. The set of all complex numbers that have the same magnitude defines a circle on the complex plane, therefore the projection involves moving the target point radially to the closest point on the circle. In practice, the Fourier transform and the inverse Fourier transform required to move the image back-and-forth between the object domain and the frequency domain is incorporated into the $P_M$ projection operator itself.

At the $n$th iteration of an IPA, the current iterate $\mathbf{x}_n$ is updated to form the next iterate $\mathbf{x}_{n+1}$. The update rule of the algorithm is a combination of the projections $P_S$, $P_M$ and the identity operator $I$, where different IPAs are distinguished by different update rules.

One of the most effective IPAs to date was proposed by Elser [9]. This so called difference map (DM) algorithm involves the use of three independent parameters $\gamma_S$, $\gamma_M$ and $\beta$ and takes the form

$$\mathbf{x}_{n+1} = \big(I + \beta\left(P_S F_M - P_M F_S\right)\big)\mathbf{x}_n \tag{10}$$

where

$$F_S = (1 + \gamma_M) P_S - \gamma_M I,$$

$$F_M = (1 + \gamma_S) P_M - \gamma_S I.$$

$F_S$ and $F_M$ are called "relaxed projections". Elser suggested as possible values for optimum performance, the values $\gamma_S = -1/\beta$ and $\gamma_M = 1/\beta$. For the DM algorithm, the iterate itself is not an estimate of the solution and an estimate of the solution $\mathbf{x}_{\text{soln}}$ can be obtained as

$$\mathbf{x}_{\text{soln}} = P_S F_M \mathbf{x}_n.$$

The DM algorithm has good global convergence properties and is thus the algorithm we chose to apply to our nanocrystallography phase retrieval problem.

## IV. Simulation Setup

Simulations are carried out in two dimensions. A 2-dimensional section of the membrane protein erythrocyte aquaporin 1 (AQP1) [10] as shown in the top left of Figure 4(a) is used as the test object in our nanocrystallography phase retrieval simulations using MATLAB. This section of a single AQP1 molecule is embedded in a 71×71 array and zero-padded to 142×142. As a result of this two-times oversampling, the diffraction data lie at the Bragg peaks and halfway in between.

The Fourier transform of the image is calculated and used as the true diffraction magnitudes. To simulate the diffraction data measured in practice, this true magnitude is corrupted by additive Gaussian noise with zero mean and amplified by

the inverse averaged interference function in 2-dimensions in accord with equation (8).

The mask for the support projection of the DM algorithm is a circle slightly larger than the circular image of the protein itself. The nanocrystals in our simulations are assumed to be 2-dimensional and square. A normal distribution with a standard deviation one-fifth of the mean is used as the probability distribution for the sizes of the nanocrystals. The origin term of the Fourier magnitude is also discarded to simulate the effect of the X-ray beam-stop.

## V. Results

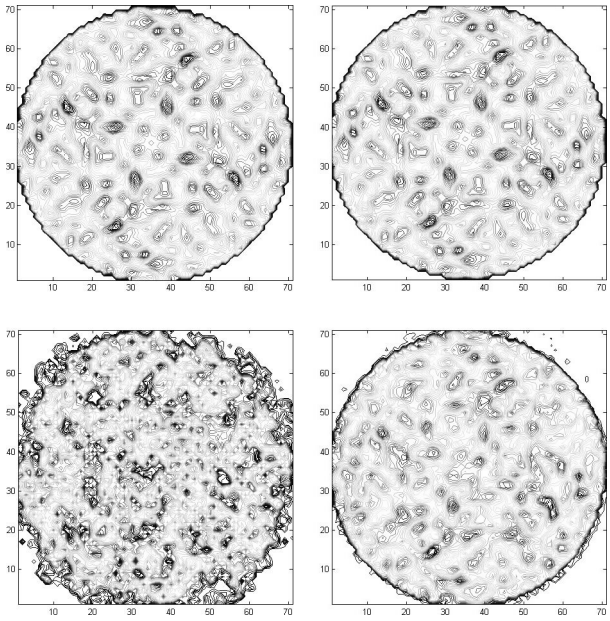
Results after the application of 1000 iterations of the DM algorithm for the case of a collection of 2-dimensional nanocrystals with a mean crystallite size of 10 are shown in Figure 4. Two metrics are used to gauge the convergence of the algorithm. The first is the difference between the next iterate and the current iterate, denoted by $\Delta = ||\mathbf{x}_{n+1} - \mathbf{x}_n||$ where the norm used is the Euclidean norm. Small values of $\Delta$ signal the convergence of the iterate and the consequent arrival at a fixed point. The second metric is the so-called R-factor and is the standard quantity used in crystallography to measure the difference between the diffraction data and that predicted by the solution. It is defined as

$$\mathrm{R} = \frac{\Sigma_u \big| \, |F(u)|_{\mathrm{meas}} - |\hat{F}(u)| \big|}{\Sigma_u \, |F(u)|_{\mathrm{meas}}}$$

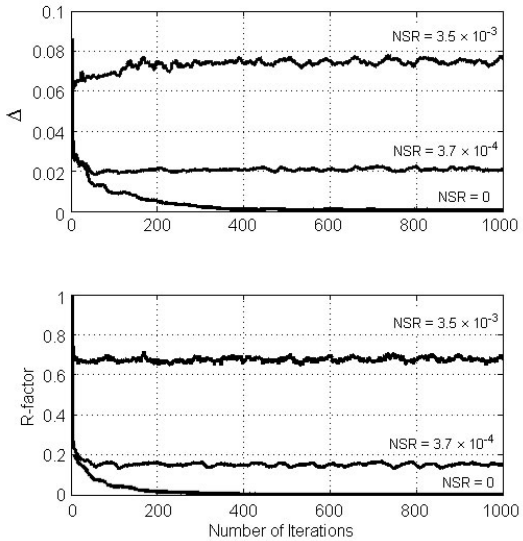
where $|F(u)|_{\mathrm{meas}}$ is the measured amplitude and $|\hat{F}(u)|$ is the magnitude of the estimate of the solution. In general, an R-factor of less than 0.25 indicates an acceptable reconstruction.

The top right-hand corner of figure 4(a) shows the reconstruction for the noiseless case acting as the control scenario for our simulation. We see that both error metrics drop rapidly with increasing iteration and the final R-factor is approximately $1 \times 10^{-3}$ indicating a perfect reconstruction. This also signals the correct functioning of the algorithm. With a slightly elevated noise level, the reconstructed image is shown in the bottom right of the same figure. The noise-to-signal-ratio (NSR) for this case is around $3.7 \times 10^{-4}$, however after amplification, the effective NSR is around 4%. The reconstruction is still fairly reasonable as indicated by the final R-factor value of 0.14. Lastly, a case where the noise level is too high for a good reconstruction is shown in the bottom left. The original NSR is $3.5 \times 10^{-3}$ whereas the NSR after amplification is around 30%. The final R-factor is 0.68.

As the mean crystallite size increases, the diffraction amplitudes between the Bragg reflections are further attenuated and the noise amplification increases. The overall noise amplification of all the data can be calculated as a function of the mean crystallite size and is shown in Figure 5. A quadratic increase in the noise amplification is seen as the mean size of the nanocrystals gets larger. Note that even for small crystallites, the noise amplification is still quite significant as a mean crystallite size of 10 gives around a hundred-fold increase in the noise level.



(a)




(b)

Fig. 4. Results for a nanocrystallography phase retrieval simulation. (a) Reconstructions obtained using the DM algorithm; clockwise from top-left is the original image, the noiseless case, NSR=$3.7\times10^{-4}$, and NSR=$3.5\times10^{-3}$. (b) Error metrics $\Delta$ and R-factor from the reconstructions in (a).

The simulations described above were repated for a range of NSRs and mean crystallite sizes and the error metrics obtained are plotted in Figure 6. These results are plotted for the iteration where the R-factor is a minimum. A distinct hyperbolic region can be seen in the contour plots. This is the convergence region of our particular nanocrystallography phase retrieval problem. The large flat region of the R-factor indicates that the metric never falls below its initial value during the 1000 iterations that the DM algorithm ran for and signals the breakdown in the convergence of phase retrieval from small crystals for these parameter values.

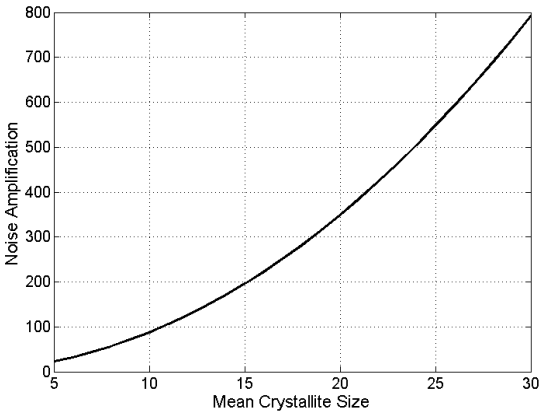



Fig. 5. Noise amplification with respect to the mean crystallite size in a collection of normally distributed nanocrystals with a standard deviation one-fifth of the mean.

Inspection of Figure 6 also shows that convergence of the algorithm, $\Delta$, deteriorates smoothly as the noise levels and crystallite sizes increase. However, the R-factor results show that acceptable reconstructions are obtained only for quite low noise levels and small crystallite sizes.

## VI. Discussion

The diffraction from a nanocrystal corresponds to the molecular transform modulated by the transform of the crystallite shape function centred at the reciprocal lattice points. This means that the molecular transform can be measured between the reciprocal lattice points, albeit the transform is attenuated; or equivalently, the noise is amplified. The result is that the amplitudes between the Bragg reflections are measured at a lower signal-to-noise ratio than for the amplitudes measured at the Bragg reflections themselves.

Simulation results on the performance of phase retrieval in the presence of a variable noise of this kind show a hyperbolic convergence region as a function of mean crystallite size and measurement noise. Successful reconstruction is very sensitive to the noise level and crystallite size. However, these results are probably a little pessimistic since the simulations were conducted in two dimensions and the problem will be better determined in three dimensions. Furthermore, a greater degree of oversampling may improve the stability. In the case presented here the crystallographic symmetry was not taken into account. Application of this symmetry should improve the reconstructions. A Poisson noise model may be more appropriate than a Gaussian one considering the low photon count in these experiments. The effect of a weaker diffraction and hence a higher noise level for smaller nanocrystals has also not been incorporated. It is possible in addition that the large surface-to-volume ratio of small crystallites may introduce disorder that could produce variations in the interference function over the diffraction pattern.

In summary, the results presented here indicate that while promising, ab initio phasing in nanocrystallography will need to overcome significant noise sensitivity.

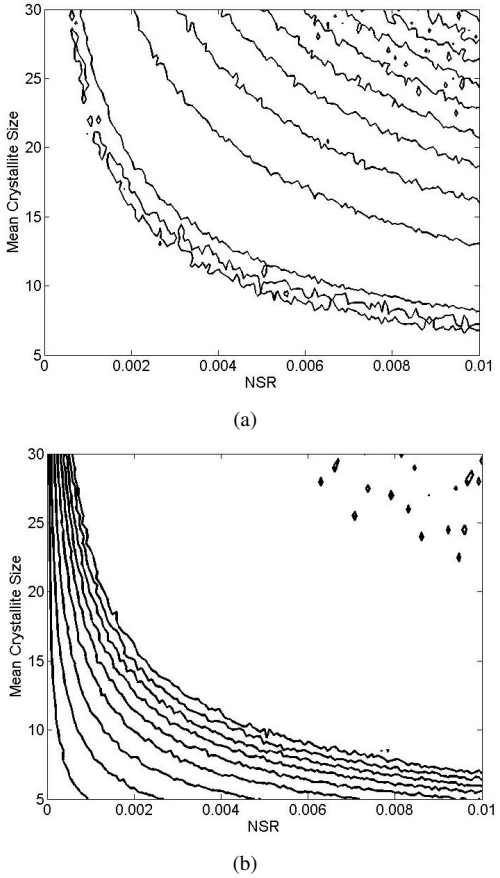



Fig. 6. Overall effect of varying the mean crystallite size and the noise level on the measured Fourier amplitude. (a) and (b) are contour plots of $\Delta$ and the R-factor, respectively. Contour levels are from 0.1 to 1.0 in steps of 0.1, increasing from the lower left to the upper right.

## Acknowledgment

The authors would like to thank Alok Mitra for providing the AQP1 electron density data used in our simulations.

## References

[1] P. Argos, "Protein crystallography in a molecular biophysics course," *American Journal of Physics*, vol. 45, pp.31-37, 1977.

[2] H. Chapman, "X-ray imaging beyond the limits," *Nature materials*, vol. 8, pp.299-301, April 2009.

[3] R. Neutze et al., "Potential for biomolecular imaging with femtosecond X-ray pulses," *Nature*, vol. 406, pp.752-757, 2000.

[4] H. Chapman et al., "Femtosecond X-ray protein nanocrystallography," *Nature*, vol. 470, pp.73-77, 2011.

[5] R. Kirian et al., "Femtosecond protein nanocrystallography - data analysis methods," *Optics Express*, vol. 18, no. 6, pp.5713-5723, 2010.

[6] R. Kirian et al., "Structure-factor analysis of femtosecond micro-diffraction patterns from protein nanocrystals," *Acta Cryst. A*, vol. 67, pp.131-140, 2011.

[7] J. Spence et al., "Phasing of coherent femtosecond X-ray diffraction from size-varying nanocrystals," *Optics Express*, vol. 19, no. 4, pp.2866-2873, 2011.

[8] R. Millane, "Phase retrieval in crystallography and optics," *J. Opt. Soc. Am. A*, vol. 7, no. 3, pp.394-411, 1990.

[9] V. Elser, "Phase retrieval by iterated projections," *J. Opt. Soc. Am. A*, vol. 20, no. 1, pp.40-55, 2003.

[10] A. Mitra et al., "Three-dimensional fold of the human AQP1 water channel determined at 4 A resolution by electron cyrstallography of two-dimensional crystals embedded in ice," *J. Mol. Biol.*, vol. 301, pp.369-387, 2000.