REVIEW

XFEL data analysis for structural biology

Haiguang Liu[1,*] and John C. H. Spence[2,*]

[1] Complex Systems Division, Beijing Computational Science Research Center, Beijing 100193, China
[2] Physics Department, Arizona State University, Tempe, AZ 85287, USA
[*] Correspondence: hgliu@csrc.ac.cn, spence@asu.edu

Received December 24, 2015; Revised February 10, 2016; Accepted February 23, 2016

X-ray Free Electron Lasers (XFELs) have advanced research in structure biology, by exploiting their ultra-short and bright X-ray pulses. The resulting "diffraction before destruction" experimental approach allows data collection to outrun radiation damage, a crucial factor that has often limited resolution in the structure determination of biological molecules. Since the first hard X-ray laser (the Linac Coherent Light Source (LCLS) at SLAC) commenced operation in 2009, serial femtosecond crystallography (SFX) has rapidly matured into a method for the structural analysis of nano- and micro-crystals. At the same time, single particle structure determination by coherent diffractive imaging, with one particle (such as a virus) per shot, has been under intense development. In this review we describe these applications of X-ray lasers in structural biology, with a focus particularly on aspects of data analysis for the computational research community. We summarize the key problems in data analysis and model reconstruction, and provide perspectives on future research using computational methods.

Keywords: X-ray Free Electron Laser; single particle scattering; serial crystallography; phase retrieval; orientation recovery

INTRODUCTION

High resolution structure determination from individual molecules without the need for crystal growth is an ultimate goal of structural biology research community. Growing high quality crystals is extremely difficult for many biomolecules that are biologically or medically significant, and may require years of effort. About 90% of structures deposited in the Protein Data Bank are determined using X-ray crystallography. The structural determination of large single molecules is possible in vitreous ice, without crystallization, at sub-nanometer resolution using cryo-electron microscopy (CryoEM). The X-ray laser, however, may offer an alternative approach, based on the "diffraction-before-destruction" mode, which in principle may allow room temperature structure determination at high spatial resolution free from radiation damage, in a native environment [1,2]. Preliminary experiments have produced hard X-ray snapshot diffraction patterns from individual viruses, in one case generating about 1.5 million elastically scattered X-rays from a single Mimivirus after being intercepted by the femtosecond X-ray pulse (70 fs pulse duration), allowing construction of a virus projection in 2D at about 32 nm resolution [3]. Furthermore, if single molecule diffraction method is coupled with specially designed pumping methods, conformational changes might be observed at high temporal resolution in time-resolved experiments. The production of a molecular movie should thus be feasible if the necessary experimental instruments can be established at X-ray Free Electron Laser (XFEL) facilities, including high brilliance X-ray pulses with a well defined sub-micron focus, stable sample delivery devices which generate little background, and fast X-ray detectors with a large dynamic range. Computational resources, both hardware and software, are equally important. The high repetition rates at XFEL facilities means high data acquisition rates, which are causing a data deluge challenge in photon science. For instance, at the Linac Coherent Light Source (LCLS), the X-ray pulses come at a frequency of 120 Hz, producing approximately 1 GB of raw data per second[1)] when two

[1)] Assuming each pixel requires 4 byte to store information and each detector is composed of 1 million pixels. Dual-detector system will produce $4\times 1\text{M}\times 120\times 2$ = 960 MB ~ 1 GB raw data in one second with 120 Hz data acquisition rate.

detectors are deployed [4]. As a consequence, this causes problems in data storage and transfer, both requiring serious investigation to find the best approach to archive management. We do not intent to discuss the data management aspect in this review. In the following sections, the challenges to computational sciences in data analysis and model reconstruction are discussed in depth. In the section of *Applications and current status*, three major applications of XFELs are summarized, and the current status is described. In the section of *Computational problems*, the challenges to the computational sciences are discussed, along with some possible solutions. Lastly, future perspectives are elaborated, with the aim of inspiring multidisciplinary research activities in the advances of XFEL applications in structure biology.

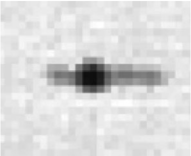
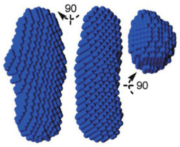
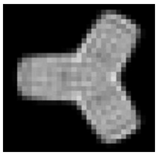
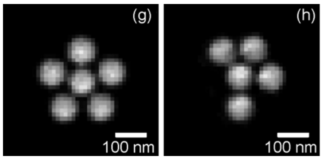
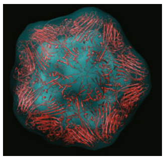
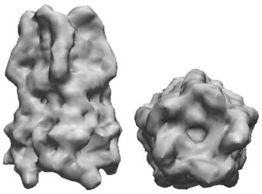
APPLICATIONS AND CURRENT STATUS

There have been three main approaches in structural analysis using an XFEL, depending on the sample type and experimental setup. The most successful method has been serial femtosecond crystallography (SFX), which records diffraction patterns from crystals of sub-micron to few-micron sizes [5–7]. Here the coherent amplification of scattering by the Bragg mechanism can provide atomic resolution data (Since the Bragg peak intensity is proportional to the square of the number of molecules, a $10\times10\times10$ unit cell nanocrystal produces a million times more scattering than one molecule at this peak. Total scattering across the Bragg rocking curve is, however, proportional to the number of molecules). Because of handling difficulties for goniometer mounting and weak diffraction, these invisible crystals could not be practically used without XFELs, whose ultrabright X-ray pulses in a beam of sub-micron focus provide diffraction intensities from these tiny crystals before the onset of radiation damage. For a sufficiently brief X-ray pulse, the flux of the pulse which produces a damage-free diffraction pattern is in principle unlimited. Samples are delivered in a continuous liquid stream, first proposed as a method for nanocrystallography by Spence and Doak, to avoid the need for goniometer mounting [8]. SFX has gone through rapid development since the commissioning of LCLS in 2009 as an emerging technology for structural studies. There is now a large literature on SFX methods and applications, covering sample preparation, crystal delivery, data collection, pre-analysis and post-refinement, all the way to final structure determination (for reviews, see [9,10]). The Philosophical Transactions of the Royal Society B and "Structural Dynamics" have each published a special issue on XFEL applications in structural biology recently. Here, we focus on data analysis and existing problems with possible solutions using advanced computational theory and algorithms.

A second approach that has aroused considerable interest in the scientific community is the single particle imaging (SPI) mode with XFELs. Here the beam diameter can be focused to generate very high photon flux that can interact with one bioparticle, such as a virus, so that the scattering from individual particle can reach higher resolutions. The X-ray SPI method could have a dramatic impact on structural biology: not only by bypassing the requirement for crystallization, but also by enabling structural determination in native molecular environments at physiological temperatures. Furthermore, because the XFEL pulse duration is on the femtosecond time scale, it is possible to probe ultrafast chemical or biological reactions. Coupled with light pumping methods, the detailed dynamics of light-sensitive biomolecules may be revealed using XFELs. It could also be used to study enzymatic reactions with advances of mixing techniques, such as fast mixing devices using microfluid instruments [11] or using photocaged ligands [12]. The computationally challenging problems in SPI are directly related to the sample heterogeneity and unknown orientations, which are entangled. We focus on the discussion of these two problems in the following section for SPI data analysis.

The third application of XFEL for non-crystallized samples is known as Fast Solution Scattering (FSS), or fluctuation X-ray scattering, or correlated fluctuation scattering (denoted as FSS in the following text). This method, in which many molecules in solution are exposed to X-rays in each shot, is similar to conventional small-angle (or wide-angle) X-ray scattering (SAXS or WAXS). However, particle rotations do not take place during each femtosecond exposure. It was shown by Kam [13] that the resulting patterns are in principle two-dimensional (even when coherent interparticle scattering is ignored), unlike the one-dimensional data produced by SAXS. They therefore contain more information than angularly isotropic SAXS patterns, facilitating the 3D reconstruction of electron density map. Structural information can be extracted by converting the raw intensity on the detector to an angular correlation function between intensities at different pixels. This method was first described by Kam and others [13,14], and has started to become feasible with the development of XFEL. Kam's method performs the remarkable feat of extracting the scattering pattern of a single oriented molecule from the scattering generated by many randomly oriented molecules in solution. This can be understood by noting that the angular correlation function obtained from a diffraction pattern from one particle is independent of the particle orientation, so that successive functions may be summed. The result can be inverted to a density map by phasing the data twice, using iterative methods, as further discussed below. In Table 1, several experimental research and simulation investigations are summarized. Model reconstruction from experimental data is feasible albeit low resolutions. On the other hand, simulation research suggested that high resolution data and structures can be achieved. The FSS method has several advantages: (i) Since the X-ray beam spans many particles, the hit rate is 100%, unlike the hit rate of the single particle imaging mode, which may be just a few percent; (ii) The structural information is represented using relative coordinates in the form of correlations, which integrates over all orientations; (iii) The method is more tolerant to a non-flat wave front, placing less stringent requirement on X-ray beam focal size and beam profile. A significant theoretical advance was published recently, in which a new iterative phasing method to invert the accumulated angular correlation function to the real-space density map is derived and implemented [20].

Table 1. Representative studies in structure determination using FSS method.

Sample	Model	Resolution	#particles/shot	Reference
	Experimental data			
Gold rod		~ 13 nm	~10	[15]
Nanorice		~ 40 nm	1	[16]
Gold particle		~13.5 nm	~20	[17]
	Simulation data			
Sphere cluster		~25 nm	10	[18]
Icosahedral virus		1.3 nm	n/a	[19]
pLGIC protein		0.48 nm	n/a	[20]

Typical experiment setups for these three types of experiments are depicted in Figure 1, composed of sample delivery equipment, XFELs, fast readout detectors, and data storage and computational analysis clusters (not shown).

COMPUTATIONAL PROBLEMS

The XFEL generates large amount of experimental data, and the analysis and model reconstruction can only be done with modern computers equipped with tailored software. For the three aforementioned types of experiments, the data analysis procedure is very similar: (i) hit-finding, picking the sub-dataset with scattering or diffraction intensity from samples; (ii) sorting data based on quality; (iii) recovering orientations and merging intensity to the volume in 3D reciprocal space; (iv) phase retrieval to get 3D electron density map for real space model; (v) model quality assessment and refinement. This is summarized in the flow chart shown in Figure 2. In the following subsections, several major challenging problems are elaborated for each of the three types of applications, namely: serial femtosecond crystallography (SFX), the single particle imaging (SPI), and fast solution scattering (FSS). The problems in conformational changes are common for these experiments, so they are discussed in the fourth subsection, the dynamics.

Serial femtosecond crystallography

Diffraction pattern identification: hit finding. The hit finding is straightforward for SFX data, because the bright Bragg spots provide clear signatures for the actual "hits" to be distinguished from the patterns resulting from X-rays scattering without hitting any crystal. Frequently used hit-finding software includes Cheetah, CASS, and cctbx.xfel [21–23]. The "hits" can result from multiple crystals that are in different orientations, or from poorly ordered crystals, or from crystals that are too small to yield strong signals. Therefore, before 3D merging, the data needs to be classified to sort out the indexable dataset that gives consistent unit cell parameters.

Intensity integration. In current XFELs, the X-ray laser pulses are generated by self amplified spontaneous emission (SASE) process. The results are limited by appreciable bandwidth (for example, the bandwidth is within 0.2% (ΔE/E) at the LCLS). Due to this reason, the Ewald sphere is a thin shell that only partially intercepts Bragg spots, whose angular profile is broadened by crystal mosaicity and the crystal shape transform (see Figure 3). Unlike conventional crystallography, in which crystals are oscillated or rotated to scan the volume of the Bragg spots for full intensity measurement, the narrow bandwidth of X-ray lasers and ultrashort exposure time limit the number of Bragg spots and measured Bragg

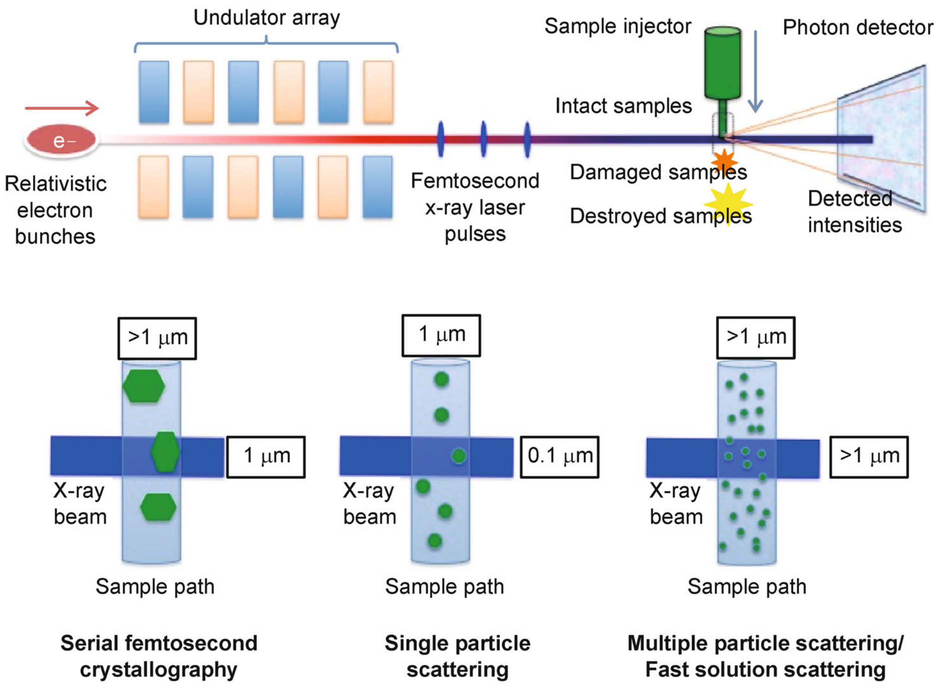


Figure 1. Diffraction-before-Destruction experiment setup at XFELs. Three types of experiments are distinguished from the nature of collected data, which are due to the sample type and the way the samples interacting with XFEL pulses.

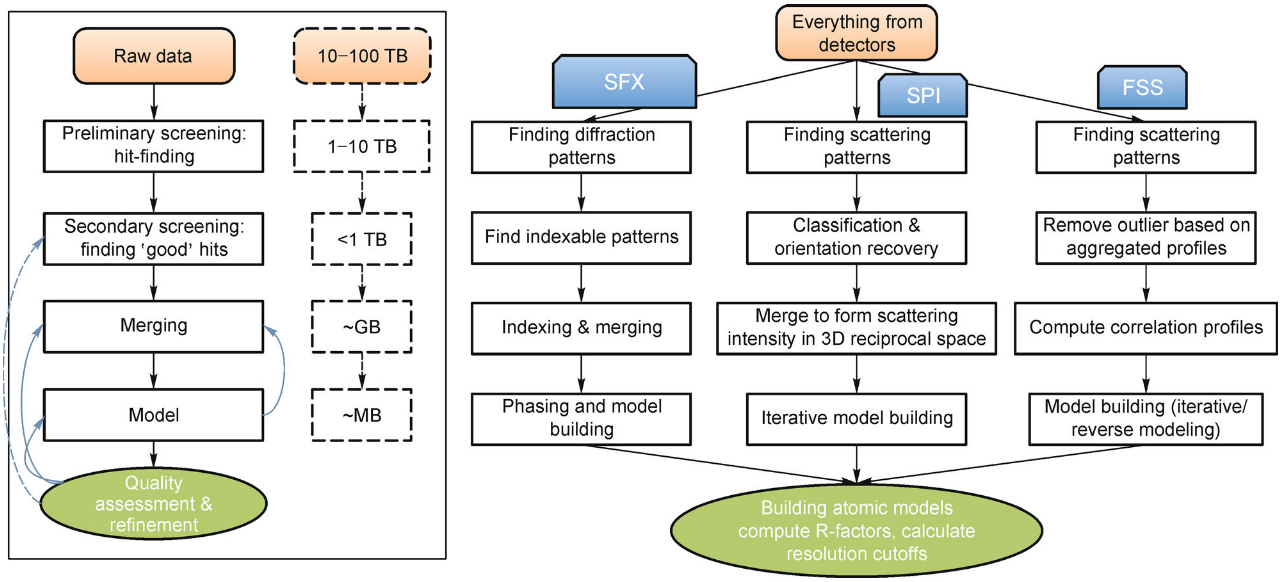

Raw data
10–100 TB
Preliminary screening: hit-finding
1–10 TB
Secondary screening: finding 'good' hits
<1 TB
Merging
~GB
Model
~MB
Quality assessment & refinement
Everything from detectors
SFX
SPI
FSS
Finding diffraction patterns
Find indexable patterns
Indexing & merging
Phasing and model building
Finding scattering patterns
Classification & orientation recovery
Merge to form scattering intensity in 3D reciprocal space
Iterative model building
Finding scattering patterns
Remove outlier based on aggregated profiles
Compute correlation profiles
Model building (iterative/ reverse modeling)
Building atomic models compute R-factors, calculate resolution cutoffs

Figure 2. Workflow for data analysis. The arrows pointing downwards are the main flow direction. The backward direction can be helpful for data selection and calibration. The backward direction analysis should be carried out with caution to avoid model bias. The pipelines corresponding to the three types of experimental data are summarized on the right panel: **SFX**, serial femtosecond crystallography; **SPI**, single particle imaging; **FSS**, Fast Solution Scattering.

volume in each individual diffraction pattern, resulting partial intensity measurements. The term partiality denotes the ratio between the partial intensity and the full intensity. Since the beginning of SFX method development, the full intensity dataset was obtained by Monte Carlo angular integration across the crystal rocking curve, which assumes that a sufficiently large number of random angular samples around each Bragg spot will eventually provide a good estimate of full scattered intensity [24,25]. Each Bragg spot has a set of indices (*h, k, l*) called miller indices. The indexing procedure is to map each Bragg spots on diffraction pattern to 3D reciprocal space. The intensity at miller index (*h, k, l*) then is obtained by summing *n* measurements:

$$I(h,k,l) = \Sigma_n I_n(h,k,l) \qquad (1)$$

The Monte Carlo method averages over stochastic fluctuations, such as variations in crystal size and quality and the shot-to-shot variation in incident X-ray flux. The resulting error in a measurement of structure-factor then falls off inversely as the square root of the number of diffraction patterns, as shown in Figure 4. As more measurements are accumulated, the mean value of all measurements on intensity at each Bragg spot converges towards the full intensity. This is true in general, but requires a large amount of data, i.e., tens of thousand indexed diffraction patterns.

Recently, algorithms have been developed to refine experimental parameters, such as beam profile, crystal properties, and orientations, which can be used to estimate the partiality for measurements at each Bragg spot, and subsequently corrections can be made to the partial intensities. It has been found that such post-refinement significantly reduced the required number of diffraction patterns. For example, efforts for partiality correction based on the estimated fraction of Bragg volume were first discussed by White [26] and by Sauter [27], following the approach proposed by Rossmann *et al.* for conventional crystallography data analysis [28]. Since crystal orientation influences the partiality evaluation, the orientation matrix obtained from indexing is subjected to post-refinement, a process that involves the optimization of the incident X-ray beam profile, in order to find the parameters which best predicts the observed number of spots in each diffraction pattern [23,29,30]. Auto-indexing method provides a rough estimate of the crystal orientation relative to the beam direction, the positions of Bragg spots on the detector change a little with small changes in crystal orientation, but intensities (which cannot be predicted without a knowledge of the crystal structure) can change greatly. The refinement of crystal orientation, detector pixel location and beam direction will improve the estimation of intensity partiality. Most algorithms follow auto-indexing with Monte-Carlo summation to give approximate angle-integrated Bragg intensities. This procedure is repeated until the results converge to optimal values. The recent method of Ginn *et al.* (2015), which minimizes the distance between the sum of partial reflection maxima and ideal (average) reciprocal lattice points for a given X-ray bandwidth profile, to

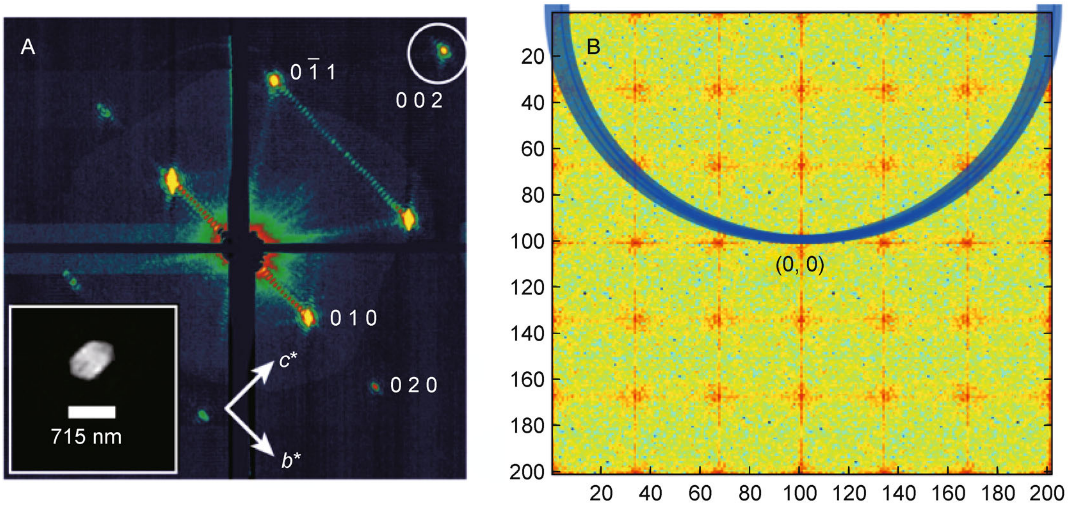


Figure 3. Diffraction pattern from nanocrystals. (A) Photosystem I crystal and its low angle diffraction pattern recorded using pnCCD detector at a far distance (564 mm), showing broad intensity distribution around Bragg spots. The shape transform is inverted to reconstruct the shape of the nanocrystal (inset). It also indicates that the intensity recorded is not full reflection because in the third dimension the intensity is also broadly distributed, and Ewald sphere only cut through part of the distribution function. (B) Schematic drawing of shape transform of crystal overlaid on lattice point. This demonstrates the cause of partial intensity due to the shape transform, only the overlapped regions of the Ewald shell (blue color) and Bragg spots are excited and produce diffracted intensity (Figure 3A is reproduced from Chapman *et al*., 2011, Ref. [5] with courtesy from Nature Publishing Group).

optimize the crystal orientation for each shot, appears to be a promising approach. Post-refinement not only reduces the number of required diffraction patterns to a few thousand, and but also improves the merged data quality. In an alternative approach, Zhang *et al.* used a profile fitting approach to estimate the full intensity by assuming that the intensity distribution around each Bragg spot follows Gaussian distributions [31]. The partiality correction can also be helpful to resolve the indexing ambiguity, to be discussed next. Besides these improvements in data analysis stage, a two-color "split and delay" experimental approach has also been suggested for both SFX and pump-probe experiments, which effectively increases the energy bandwidth of X-ray pulses to give a thicker Ewald sphere (a two-color laue diffraction) [32].

The experimental geometry and detector metrology (e.g. the tiling of detector panels in a 2D CSPAD detector) are closely related to partiality correction and intensity integration. D. Loh conducted a systematic survey on the stability of the X-ray incidence direction, and found that pointing directions fluctuated over an angular range from −0.5 mrad to 0.5 mrad for the case of AMO station at LCLS [33]. These fluctuations will affect the indexing and integration analysis, so that re-centering for each diffraction pattern might be necessary. The metrology calibration must be carried out when the panel arrangement is modified. Hattne *et al.* reported that metrology corrections needed to be achieved to sub-pixel level precision to obtain optimal indexing and integration results [23].

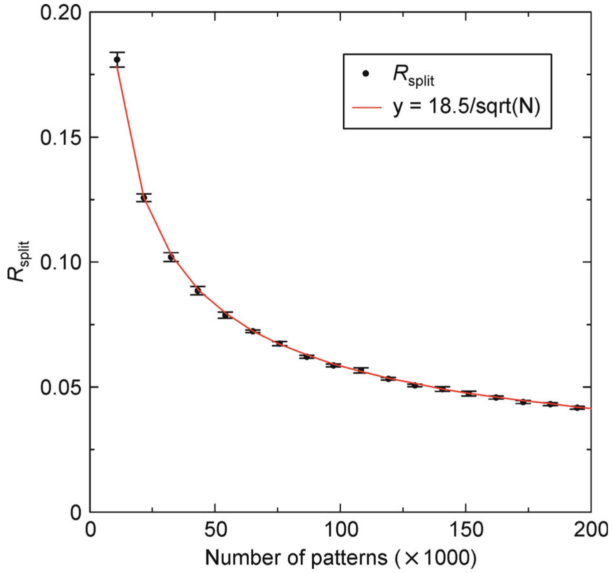


Figure 4. The accuracy of merged intensity depends on the number of diffraction patterns in SFX. Photosystem II protein complex (PDB 3WU2) is used in this simulation and the partial_sim program in CrystFEL is used to generate diffraction data. R_{split} shows the consistency between merged intensity distributions from two half-datasets that are randomly splitted from the full dataset.

The most common integration methods consist of two steps: (i) integrate the intensity around each Bragg spot on the 2D detector (in a confined area surrounding the pixel with peak intensity) and subtract background estimated from a larger annular region around the spot, yielding partial measurement of the reflection (the intensity at Bragg spot); (ii) index the dataset to determine the crystal orientation, and merge the partial intensities (summing Bragg intensities corresponding to the same Miller indices from all diffraction patterns) to obtain full angle-integrated intensities. The post-refinement discussed in the previous paragraphs mainly targets the second step, which requires iterations to optimized orientation, partiality corrections, and so on. Yefanov *et al.* proposed the idea of merging intensity from all pixels in each diffraction pattern to 3D reciprocal space volume, instead of mapping the Bragg intensities [34]. This approach is the same as the analysis for single particle scattering that is discussed in the next section. It has the following advantages: (i) the intensity information is utilized besides the geometry relations of Bragg spots during the indexing procedure, so that the indexing ambiguity is automatically resolved; (ii) the intensity between Bragg spots can be retained for phase retrieval using the idea of shape transforms (to be discussed in the phase retrieval section); (iii) the method works well with ultra-thin Ewald spheres, when the beam bandwidth is very narrow. As reported by Yefanov *et al.*, the 3D merging approach gave higher resolution than the Monte Carlo integration method for the published data of cathepsin B (see [7]).

Resolving the indexing ambiguity. Indexing ambiguity is a problem associated with serial crystallography, which occurs when the Bravais lattice symmetry is higher than the space group symmetry. In such cases, there are multiple equivalent assignments of miller indices to the reciprocal lattice, however the Bragg intensities are not equivalent. For the simple case of merohedral twinning, where two indexing schemes exist, there is a probability that a second crystal will be indexed as a "virtual twin" of the first if the lattice information alone is considered rather than together with associated Bragg intensities. In order to ensure that the indexing is carried out in a consistent way that maps the correct intensities to the corresponding reciprocal point, algorithms that use the intensity information at each point must be developed in addition to the programs that exploit geometry information of Bragg spots. It is worthwhile noting that each diffraction pattern records (different) partial intensities for each Bragg spot, therefore the datasets extracted from individual patterns do not provide full intensity as references, complicating the assignment of indices. For example, the effects of partiality may be greater than those of a twin assignment. Several methods have been proposed and found to be useful to resolve this ambiguity: the BD algorithm developed by Brehm & Diederichs used the correlations between datasets and a distance metric between diffraction patterns to separate the indexing groups [35]; following this idea, the *ambigator* program has been implemented to utilize the relation between each pattern and a reference obtained from an ensemble of patterns to iteratively resolve the ambiguity [25]. A third algorithm uses the idea of expectation maximization that has been used in orientation recovery for single particle imaging, allowing a merged intensity model to be refined, while orientation assignments are improved with the model obtained from the previous step, and eventually the model and orientations all converge to the correct solution [36].

Phase retrieval. Obtaining accurate diffraction intensity paves the way for subsequent model reconstruction and refinement. It is known, however, that phase information dominates the 3D model reconstruction process, so that direct measurement of phase is highly desirable to avoid model bias, and to determine new structures entirely on experimental measurements. The molecular replacement (MR) method relies on phase information from a molecule which is similar to the target molecule in structure and sequence. Other phasing methods, such as single or multiple wavelength anomalous dispersion (SAD or MAD), have not been widely applied in SFX yet. The anomalous signal from SAD, for instance, is very small, thus requiring high accuracy in the measurement of the full angle-integrated intensity at Bragg spots. Berends *et al.* carried out the first SAD phasing experiment with XFEL on lysozyme crystals soaked with gadolinium, which has a strong anomalous signal (contributing about 10% signals to the total amplitude at about 2Å resolution, whereas the anomalous signal is only about 1.5% if sulfur is used) [37]. In that study, about 60,000 indexed patterns were integrated using the Monte Carlo approach implemented in CrystFEL. The development in post-refinement for intensity integration can be expected to reduce the requirements on data quantity and improve the accuracy of the diffraction intensity measurement. It is therefore worthwhile re-analyzing all the data that has been published, including that from the SAD experiments, using various post-refinement approaches. The compiled dataset will form a benchmark and basis for future algorithm developments.

An *ab initio* method of phasing for nanocrystals has been proposed by Spence *et al.*, based on the Fourier Transform of nanocrystals, which can be shown to be laid down around every reciprocal lattice point, i.e., the Bragg spots [38]. This "shape transform" produces fine fringes running between the Bragg reflections, which sample the transform of the contents of the primitive unit cell, often referred to as the "molecular transform". The molecular transform envelope can therefore be evaluated from the intensities between Bragg spots, and this can be phased using the iterative methods of Coherent Diffractive Imaging (CDI).

The scattered intensity for n-th pattern at scattering vector $\Delta\boldsymbol{k} = \boldsymbol{k}_{\mathrm{i}} - \boldsymbol{k}_{\mathrm{o}}$, where $\boldsymbol{k}_{\mathrm{i}}$ and $\boldsymbol{k}_{\mathrm{o}}$ are incident and scattered wave vectors, is

$$I_n(\Delta\boldsymbol{k}) = J_0 r_e^2 P |F(\Delta\boldsymbol{k})|^2 |S_n(\Delta\boldsymbol{k})|^2 \Delta\Omega \qquad (2)$$

Here, J_0 is the incident photon flux density (photons/pulse/area), $F(\Delta\boldsymbol{k})$ is the structure factor depending on the molecules, r_e is the electron scattering radius, P is a polarization factor depending on the X-ray beam, and $\Delta\Omega$ is the solid angle spanned by the detector pixel measuring the $\Delta\boldsymbol{k}$ diffraction beam. It is obvious that $|F(\Delta\boldsymbol{k})|^2$ can be separated from the shape transform $|S_n(\Delta\boldsymbol{k})|^2$ by "dividing out" the shape-transform factor. However this is best done after merging into a 3D data set. The structure factors provide molecule structural information. If the shape transform can be determined, and because all other terms are constants or known parameters, the structure factors can be evaluated as

$$|F(\Delta\boldsymbol{k})|^2 \propto \frac{I_n(\Delta\boldsymbol{k})}{|S_n(\Delta\boldsymbol{k})|^2} \qquad (3)$$

Spence *et al.* have derived a way to estimate the shape transform function $|S_n(\Delta\boldsymbol{k})|^2$, which can be calculated from diffraction volume obtained using the procedure described by Yefanov *et al.* [34], based on the fact that the modulation function is the same around each Bragg spot (see Figure 5). If we partition reciprocal space into Wigner-Seitz cells around Bragg spots, then shift all Wigner-Seitz cells to the origin, and sum up the intensity in the same location, the resulting distribution well approximates the shape transform. Subsequently the shape transform modulation can be removed from diffraction data, so that an oversampled intensity distribution map can be obtained for iterative phasing, using the algorithms that utilize information from both real space constraints (positivity/isolated/continuity) and reciprocal space amplitude constraints, as pointed out by Sayre [39]. These algorithms have already been developed for coherent X-ray diffraction imaging [40]. The foundation behind these algorithms lies in information

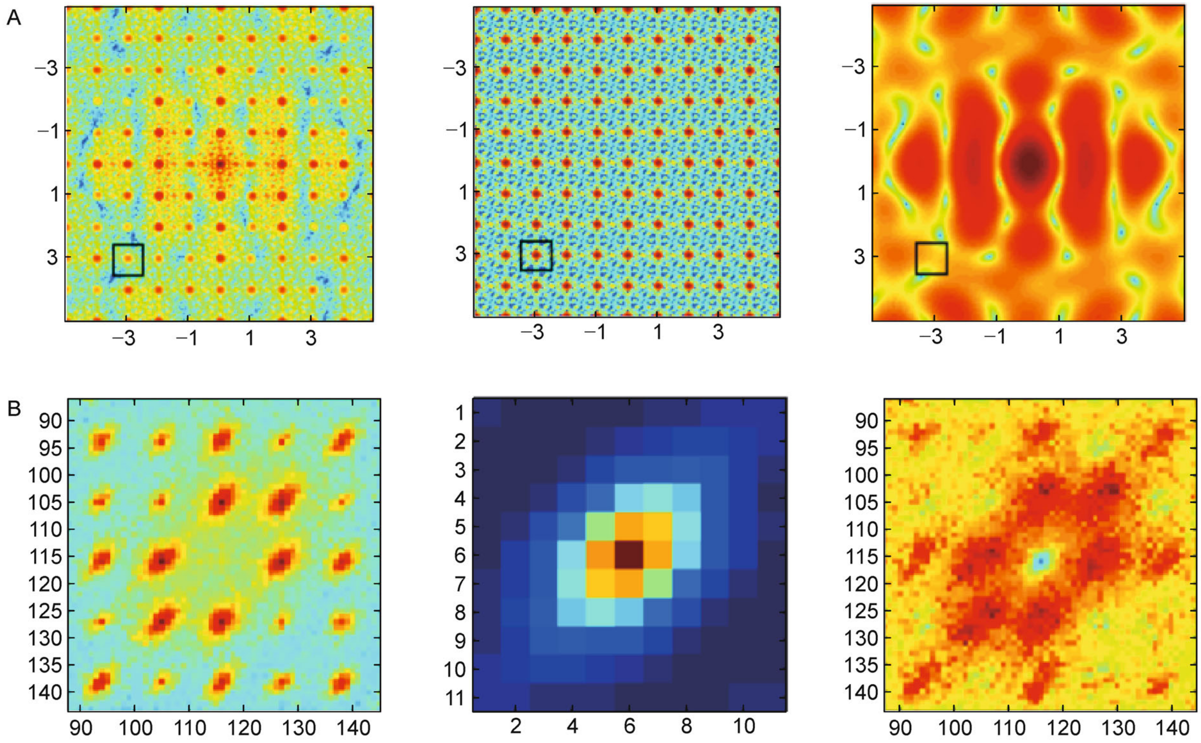


Figure 5. Phase retrieval using shape transform method. (A) Simulated data, and from left to right: diffraction pattern, shape transform (modulation function) arranged on lattice points covering the same reciprocal space, and the de-modulated structure factors that can be used for iterative phasing. (B) SFX experimental data for photosystem-I nanocrystals. The arrangements are the same as in (A), except that the middle panel shows one copy of the shape transform.

theory: the oversampled intensities together with knowledge about the model in real space compensate for the unknown information on phases. More details will be elaborated below in the single particle scattering section. The shape transforms of nanocrystals was observed during the first nanocrystallography beamtime at LCLS, using photosystem I crystals as samples (Figure 5) [5]. The feasibility of this shape transform approach was initially investigated using simulated data [38], and then attempted at LCLS using nanocrystals of polyhedrin proteins from granulovirus with little success initially. However recently successful phasing by this method has been demonstrated for artificial 2D crystals at the FLASH facility [41]. The main challenges with the method are: (i) the limited dynamic range of detectors; (ii) the X-ray beam brilliance; (iii) crystal quality. The second issue is easy to understand: higher brilliance allows signal collections to higher resolutions, especially for very small crystals whose shape transforms are more pronounced. We now elaborate the other two issues.

The CSPad detector has a dynamic range of 10^3, which is not enough to simultaneously measure both the strong scattering intensities at Bragg peak and the weak intensities in between Bragg spots. To tackle this problem, Millane and coworkers have demonstrated that the requirement on intensity sampling can be relaxed to the points where signal-noise-ratio are the highest within the space between Bragg spots, and their simulation results suggest that oversampling ratio of three should be sufficient for iterative phase retrieval [42,43]. In another study, Elser proposed a different approach to utilize the in-between Bragg intensities: by combining intensity and gradient at each Bragg spot, the phase information can be recovered using his algorithm [44]. Both algorithms are waiting for validations with experimental data.

Crystal quality here means that the crystals should have similar sizes and shapes in order to optimize the convergence of the shape transform calculations. Another issue concerning crystal quality results from edge or termination effects. It has been found that the crystals may terminate at their surface with an incomplete unit cell (for example in the case that the primitive cell contains more than one molecule). A crystal with two molecules per primitive cell may contain an odd number of molecules across its width. The additional molecule will have a large effect on the inter-Bragg scattering. This edge effect was first studied by Liu *et al.,* who concluded that the partial unit cells should not affect the reconstruction of the oversampled intensity from single unit cell if the incomplete unit cells are randomly distributed over the edge or surface of the crystals [45]. Kirian *et al.* have demonstrated successful experimental phasing by this method for the first time, using a modified shape transform algorithm that allows for the existence of incomplete unit cells at the crystal boundary. They take account of this effect by allowing several choices of unit cell for the nanocrystal, the unit cell type that correctly generates the finite crystal can emerge from phase retrieval procedure [41].

Very recently, a new method for phasing and resolution enhancement has been described for SFX data [46]. For protein crystals whose disorder consists solely of translational displacements from an ideal lattice, the Debye-Einstein theory of diffuse scattering can be used to show that diffraction patterns will show, in addition to sharp Bragg spots, diffuse scattering between the Bragg spots whose envelope is the modulus of the molecular transform (the fourier transform of the unit cell). This resulting continuous diffraction pattern is the incoherent sum of diffraction signals from rigid molecules enclosed by individual unit cells, while the sharp bragg intensities are the results of coherent sum of diffraction signals. More importantly, this diffuse scattering extends beyond the diffraction signals, and so may be used to improve resolution. The continuous scattering also provides the “oversampling” needed to solve the phase problem by iterative methods.

Single particle scattering

The determination of high resolution structures using X-ray scattering from individual particles (for example with one virus per shot) not only avoids the crystal growth process, but also allows the study of protein structures under physiological conditions. Single particle experimental data is not expected to be homogeneous, and the dataset should provide information about an ensemble of conformations. One of the earliest experiments conducted at the LCLS was the detection of the X-ray scattering from a single virus particle (the Mimivirus, with a diameter of about 0.45 microns) by Seibert *et al.*, using a 3 micron diameter beam containing about 10^{12} photons per pulse (about 10^{10} times brighter in terms of peak brilliance, compared to advanced synchrotron radiations). This demonstrated the feasibility of collecting scattering from single particles, and further shown that the phase retrieval can be achieved from the oversampled scattering patterns [3]. The 2D projections were constructed at about 32 nm resolution and showed the dense genetic materials enclosed in the virus. The first 3D reconstruction from single particle scattering data using XFEL data was only completed after five years of continuous development. Ekeberg *et al.* reported their 3D model of mimivirus built from scattering data merged from 198 scattering patterns [47]. However, the resolution of the reconstructed model is unexpectedly low (about 125 nm for full-period resolution) due to difficulties in merging 2D diffraction patterns in correct orientations into 3D diffraction volume. Although the determination of molecular structures to high resolutions using the diffraction-before-destruction approach with XFEL pulses is feasible in theory, there still remain many challenges for sample delivery, detector calibration, algorithm development for sorting samples and so on. To solve these technical challenges, researchers have formed an international collaboration under the leadership of the LCLS — the "single particle imaging initiative" (SPI-i) [48]. It is hoped that this collective effort can speed up the development of single particle scattering technology using XFELs, and progress has been made in many areas, including reduction of stray X-ray background in the beamlines and detector metrology. In the following sections, we focus on two major challenges to computational science: the sample heterogeneity problem, and unknown orientation problem. These two problems are entangled and difficult to be solved separately without *a priori* information about the sample structures.

Scattering pattern picking and screening. The first step in single-particle data analysis is to identify and extract scattering signals. This is not as straightforward as in the case of Bragg diffraction pattern recognition. The scattered intensity distribution from non-crystalline samples produces a continuous distribution, so the scattering signals spreading out in reciprocal space is thus much weaker than Bragg scattering intensities. For this reason, the actual scattering patterns from single particles are harder to be distinguished from background, especially when the sample is delivered using liquid solvent that also scatters X-rays. Since actual scattering pattern may be identified clearly from the low-angle scattering (which, from an icosahedral virus, shows a series of rings, as shown in Figure 6), a comparison of hits and misses (from jet solvent alone) can be used to identify high-angle scattering from bioparticles. An analysis of this difference signal has recently indicated the presence of scattering out to sub-nanometer resolution in scattering patterns from individual viruses, however this provides insufficient signal to reconstruct an image. To reduce the solvent contribution, a smaller liquid jet has to be used whenever possible — the smallest droplet beam so far generated produces droplets of about 0.3 micron in diameter [49]. Improvements in aerosol injecting methods can potentially help reduce this solvent scattering to a level that model reconstruction algorithms can tolerate [48]. Our current understanding of single particle scattering is that the signal comes from the sample particle with a very thin solvent layer in vacuum, which preserves the molecules in their native hydrated states. Computational software is under development to analyze such datasets. If the samples are delivered using a continuous liquid injection system, the solvent background needs to be treated carefully. The data processing methods for SAXS/WAXS can be borrowed for background subtraction. In order to enhance the signal, a multiple particle scattering approach might be adapted (with many particles per shot), and the signal extraction can be accomplished with the angular correlation function approach proposed by Kam [13].

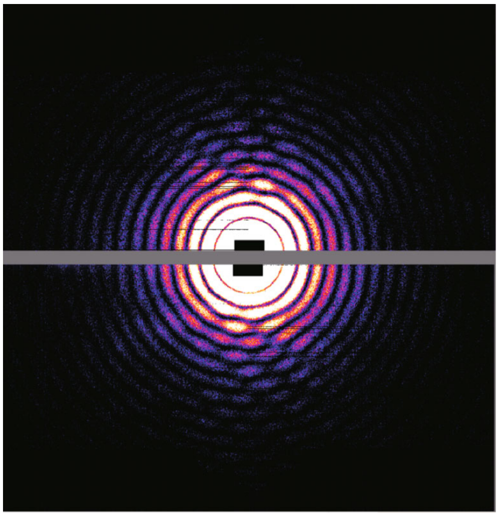


Figure 6. Scattering pattern from a single paramecium bursaria chlorella virus particle, collected at AMO beam station with LAMP chamber. (The data is kindly provided by Ilme Schlichting from Max Planck Institute for Medical Research, Heidelberg).

Orientation information. First, let's simplify the problem by assuming that particles are identical within the resolution limit of interest, several algorithms can be directly applied to recover the relative orientation between scattering patterns. The "common arc" method relies on the analysis of the intersecting lines of any two scattering patterns [50]. It is known that any pair of scattering patterns must intersect along a curve running through the origin of reciprocal space. A third pattern then fixes, by its intersections with the other two patterns, the orientation relationship among these three scattering patterns. Kassemeyer *et al.* developed a method (Geodesic and In-Plane Rotations ALgorithm, GIPRAL) to map similarity between patterns to geodesic distance on a sphere and applied this method in the reconstruction of an ellipsoidal nanoparticle scattering data [51]. The GIPRAL can throw out outliers if the pattern is from an object that is sufficiently different from the object of the aligned patterns. Expectation-maximization algorithms find orientation information and merge intensities using an iterative procedure, during which the expectation of fitness to a model is maximized gradually until the merged data converges to a model that optimally arranges all patterns consistently [52,53]. Manifold embedding method maps each pattern to higher dimension space and exploits the relations to find orientations. Here each diffraction pattern containing N pixels is represented by a vector in N-dimensional Hilbert space. The high dimensional space is then reduced to SO3 rotation space, where similar patterns will be found close to each other, since their difference vector forms the Euclidean metrics used in the least-squared optimization method [54]. It is worth to note that these algorithms were used to solve orientation problems in cryo-electron microscopy (cryoEM) single particle imaging experiments, in which projection images need to be arranged in SO3 rotational space.

The common arc method is fast but is not robust when the signals are weak, because the intersecting lines utilize only a fraction of the information contained in each pattern. The other approaches use all information contained in 2D patterns and the relation with all other patterns in the whole dataset. The expectation maximization method implemented by Loh and Elser in their expansion-maximization-compression (EMC) algorithm has demonstrated its applicability in recovering orientations of scattering patterns from simulated and experimental data [47,53]. The reconstructed 3D model of a mimivirus was determined by utilizing a modified version of the EMC program [47]. This approach can also be used to classify the scattering patterns into several discrete conformations, similar to the approach cryoEM data analysis [52]. The manifold embedding method has recently been applied to cryoEM images, and so used to map out the continuous conformational changes of a large molecular complex, the ribosome [55]. The frequency of occurrence of each conformation in image dataset can be converted to free energy via Boltzmann relations. Therefore, it is possible also to map out energy landscape for protein dynamics.

Sample heterogeneity. Conformational changes are fundamental to molecular function. Therefore, without the conformational filtering (such as in crystallization), particles (especially proteins) are usually not identical in SPI, so it is necessary to sort scattering patterns into distinct conformations. For example, in ligand binding studies, the molecule could be in the closed state when the ligand molecule bounded, but in an open state without ligand. Synthetic nanoparticles cannot avoid the heterogeneity introduced by inadequate control of synthesis conditions and growth process. Although the orientation recovery algorithms can be applied in such cases, the result and performance remain to be subjected to extensive testing with experimental data. Synthetic nanoparticles often grow into different sized particles with similar shapes as designed if the growth method is based on a layer addition mechanism. In their study on the nanorice particle scattering, Kassemeyer *et al.* found that scattering from a different sized nanoparticle need to be put aside during orientation recovery and merging. They reported that outliers formed an un-merged dataset, which could be subsequently merged to another model with larger size [51]. Some nanoparticles have nice features that can facilitate particle size analysis. For instance, facetted nanocube particles generate scattering patterns with streaks that shown clear fringes in orientations that X-ray incidence direction is normal to a cube face. The spacing between the fringes can be used to calculate the size of the particles [56]. For other orientations that have a tilting angle relative to the normal incidence case, a model with simulated patterns can help to find particle size and orientation. This approach utilizes priori information on the particles and has been used to study a set of core-shell nanoparticle scattering experimental data. The size distribution function obtained using this approach is consistent with real-space images from electron microscopy (Li *et al.*, unpublished data). This work shows the value of priori knowledge for single particle scattering data analysis.

Multiple particle scattering

It is known that SAXS/WAXS experiments can provide limited but valuable structural information. Because many molecules in random orientations are exposed to X-rays for a time duration that is much longer than the rotational diffusion time constant (second vs microsecond), the resulting SAXS scattering patterns are angularly isotropic, yielding only one-dimensional curves. If the exposure time is much shorter than rotational diffusion time constant (or molecules were held stationary in ice or on a solid support during illumination), the scattering patterns become anisotropic, and intensity fluctuations can be observed around each q-ring. This was firstly pointed out by Kam in 1977, and ultrashort XFEL pulses now make it technically feasible to collect snapshot scattering patterns from stationary particles. A review of the method can be found in the article by Kirian [57]. The term “fluctuation scattering” (or fast solution scattering, FSS) reflects these intensity variations in scattering pattern recorded from such experiments. These two-dimensional patterns contain more structural information than 1D SAXS or WAXS patterns, however it has been shown [58] that inversion from 2D to 3D in this case (although greatly preferable to inversion to 3D from 1D) is incompletely specified, unless additional information such as symmetry or a good initial model is provided. In order to extract structure information from FSS patterns, Kam developed an angular correlation analysis, using relative coordinates to encode the structural information. For example, the auto-correlation between intensities on the same q-ring ($q = 4\pi\sin\theta/\lambda$ and 2θ is the scattering angle) can be defined as following:

$$C2(q,\Delta\phi) = \int_{0}^{2\pi} I(q,\phi)I(q,\phi + \Delta\phi)d\phi \qquad (4)$$

This is the autocorrelation for a single scattering pattern, and Kam showed that if summing over many autocorrelations each calculated from molecules in random orientations, the resulted profile converges to the autocorrelation for a single molecule.

A simple way to understand the Kam method (outlined in [57]) is to note that, in two-dimensions, for particles which differ only by rotation about the beam direction, the angular correlation function is independent of the particle's orientation (analogous to that spatial correlation function, or Patterson function for a crystal, is independent of origin or translational shift). A sum of angular correlation functions from particles (with one particle per shot) can therefore be formed, equal to that of one particle, but with steadily improving single-to-noise ratio. If more than one particle per shot is considered, the same result applies, however an isotropic background is added, which can be shown to be equal to the SAXS pattern. Since this background is isotropic, it can be separated from the single-particle angular correlation function. The remaining problem is to invert this function to the electron density map.

Looking at Equation (4), we can see that structural information is encoded in the internal coordinates (q,$\Delta\Phi$). There have been several different approaches for 3D model reconstructions from correlation profiles. Saldin *et al.* demonstrated a 2D case using scattering data from many identical golden rod particles lying sideways on a membrane, using the two stage phasing approach: first to find the scattering intensity by combining the circular expansion components that are obtained from the correlation profiles, and then by reconstructing the model using iterative phasing methods [15]. A similar method was used by Starodub *et al.* for the reconstruction of a 3D dumbbell model from the scattering data collected at LCLS [59]. It is noteworthy to point out that the orientation of dumbbell is defined by two Euler angles (unlike one rotation angle for the gold rods). Saldin and coworkers have also exploited the symmetry properties of virus particles to facilitate the reconstruction of icosahedral virus capsids and helical viruses [19,60]. In a different approach, Liu *et al.* demonstrated the feasibility of reverse modeling method for model reconstruction using a simulated annealing algorithm [16]. From LCLS experimental data, the angular correlation profiles were extracted and then used as the blue-print for model building (a process similar to sculpturing). The model was adjusted until its correlation function gave the best fit to correlation function extracted from experimental data. The reconstruction eventually converges to a model that satisfied constraints imposed by the experimental data (see Figure 7). Donatelli and others have recently developed an iterative approach for model reconstruction that is similar to the hybrid-input-output algorithm for phase retrieval from oversampled single particle scattering data [20]. This approach is promising because it requires only a single phasing step, and converges rapidly, allowing high-resolution reconstruction, as demonstrated using simulation data.

Several obstacles are still limiting the application of FSS method. The first challenge is the treatment of solvent background. The solvent contribution has to be carefully characterized and corrected before extracting useful information from experimental data. The liquid injection method generates a thin water column, which may vary from shot to shot, due to fluctuations of the jetting column, making it difficult to quantify the solvent scattering. If a thin and reproducible sheet of solvent could be generated, similar to case of the cryoEM experiment, the background subtraction would be improved. However, the hit rate would be greatly reduced, since the concentration of particles per unit area in a sheet jet is much lower than that in the column of liquid from which the sheet originated. The other possibility is to use an aerosol injector as in single particle scattering experiments.

It is interesting to compare these two approaches (SPI or FSS) for noncrystalline particle imaging. Firstly, we note that for few-particles per shot, the dominant intensity modulations in the diffraction patterns will be due to coherent inter-particle interference. For example, with two identical particles, the patterns will be crossed with strong Young's fringes. These must be removed in order to apply the Kam method. Saldin has suggested that, because the interparticle spacings (and corresponding fringe spacings) vary over a wide range when correlation functions are summed, they will wash out [61]. Secondly, we need to find the optimum number of particles per shot. It has been shown that the addition of more particles per shot does not help improve the signal to noise ratio (SNR), which is independent of the number of particles per shot [62]. This occurs because both the signal (the angular correlation function) and the Poisson shot noise in the beam vary as the square root of the number of particles in the beam. This suggests that if the use of one particle per shot does not provide sufficient resolution, the addition of more particles will not improve matters. The addition of more patterns, however, does improve SNR. Finally, we may compare the hit rates and signal collection when using a large beam spanning many particles (FSS) with the same beam focused down to the size of one particle (with the same number of photons per shot, in SPI mode). There will be some inevitable lateral jitter in beam position, so that many of the shots in SPI will simply drill holes in the buffer solution between

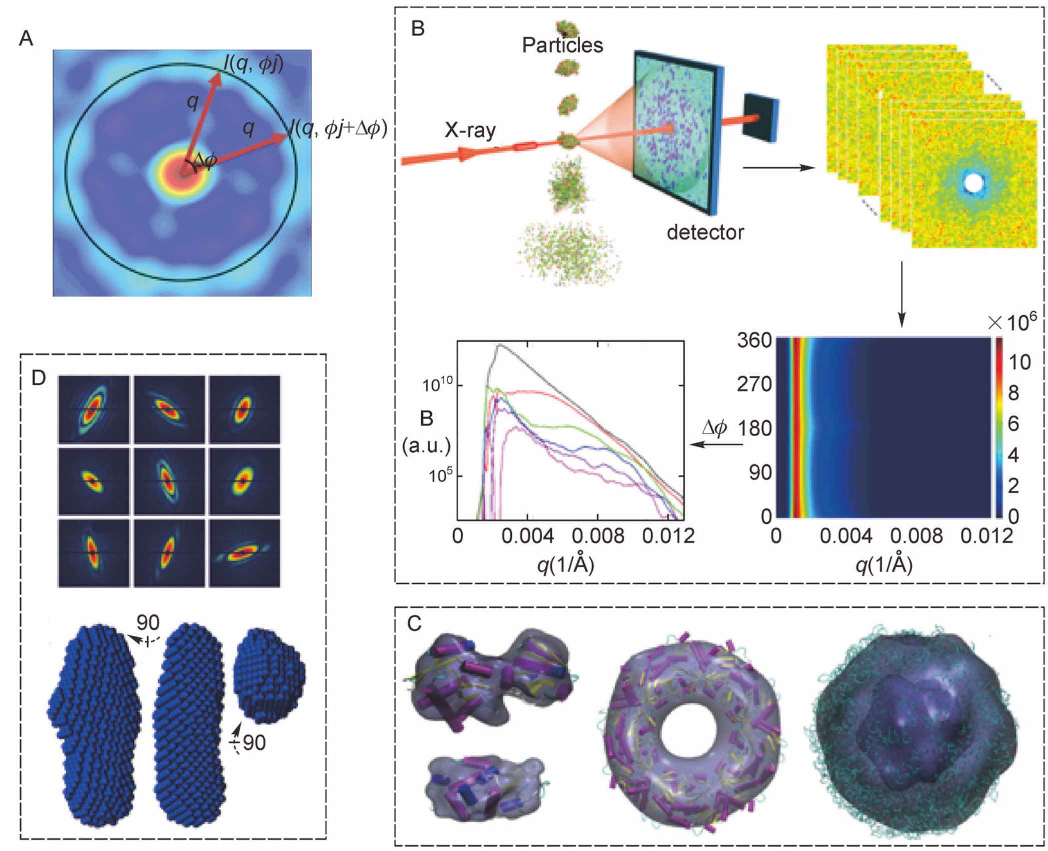


Figure 7. Fast solution scattering data analysis. (A)The correlation between two intensities separated by ΔΦ on the ring with scattering angle. (B) Data collection and correlation profile calculation. (C) Reconstructed models using reverse modeling approach from simulation data. (D) Scattering patterns from nanorice particles and the reconstructed 3D model (Figure reproduced from Liu *et al.*, 2013, Ref. [16] with permission).

sample particles. Clearly the hit rate for a defocussed wide beam in FSS mode is 100%, while the hit rate in SPI mode is low, due to fluctuations in beam position and, equivalently, the motion of particles flowing past the beam — many shots will pass between particles. Using simple cross sections for scattering and assuming a thin sheet jet wider than the beam, we may compare the total scattering for defocussed (FSS) and focused (SPI) beams with the same areal concentration of particles. We find that the total particle scattering intensity is the same in both cases. As the beam is broadens out, the hit rate increases, but the incident intensity per particle decreases, and these effects cancel.

The second challenge is to quantify the distributions of sample orientations. In Kam's original proposal, the distribution is assumed to be uniform in rotational space; therefore a summation of correlation functions calculated from each scattering pattern is equivalent to the integration over all orientations. However, this requires a very large number of scattering patterns. Furthermore, the molecules may have an orientation preference due to the experimental setup, such as a flow-alignment effect in a liquid jet. If the orientation distribution can be computationally quantified from a small amount of data, this can make the correlation scattering method more attractive and practical.

The third challenge is the treatment of interference between sample particles. XFELs provide full coherent X-ray beam, meaning that everything in the beam contributes to the scattering pattern coherently. The interference between particles is more crucial than in the case of SAXS or WAXS. Based on this observation, Kodama and Nakasako proposed an "X-ray diffraction microscopy" approach, which solves the phase problem from oversampled scattering patterns in 2D projections, which are then assembled to build 3D models using the existing cryoEM model reconstruction method [63]. This works only for the regime where Ewald sphere is approximately flat, such that the scattering patterns can be phased to 2D projections (intensities on curved Ewald sphere do not lie in a central slice of 3D reciprocal space). In another study, Kurta *et al.* showed that the coherent scattering from multiple particles can be treated in the same way as in the cases of incoherent scattering, as long as the interferences between particles are not correlated [18]. To rule out the correlated interferences, sample concentrations need to be very dilute, as in the SAXS experiments.

Finally, the most important challenge is the limited resolution found in FSS experiments, which is at best about 20 nm (simulation data shows sub-nanometer resolution can be achieved). Clearly, there is a large gap between actual experimental and simulations. We need to note that many experimental artifacts (such as stray scattering from apertures in the beamline, jet scattering, detector artifacts etc.) may produce anisotropic effects on the FSS patterns that can be much larger than those due to the Kam fluctuations from the sample itself. These noises are not well modeled in most simulation studies. Some of these effects can be reduced or corrected computationally. Researchers in imaging processing are encouraged to bring in various specialties to help advance the data analysis.

Dynamics

Dynamics study is the most exciting application area for XFELs, which not only extends the experimental limits to smaller crystals and single particles, but also provides much higher time resolution than that is possible at synchrotrons. At present, in order to obtain femtosecond time resolution combined with near-atomic resolution, one needs to take advantage of the “Bragg boost” as in SFX, whereby at its peak the intensity of Bragg scattering from a crystal is proportional to the sixth power of the number of molecules on a side of the crystal — thus one has a million times more scattering from a nanocrystals just ten molecules on a side than from a single molecule. This has been spectacularly demonstrated in the recent work of the Schmidt collaboration [64], where a molecular movie of the light-sensitive bacteria Purple Yellow Protein has recently been obtained with 0.16 nm spatial resolution and 10 ns time resolution (this has been extended to 100 fs time resolution in more recent unpublished work). The results show the presence of various stable intermediate molecular species, which appear and disappear during the photo-induced chemical reaction.

Smaller crystals allow faster and more uniform response to external stimuli, such as light activation, since the nanocrystal size is comparable with the optical absorption length. Single molecule or particles are ideal samples, allowing investigation of the enzymatic dynamics up to the time scale of the mixing diffusion limit. For nanocrystals in a mixing injector, diffusive mixing is therefore possible (this diffusion of a substrate into a submicron enzyme crystal sets the time-resolution limit of the method). A mixing jet, which will allow time-resolved tracking of chemical reactions in this way using an XFEL has been described in literature [11]. While pumping molecules to activated states using light is instantaneous and convenient, this pump-probe method is only applicable to proteins that are optically sensitive (either naturally or through genetic engineering). The single or multiple particle scattering methods therefore have great potentials in exploring the dynamics of molecules once the method is extended to higher resolution. The scattering from a large ensemble of molecules, i.e., SAXS or WAXS at XFELs, can reveal rich dynamics at high time resolution with computational modeling and simulations. This approach has been demonstrated in the work of Arnlund *et al*., who collected time resolved WAXS data using the LCLS and extracted the ultrafast picosecond global conformational changes of the photosynthetic reaction center [65]. The one dimensional WAXS data alone is not sufficient to allow a complete understanding of structure and dynamics. Based on an atomic structure determined from crystallography, the ground state is modeled as the initial structure for the simulation of protein dynamics after photon absorption. Molecular dynamics simulations were carried out for detailed analysis of the conformational changes of co-factor, protein, lipid and buffer, upon absorption of energy from a photon. In conjunction with the WAXS profiles, a clear molecular movie was compiled to explain the experimental data and the heat dissipation molecular mechanism. This approach can be generalized to the study of the dynamics and kinetics of many molecular machines. Another take home message from this case study is that information from a single type of experiment (X-ray Crystallography, NMR, SAXS/WAXS, SPI using X-rays or electrons, etc.) should be considered as incomplete; and different approaches probe molecules from its own perspectives. A generalized framework that integrates incomplete information will enable a full visualization of the molecular machines (see the work of constructing nuclear core complex by F. Alber *et al.* [66]). An important task in computational structural biology is to establish such an integrative modeling framework. For XFEL data analysis, the software and computational tools should be developed to facilitate this integration.

Other issues. Besides the major challenges to computational sciences in XFEL data analysis, there are several other instances where computational modeling can provide insights and predictions. (i) The effects of the X-ray intensity distribution at the beam focus. As the X-ray beam focus approaches the nanometer scale, the spatial profile of the X-ray beam becomes crucial for interpreting the scattering intensity from a molecule that is also at nanometer scale. The intensity distribution of the incident X-ray at the focal point is not flat, making the resulting scattering pattern depending on the sample molecule position in the cross section of the X-ray beam. One solution is to use a larger X-ray focus to flatten the incident photon distribution across the molecule, i.e., using FSS mode. (ii) X-ray pulse duration. Shorter X-ray pulses allow cleaner signals from intact molecules. On the other hand, shorter pulses (a few femtoseconds) yield smaller photon flux compared to longer pulses (such as 100 femtoseconds). The optimal pulse duration for a given sample requires systematic investigation, and information obtained from computational simulations will be very valuable. (iii) Real time monitoring and feedback. Preliminary data analysis using existing software, such as Cheetah, CASS, cctbx.xfel, and the programs developed at SLAC, provides information for improving experimental group during experiments. Converting preliminary data to actual models often requires many weeks after completing an experiment. If the models can be solved (even at lower resolutions) during the experiments, the throughput of each beam time can be significantly enhanced.

FUTURE PERSPECTIVES

The application of XFEL in structure biology has been extensively developed in the past five years. The computational tools and software have made it possible to process unprecedented amounts of data. The data deluge problem was not as severe as initially anticipated, partly because of the utilization of high performance computing facilities, mainly the HPC cluster at SLAC, and also because of fast "hit-finder" routines which can pick out useful data from blank shots. In future, higher data acquisition rates will be expected, for example 27,000 Hz at the European XFEL [67], if suitable detectors can be made to acquire data at this speed. The challenge to data analysis remains in this case. The computational problems mentioned here need better solutions. Scientists with computational skills should look for answers to this open question: what can computational sciences contribute to the development of XFEL applications? In the final section of this paper, we provide some suggestions. The thinking should not be limited by this list, which only summarizes major components of this field based on perspectives from the authors.

(1) Lower sample consumption. Improved algorithms for hit-finding, background correction, intensity refinement and merging can reduce the demands to the amount of experimental data.

(2) Sample delivery. Various sample delivery methods have been developed, including gas dynamic virtual nozzle, liquidic cubic phase injector, electronic spin injector, aerosol injector, and scanning through samples mounted on thin support or goniometer. The background scattering requires background correction algorithms. Real time monitoring also can help improve the hit-rate by maximizing the overlap between X-ray beam and sample streams.

(3) High throughput experiments. This requires the development of several configurations for standard experiments to allow quick switching between experimental modes and to reduce the time needed for fine-tuning. Real time monitoring and feedback from both preliminary results and model reconstructions are extremely valuable.

(4) Beam profile characterization. A clean beam path, knowledge of the intensity distribution at focus, and pulse duration, needs computational modeling and analysis to identify the optimal parameters and setup.

(5) Algorithm development and testing for data analysis, from raw data to model reconstruction. The coherent X-ray imaging data bank (CXIDB) was launched to aid such development by feeding the actual experimental data to computational science community [68]. The data from CXIDB has been used for several important method developments. Computational scientists who want to understand XFEL-related problems and make contributions are highly encouraged to study and analyze the data on this site in order to develop new methods.

(6) Combining experiments and simulations. The simulation and experimental time scales starting to overlap. This is an exciting moment for both fields. However many experimental parameters have yet to be fully parametrized in simulations, especially details of detector performance and metrology, the XFEL bandwidth and flux profile in each pulse, and the details (e.g., impact parameter, background) of the beam-specimen interaction. Eventually, we expect that the mechanisms of biomolecular interaction will be revealed by combining all sources of information from experimental data with computational modeling.

(7) Guide the development of *de novo* methods that can exploits XFEL properties, or help design future XFELs. Simulations allow *in silico* experiments before experimental instruments are available. The parameters obtained from such studies will be useful to design better equipped XFEL facilities, which is crucial because each facility may cost billions of dollars to build and to keep it running efficiently.

(8) The development of methods for collecting snapshot absorption and emission spectroscopy in registration with diffraction data has been an important development [69]. Already this has allowed us to track the chemical state of atoms during a reaction, in correlation with the

corresponding change-density maps, without damage, at atomic resolution and femtosecond time resolution for nanocrystalline samples. The extension of this to single-particle data is a major project for future development. This summary aims to lay out the major challenges for the computational sciences, which support XFEL data analysis in biology. It is hoped that more experts from different fields, such as information theory, image processing, "big-data" science, algorithms and physics, will now contribute to the development of structural biology using the unique capabilities of XFELs.

ACKNOWLEDGEMENTS

HG Liu acknowledges the support from the national 1000-talent program and the Natural Science Foundation of China (No.11575021). JCH Spence acknowledges the funding from NSF to BioXFEL STC center, award No. 1231306. The authors are grateful to Dr. I. Schlichting for critical comments and valuable discussions.

COMPLIANCE WITH ETHICS GUIDELINES

The authors Haiguang Liu and John C. H. Spence declare that they have no conflict of interests.

This article does not contain any studies with human or animal subjects performed by any of the authors.

REFERENCES

1. Solem, J. C. (1986) Imaging biological specimens with high-intensity soft x rays. J. Opt. Soc. Am. B, 3, 1551–1565
2. Neutze, R., Wouts, R., van der Spoel, D., Weckert, E. and Hajdu, J. (2000) Potential for biomolecular imaging with femtosecond X-ray pulses. Nature, 406, 752–757
3. Seibert, M. M., Ekeberg, T., Maia, F. R. N. C., Svenda, M., Andreasson, J., Jönsson, O., Odić, D., Iwan, B., Rocker, A., Westphal, D., *et al.* (2011) Single mimivirus particles intercepted and imaged with an X-ray laser. Nature, 470, 78–81
4. Emma, P., Akre, R., Arthur, J., Bionta, R., Bostedt, C., Bozek, J., Brachmann, A., Bucksbaum, P., Coffee, R., Decker, F.-J., *et al.* (2010) First lasing and operation of an angstrom-wavelength free-electron laser. Nat. Photonics, 4, 641–647
5. Chapman, H. N., Fromme, P., Barty, A., White, T. A., Kirian, R. A., Aquila, A., Hunter, M. S., Schulz, J., DePonte, D. P., Weierstall, U., *et al.* (2011) Femtosecond X-ray protein nanocrystallography. Nature, 470, 73–77
6. Liu, W., Wacker, D., Gati, C., Han, G. W., James, D., Wang, D., Nelson, G., Weierstall, U., Katritch, V., Barty, A., *et al.* (2013) Serial femtosecond crystallography of G protein-coupled receptors. Science, 342, 1521–1524
7. Redecke, L., Nass, K., DePonte, D. P., White, T. A., Rehders, D., Barty, A., Stellato, F., Liang, M., Barends, T. R., Boutet, S., *et al.* (2013) Natively inhibited *Trypanosoma brucei* cathepsin B structure determined by using an X-ray laser. Science, 339, 227–230
8. Spence, J. C. H. and Doak, R. B. (2004) Single molecule diffraction. Phys. Rev. Lett., 92, 198102
9. Spence, J. C. H., Weierstall, U. and Chapman, H. N. (2012). X-ray lasers for structural and dynamic biology. Rep. Prof. Phys., 75, 102601
10. Schlichting, I. (2015) Serial femtosecond crystallography: the first five years. IUCrJ, 2, 246–255
11. Wang, D., Weierstall, U., Pollack, L. and Spence, J. (2014) Double-focusing mixing jet for XFEL study of chemical kinetics. J. Synchrotron Radiat., 21, 1364–1366
12. Mayer, G. and Heckel, A. (2006) Biologically active molecules with a "light switch". Angew. Chem. Int. Ed. Engl., 45, 4900–4921
13. Kam, Z. (1977) Determination of macromolecular structure in solution by spatial correlation of scattering fluctuations. Macromolecules, 10, 927–934
14. Kam, Z., Koch, M. H. J. and Bordas, J. (1981) Fluctuation x-ray scattering from biological particles in frozen solution by using synchrotron radiation. Proc. Natl. Acad. Sci. USA, 78, 3559–3562
15. Saldin, D. K., Poon, H. C., Bogan, M. J., Marchesini, S., Shapiro, D. A., Kirian, R. A., Weierstall, U. and Spence, J. C. (2011) New light on disordered ensembles: *ab initio* structure determination of one particle from scattering fluctuations of many copies. Phys. Rev. Lett., 106, 115501
16. Liu, H., Poon, B. K., Saldin, D. K., Spence, J. C. H. and Zwart, P. H. (2013) Three-dimensional single-particle imaging using angular correlations from X-ray laser data. Acta Crystallogr. A, Foundations of crystallography, 69, 365–373
17. Pedrini, B., Menzel, A., Guizar-Sicairos, M., Guzenko, V. A., Gorelick, S., David, C., Patterson B. D., and Abela, R. (2013). Two-dimensional structure from random multiparticle X-ray scattering images using cross-correlations. Nat. Commun., 4, 1647
18. Kurta, R. P., Dronyak, R., Altarelli, M., Weckert, E. and Vartanyants, I. A. (2013) Solution of the phase problem for coherent scattering from a disordered system of identical particles. New J. Phys., 15, 013059
19. Saldin, D. K., Poon, H.-C., Schwander, P., Uddin, M. and Schmidt, M. (2011) Reconstructing an icosahedral virus from single-particle diffraction experiments. Opt. Express, 19, 17318–17335
20. Donatelli, J. J., Zwart, P. H. and Sethian, J. A. (2015) Iterative phasing for fluctuation X-ray scattering. Proc. Natl. Acad. Sci. USA, 112, 10286–10291
21. Barty, A., Kirian, R. A., Maia, F. R. N. C., Hantke, M., Yoon, C. H., White, T. A. and Chapman, H. (2014) Cheetah: software for high-throughput reduction and analysis of serial femtosecond X-ray diffraction data. J. Appl. Cryst., 47, 1118–1131
22. Foucar, L., Barty, A., Coppola, N., Hartmann, R., Holl, P., Hoppe, U., Kassemeyer, S., Kimmel, N., Küpper, J., Scholz, M., *et al.* (2012) CASS—CFEL-ASG software suite. Comput. Phys. Commun., 183, 2207–2213
23. Hattne, J., Echols, N., Tran, R., Kern, J., Gildea, R. J., Brewster, A. S., Alonso-Mori, R., Glöckner, C., Hellmich, J., Laksmono, H., *et al.* (2014) Accurate macromolecular structures using minimal measurements from X-ray free-electron lasers. Nat. Methods, 11, 545–548
24. Kirian, R. A., Wang, X., Weierstall, U., Schmidt, K. E., Spence, J. C. H., Hunter, M., Fromme, P., White, T., Chapman, H. N. and Holton, J. (2010) Femtosecond protein nanocrystallography-data analysis methods. Opt. Express, 18, 5713–5723
25. White, T. A., Kirian, R. A., Martin, A. V., Aquila, A., Nass, K., Barty, A. and Chapman, H. N. (2012) CrystFEL : a software suite for snapshot serial crystallography. J. Appl. Cryst., 45, 335–341
26. White, T. A. (2014) Post-refinement method for snapshot serial crystallography. Philos. Trans. R. Soc. Lond. B Biol. Sci., 369, 20130330

27. Sauter, N. K. (2015) XFEL diffraction: developing processing methods to optimize data quality. J. Synchrotron Radiat., 22, 239–248
28. Rossmann, M. G., Leslie, A. G. W., Abdel-Meguid, S. S. and Tsukihara, T. (1979) Processing and post-refinement of oscillation camera data. J. Appl. Cryst., 12, 570–581.
29. Ginn, H. M., Brewster, A. S., Hattne, J., Evans, G., Wagner, A., Grimes, J. M., Sauter, N. K., Sutton, G. and Stuart, D. I. (2015) A revised partiality model and post-refinement algorithm for X-ray free-electron laser data. Acta Crystallogr. D Biol. Crystallogr., 71, 1400–1410
30. Uervirojnangkoorn, M., Zeldin, O. B., Lyubimov, A. Y., Hattne, J., Brewster, A. S., Sauter, N. K., Brunger, A. T. and Weis, W. I. (2015) Enabling X-ray free electron laser crystallography for challenging biological systems from a limited number of crystals. eLife, 4, e05421
31. Zhang, T., Li, Y. and Wu, L. (2014) An alternative method for data analysis in serial femtosecond crystallography. Acta Crystallogr. A Found. Adv., 70, 670–676
32. Li, C., Schmidt, K. and Spence, J. C. (2015) Data collection strategies for time-resolved X-ray free-electron laser diffraction, and 2-color methods. Struct. Dyn., 2, 041714
33. Loh, N. D., Starodub, D., Lomb, L., Hampton, C. Y., Martin, A. V., Sierra, R. G., Barty, A., Aquila, A., Schulz, J., Steinbrener, J., *et al.* (2013) Sensing the wavefront of x-ray free-electron lasers using aerosol spheres. Opt. Express, 21, 12385–12394
34. Yefanov, O., Gati, C., Bourenkov, G., Kirian, R. A., White, T. A., Spence, J. C. H., Chapman, H. N. and Barty, A. (2014) Mapping the continuous reciprocal space intensity distribution of X-ray serial crystallography. Philos. Trans. R. Soc. Lond. B Biol. Sci., 369, 20130333
35. Brehm, W. and Diederichs, K. (2014) Breaking the indexing ambiguity in serial crystallography. Acta Crystallogr. D Biol. Crystallogr., 70, 101–109
36. Liu, H., & Spence, J. C. H. (2014). The indexing ambiguity in serial femtosecond crystallography (SFX) resolved using an expectation maximization algorithm. IUCrJ, 1, 393–401
37. Barends, T. R. M., Foucar, L., Botha, S., Doak, R. B., Shoeman, R. L., Nass, K., Koglin, J. E., Williams, G. J., Boutet, S., Messerschmidt, M., *et al.* (2014) *De novo* protein crystal structure determination from X-ray free-electron laser data. Nature, 505, 244–247
38. Spence, J. C. H., Kirian, R. A., Wang, X., Weierstall, U., Schmidt, K. E., White, T., Barty, A., Chapman, H. N., Marchesini, S. and Holton, J. (2011) Phasing of coherent femtosecond X-ray diffraction from size-varying nanocrystals. Opt. Express, 19, 2866–2873
39. Sayre, D. (1952) Some implications of a theorem due to Shannon. Acta Crystallogr., 5, 843
40. Fienup, J. R. (1982) Phase retrieval algorithms: a comparison. Appl. Opt., 21, 2758–2769
41. Kirian, R. A., Bean, R. J., Beyerlein, K. R., Barthelmess, M., Yoon, C. H., Wang, F., Capotondi, F., Pedersoli, E., Barty, A. and Chapman, H. N. (2015) Direct phasing of finite crystals illuminated with a free-electron laser. Phys. Rev. X, 5, 011015
42. Chen, J. P. J., Spence, J. C. H. and Millane, R. P. (2014) Direct phasing in femtosecond nanocrystallography. I. Diffraction characteristics. Acta Crystallogr. A Found. Adv., 70, 143–153
43. Chen, J. P. J., Spence, J. C. H. and Millane, R. P. (2014) Direct phasing in femtosecond nanocrystallography. II. Phase retrieval. Acta Crystallogr. A Found. Adv., 70, 154–161
44. Elser, V. (2013) Direct phasing of nanocrystal diffraction. Acta Crystallogr. A, 69, 559–569
45. Liu, H., Zatsepin, N. A. and Spence, J. C. H. (2014) *Ab-initio* phasing using nanocrystal shape transforms with incomplete unit cells. IUCrJ, 1, 19–27
46. Ayyer, K., Yefanov, O., Oberthür, D., Roy-Chowdhury, S., Galli, L., Mariani, V., Basu, S., Coe, J., Conrad, C., Fromme, R. (2015) Macromolecular imaging using scattering from disordered crystals. Nature, 530, 202–206
47. Ekeberg, T., Svenda, M., Abergel, C., Maia, F. R. N. C., Seltzer, V., Claverie, J. -M., Hantke, M., Jönsson, O., Nettelblad, C., van der Schot, G., *et al.* (2015) Three-dimensional reconstruction of the giant mimivirus particle with an x-ray free-electron laser. Phys. Rev. Lett., 114, 098102
48. Aquila, A., Barty, A., Bostedt, C., Boutet, S., Carini, G., dePonte, D., Drell, P., Doniach, S., Downing, K. H., Earnest, T., *et al.* (2015) The linac coherent light source single particle imaging road map. Struct. Dyn., 2, 041701
49. Deponte, D. P., McKeown, J. T., Weierstall, U., Doak, R. B. and Spence, J. C. H. (2011) Towards ETEM serial crystallography: Electron diffraction from liquid jets. Ultramicroscopy, 111, 824–827
50. Bortel, G. and Tegze, M. (2011) Common arc method for diffraction pattern orientation. Acta Crystallogr. A, 67, 533–543
51. Kassemeyer, S., Jafarpour, A., Lomb, L., Steinbrener, J., Martin, A. V. and Schlichting, I. (2013) Optimal mapping of x-ray laser diffraction patterns into three dimensions using routing algorithms. Phys. Rev. E Stat. Nonlin. Soft Matter Phys., 88, 042710
52. Scheres, S. H. W. (2012) RELION: implementation of a Bayesian approach to cryo-EM structure determination. J. Struct. Biol., 180, 519–530
53. Loh, N.-T. D. and Elser, V. (2009) Reconstruction algorithm for single-particle diffraction imaging experiments. Phys. Rev. E Stat. Nonlin. Soft Matter Phys., 80, 026705
54. Fung, R., Shneerson, V., Saldin, D. K. and Ourmazd, A. (2009) Structure from fleeting illumination of faint spinning objects in flight. Nat. Phys., 5, 64–67
55. Dashti, A., Schwander, P., Langlois, R., Fung, R., Li, W., Hosseini-zadeh, A., , Liaob, H., Pallesenc, J., Sharmab, G., Stupinad, V. *et al.* (2014) Trajectories of the ribosome as a Brownian nanomachine. Proc. Natl. Acad. Sci. USA, 111, 17492–7
56. Takahashi, Y., Suzuki, A., Zettsu, N., Oroguchi, T., Takayama, Y., Sekiguchi, Y., Kobayashi, A., Yamamoto, M. and Nakasako, M. (2013) Coherent diffraction imaging analysis of shape-controlled nanoparticles with focused hard X-ray free-electron laser pulses. Nano Lett., 13, 6028–6032
57. Kirian, R. A. (2012) Structure determination through correlated fluctuations in x-ray scattering. J. Phys. At. Mol. Opt. Phys., 45, 223001
58. Elser, V. (2011) Strategies for processing diffraction data from randomly oriented particles. Ultramicroscopy, 111, 788–792
59. Starodub, D., Aquila, A., Bajt, S., Barthelmess, M., Barty, A., Bostedt, C., Bozek, J. D., Coppola, N., Doak, R. B., Epp, S. W., *et al.* (2012) Single-particle structure determination by correlations of snapshot X-ray diffraction patterns. Nat. Commun., 3, 1276
60. Poon, H.-C., Schwander, P., Uddin, M. and Saldin, D. K. (2013) Fiber diffraction without fibers. Phys. Rev. Lett., 110, 265505
61. Saldin, D. K., Shneerson, V. L., Howells, M. R., Marchesini, S., Chapman, H. N., Bogan, M., Shapiro, D., Kirian, R. A., Weierstall, U., Schmidt, K. E., *et al.* (2010) Structure of a single particle from

scattering by many particles randomly oriented about an axis: toward structure solution without crystallization? New J. Phys., 12, 035014

62. Kirian, R. A., Schmidt, K. E., Wang, X., Doak, R. B. and Spence, J. C. H. (2011) Signal, noise, and resolution in correlated fluctuations from snapshot small-angle x-ray scattering. Phys. Rev. E Stat. Nonlin. Soft Matter Phys., 84, 011921
63. Kodama, W. and Nakasako, M. (2011) Application of a real-space three-dimensional image reconstruction method in the structural analysis of noncrystalline biological macromolecules enveloped by water in coherent x-ray diffraction microscopy. Phys. Rev. E Stat. Nonlin. Soft Matter Phys., 84, 021902
64. Tenboer, J., Basu, S., Zatsepin, N., Pande, K., Milathianaki, D., Frank, M., Hunter, M., Boutet, S., Williams, G. J., Koglin, J. E., *et al.* (2014) Time-resolved serial crystallography captures high-resolution intermediates of photoactive yellow protein. Science, 346, 1242–1246
65. Arnlund, D., Johansson, L. C., Wickstrand, C., Barty, A., Williams, G. J., Malmerberg, E., Davidsson, J., Milathianaki, D., DePonte, D. P., Shoeman, R. L., *et al.* (2014) Visualizing a protein quake with time-resolved X-ray scattering at a free-electron laser. Nat. Methods, 11, 923–926
66. Alber, F., Dokudovskaya, S., Veenhoff, L. M., Zhang, W., Kipper, J., Devos, D., Suprapto, A., Karni-Schmidt, O., Williams, R., Chait, B. T., *et al.* (2007) The molecular architecture of the nuclear pore complex. Nature, 450, 695–701
67. Vartanyants, I. A., Robinson, I. K., McNulty, I., David, C., Wochner, P. and Tschentscher, T. (2007) Coherent X-ray scattering and lensless imaging at the European XFEL Facility. J. Synchrotron Radiat., 14, 453–470
68. Maia, F. R. N. C. (2012) The Coherent X-ray Imaging Data Bank. Nat. Methods, 9, 854–855
69. Kern, J., Alonso-Mori, R., Tran, R., Hattne, J., Gildea, R. J., Echols, N., Glöckner, C., Hellmich, J., Laksmono, H., Sierra, R. G., *et al.* (2013) Simultaneous femtosecond X-ray spectroscopy and diffraction of photosystem II at room temperature. Science, 340, 491–495