Chapter 7

Climbing the data mountain: processing of SFX data

Chun Hong Yoon, Linac Coherent Light Source, SLAC National Accelerator Laboratory, 2575 Sand Hill Rd, Menlo Park, CA 94025 USA. yoon82@stanford.edu

Thomas A. White, Center for Free-Electron Laser Science, Deutsches Elektronen-Synchrotron DESY, Notkestraße 85, 22607 Hamburg, Germany. taw@physics.org

Correspondence may be addressed to either author

This is the authors' manuscript version of a chapter from "X-ray Free Electron Lasers — A Revolution in Structural Biology", Editors Sbastien Boutet, Petra Fromme and Mark S. Hunter, ISBN 978-3-030-00550-4 (print) 978-3-030-00551-1 (online). There are minor differences between this version and the final published version.

Abstract Serial femtosecond crystallography experiments produce mountains of data that require FEL facilities to provide many petabytes of storage space and large compute clusters for timely processing of user data. The route to reach the summit of the data mountain requires peak finding, indexing, integration, refinement, and phasing. Those who reach the summit get a *crystal clear* view of the radiation damage-free structure of a protein that is most consistent with the observed measurements. Data processing plays a critical role in the ability to measure accurate structure factor intensities from individual diffraction snapshots and combine them in 3-dimensional space. Current developments in SFX aim to take into account the huge complexity of SFX experiments, modeling variations in the beam and crystals, uncertainties in geometry, partiality, mosaicity, and figures of merit that are unique to SFX.

7.1 The data mountain

7.1.1 Why does Serial Femtosecond Crystallography produce so much data?

Serial crystallography represents a paradigm shift in macromolecular crystallography from the rotation method for data collection. It brings many benefits as described elsewhere in this book, however it also brings a steep increase in the data volume. The main reason for this is simple statistics. Systematic rotation of a single crystal allows all the Bragg peaks, required for structure determination, to be swept through and recorded. Serial collection is a rather inefficient way of measuring all these Bragg peak intensities because each snapshot is from a randomly oriented crystal, and there are no systematic relationships between successive crystal orientations. In this chapter, we will elaborate on the quantities of data required, and how one can climb this data mountain to yield valuable and meaningful results.

Consider a game of picking a card from a deck of all 52 cards until all the cards in the deck have been seen. The rotation method could be considered as analogous to picking a card from the top of the deck, looking at it and then throwing it away before picking the next, *i.e.* sampling without replacement. In this analogy, the faces of the cards represent crystal orientations or Bragg reflections. Only 52 turns are required to see all the cards in this case (Fig. 7.1 top), or analogously to acquire a complete dataset where all the symmetrically unique reflections (up to some resolution limit) have been measured. Serial collection is akin to randomly picking a card and then putting the card back in the deck before choosing the next card, *i.e.* sampling with replacement (Fig. 7.1 bottom). How many cards are needed to be drawn before all 52 have been seen? Intuitively, we can see there is no guarantee that all cards will ever be observed. However, statistically speaking, the expected number of turns to complete the task, c , is given by

$$c = n \sum_{k=1}^n \frac{1}{k},$$

where n is the total number of cards. For large n , c converges to $n \log(n)$. That is, for $n = 52$, it can reasonably be expected that all 52 cards will be observed only after about 236 turns! The problem is further exacerbated because a fraction of the images obtained in an SFX experiment will be blank because the x-ray pulse did not hit a crystal. This fraction varies depending on the sample preparation and delivery methods (see **Chapters 3-5**), but is often higher than 60%. The random orientation of crystals and the random picking of this orientation on every measurement represent the primary reasons why SFX data volumes are inherently larger than rotation series data.

The second reason why SFX data volumes are so high is the high variability of many experimental parameters. The self-amplified spontaneous emission (SASE) process, by which the x-ray pulses are generated in the FEL, essentially amounts to amplifying a random fluctuation in the electron bunch by many orders of magnitude. This randomness becomes imprinted on the x-ray pulses, such that each one

Fig. 7.1 Sampling without replacement (top) vs sampling with replacement (bottom) for a deck of cards. The expected number of turns to observe all cards by sampling one card at a time is 52 and 236, respectively. Figure ©The Author, licensed under CC-BY-4.0.

of them has a different intensity profile and photon energy spectrum. There may also be a wide variability in the crystals: their size, shape, crystalline order and even their crystal structure. In effect, each frame in an SFX experiment is from a completely separate experiment to the others. As described later in this chapter, great progress has been made in compensating for this variation, but the main method for mitigating it is still to average intensity measurements from a large number of crystals.

Over the years, FEL facilities have been built or upgraded with higher repetition rates and larger detectors which help reduce the data collection time, but do nothing to reduce the data quantity required.

7.1.2 Facilities, data rates and detectors

In 2005, a series of proof-of-principle studies were performed at the Free electron LASer in Hamburg (FLASH) facility in Hamburg, Germany, demonstrating “diffraction-before-destruction” [Chapman et al(2006), Barty et al(2008)], the concept of side-stepping classical radiation damage limits by using x-ray pulses shorter than the damage processes (See **Chapter 6**). These results motivated high resolution experiments using shorter wavelength x-rays. In 2009, the Linac Coherent Light Source (LCLS) in Stanford, USA, started producing x-rays in the sub-nanometer range, and the first SFX experiments on protein crystals were performed [Chapman et al(2011)], starting at a rate of 30 x-ray pulses per second. Shortly after,

the LCLS repetition rate was increased to 120 pulses per second with the data acquisition infrastructure capable of reading out 5 GB/s per instrument. Table 7.1 shows the repetition rates of the currently operating (and soon to be operating) x-ray FEL facilities. In 2011, SPring-8 Angstrom Compact free electron LAser (SACLA) in Japan soon followed as a compact x-ray FEL facility that can be operated below 1 Å wavelength. In 2017, Pohang Accelerator Laboratory XFEL (PAL-XFEL) in Korea and European XFEL in Germany started user operation. Superconducting accelerator technology has lead to much higher repetition rates that exceed the detector read-out rates. The European XFEL produces bursts of 2,700 pulses with only a few hundred nanoseconds between them, ten times a second, for a total of 27,000 pulses per second. LCLS-II, which is still a few years away, will produce a constant rate of x-ray pulses at a rate approaching 1 million pulses per second. **Chapter 16** will present an outlook of x-ray FELs and discuss these new machines further.

Table 7.1 Photon pulses per second at facilities around the world.

Facility	Pulses per second
FLASH (EUV and soft x-rays)	8000
LCLS	120
SACLA	60
FERMI (UV and soft x-rays)	50
PAL-XFEL	60
SwissFEL (from 2018)	100
European XFEL	27,000
LCLS-II (from 2020)	1,000,000

Data rates will also increase with improvements in detector technology. Detectors will become larger and be able to read-out data at higher rates. The short femtosecond pulses from x-ray FELs require the use of integrating detectors. Single-photon-counting detectors, such as the PILATUS [Kraft et al(2009)] and the EIGER [Casanas et al(2016)] detectors used at synchrotrons, offer many excellent features such as essentially zero background noise, but cannot be used with femtosecond pulses. The reason for this is that the ability to count photons requires that the electronic pulse from each photon be distinguishable from the others. When all the photons arrive within a few femtoseconds, a counting detector can only count zero or one photon. Instead, integrating detectors are used, in which the total charge created by the photons in the detector sensor is summed and read out after the pulse. Detectors used for crystallography are typically of multi-megapixel scale, which makes for formidable data rates at the repetition rates achievable for x-ray FELs. Detectors for SFX experiments need to have low noise and high dynamic range capable of capturing the Bragg peak intensities. The first detector specifically developed for LCLS was a 2.3 megapixel Cornell-SLAC Pixel Array Detector (CSPAD) which has a dynamic range of > 2500 photons at 8 keV in low gain mode (Fig. 7.2) [Carini et al(2014), Blaj et al(2015), Gruner]. Even with the high dynamic range, the Bragg spots produced from the ultra intense FEL pulses can saturate the detec-

Fig. 7.2 The Cornell-SLAC Pixel Array Detector (CSPAD) is an example of a PAD designed to operate at the conditions of a hard x-ray FEL. Photon counting PADs are not usable for femtosecond pulses unless the expected signal is ≤ 1 photon per shot. The CSPAD has a dynamic range of > 2500 photons at 8 keV in low gain mode with a total of 2.3 megapixels.

tor pixels and Fourier amplitude of the protein structure can not be determined in this case. The Rayonix (MX225-HS) detector at PAL-XFEL uses scintillators coupled to fibre-optics that indirectly transmit signal to a light sensitive camera to boost the dynamic range up to 45,000 photons at 12 keV. European XFEL is equipped with a smarter detector that can adjust the gain depending on the charge deposited on the pixel, called Adaptive Gain Integrating Pixel Detector (AGIPD), giving it both single-photon sensitivity and a dynamic range of more than 10,000 photons at 12.4 keV [Allahgholi et al(2015)]. Similarly, SwissFEL will have the adJUstiNg Gain detector FoR the Aramis User station (JUNGFRAU) which also automatically adjusts the gain depending on signal [Mozzanica et al(2014)].

Instead of using a beam stop placed in front of the detector, x-ray FEL detectors typically allow the beam to go through a central gap. This allows the beam to be stopped far downstream of the detector, where due to its typical divergence it will be much larger and cause less damage. It further allows the “spent” beam to be re-used for a second experiment in a downstream position, as shown in Fig. 7.3. The CXI beamline at LCLS operates in this configuration to double the number of experiments [Boutet et al(2015), Hunter et al(2016)]. A similar serial setup is planned for the SPB/SFX beamline at the European XFEL [Mancuso et al(2013)]. There are also other ways that make more efficient use of the x-ray beam, such as splitting the beam using a monochromator into “monochromatic” and “rejected” components, both of which can be used for SFX or other types of experiments [Zhu et al(2014)]. Since an SFX experiment is not very sensitive to pauses in the data acquisition, the pauses leading only to wasted sample as it flows past, the beam can be switched

between an SFX experiment on a timescale of seconds. A common application of these multiplexing techniques is to “screen” sample batches in parallel to data collection on a primary sample. Of course, all this acts to further increase the amount of SFX data.

Fig. 7.3 Overview of the serial SFX setup at the CXI beamline, LCLS; (a) Data are collected simultaneously in two sample chambers using a serial SFX setup. The x-rays are focused only using Be lenses with the Kirkpatrick-Baez mirrors (KB1) moved out. The x-rays enter the first sample chamber (SC1) and scatter from the protein crystal. The unscattered x-rays exit through the central hole in the CSPAD detector, and are then refocused for another scattering experiment in the downstream sample chamber (SC3) followed by the diagnostic (Diag). Reproduced from Hunter et al, “Selenium single-wavelength anomalous diffraction de novo phasing using an X-ray-free electron laser”, in *Nature Communications* 7:13, 388 (2016).

In an SFX experiment, the meeting of a crystal with an x-ray pulse is a chance event. As a result, there will be blank shots in an SFX data set. The fraction of detector readouts that correspond to a crystal interacting with the pulse and producing a diffraction pattern is known as the hit rate and can be controlled by altering parameters such as the concentration of crystals or the thickness of the liquid jet, for example. To ease the data processing, we normally aim to maximize the number of “single hits,” where only one crystal is hit by the x-ray pulse. The theoretical maximum for this rate is 37%, given by Poisson statistics as shown in Fig. 7.4, at which 23% of the hits will contain more than one overlapping diffraction pattern to produce an overall hit rate of 60% [Hunter et al(2014)]. In most SFX experiments performed to date, the actual hit rate has been much lower, with a large fraction of blank shots. The first stage of data processing, described in the next section, is to identify the hits so that the blank frames can be ignored in the later stages of processing. In the high-throughput regime of future x-ray FELs, it may become necessary to perform hit finding prior to writing it to persistent storage, *i.e.* immediately and permanently deleting the blank frames.

Fig. 7.4 Optimum hit rate for most single hits assuming Poisson statistics [Hunter et al(2014)]. Image ©Takanori Nakane, reproduced by kind permission.

7.2 Reducing the mountain to a hill: Hit finding

7.2.1 Realities of experimental data

The aim of hit finding in SFX is to determine whether the snapshot contains Bragg spots or not. All the later processing stages are based on Bragg spots, and so frames which do not contain any of them are useless, at least as far as crystallographic data processing is concerned. Conceptually, hit finding seems trivial. However, in practice it can be challenging.

In an ideal case shown in Fig. 7.5a, the peaks are intense and there is no background noise. In this case, even a simple thresholding algorithm can locate the peaks. Unfortunately, real life is not so simple, and there are many additional features on the detector. The medium in which the crystals are embedded, necessary for them to stay hydrated, leads to coherent scattering. X-ray fluorescence from the sample can be significant for some elements and some photon energies, as can parasitic scattering from x-ray apertures and focusing optics. These effects all combine to give an overall background as shown in cyan in Fig. 7.5b. The detector will inevitably contain a small number of defective pixels, producing either very high or very low readouts.

Fluctuations due to beam fluence, photon energy and crystal size affect both the signal and background which makes it hard to set the right peak finding parameters that will work for all the images. Given that each event is unique, peak finding parameters will also be unique for each event. Indeed, a grid search over the peak

finding parameter space for each individual image can yield better results than a one-size-fits-all approach [Lyubimov et al(2016)]. Shadowing due to particular experimental setup, such as upstream apertures or diagnostics including viewing camera optics close to the beam, can add abrupt non-uniform changes to the background, as shown in green in Fig. 7.5b. Multiple crystals in the beam can introduce extra Bragg spots that influence the local noise estimation (Fig. 7.5b). Some of these artifacts can be reduced by a series of calibration steps, as described below.

Fig. 7.5 Illustration of challenges associated with peak finding. (a) Ideal case, the Bragg spots (magenta) have no background noise apart from the detector noise due to pedestal (gray) and common mode (red). (b) Background scattering (cyan) from optics/crystal buffer/fluorescence can be significant. Bad pixels can often be brighter than the Bragg spots. Shadowing due to the chamber setup such as upstream aperture/camera wires add abrupt changes, non-uniform noise to the background (green). Multiple crystals in the beam can introduce extra Bragg spots that influence the local noise estimation (blue). Figure ©The Author, licensed under CC-BY-4.0.

7.2.2 Correcting detector artifacts and removing background

Raw pixel values are read out by the detector for each event and calibrated in the following order: pedestal correction, common mode correction, and gain correction. Pedestal refers to a large additive noise due to the electronic readout that is fairly constant over time (which slowly drifts during the course of an experiment); pedestals are independent of signal being measured. Before and during an exper-

iment, so called dark images are collected (detector on, x-ray off) to measure the average pedestal. Dark runs also provide information about bad pixels.

Common mode is another form of additive noise that arises when a detector tile is bombarded by many photons producing a 'common' offset for all pixels in the tile (Fig. 7.5a shown as red). Since this is signal-dependent noise, each tile has to be corrected per event basis by evaluating the offset experienced by the pixels with no photons. If all the pixels are illuminated the common mode can not be evaluated. A way to avoid this problem was devised in the CSPAD detector where unbonded pixels are strategically placed to readout the offset. Unbonded pixels are pixels where the x-ray sensor material is not connected to the electronics of the pixel and therefore can not detect photons but still experience the common offset. It is important to carefully mask out the unbonded pixels during analysis.

Gain correction normalizes the response of the individual pixel. This is multiplicative noise that is proportional to the photons seen by the pixels.

For an i^{th} event, the calibration step can be expressed as:

$$C_i = \frac{R_i - P - CM_i}{G}, \quad (7.1)$$

where C_i is the i^{th} calibrated image, R_i is the i^{th} raw image, P is the pedestal, CM_i is the common mode due to R_i , and G is the gain factor.

The calibrated images are (to first approximation) devoid of electronic noise. However, the diffraction patterns still contain other forms of noise, such as the solvent scattering from crystal buffer, secondary scattering from experimental components around the interaction region, fluorescence and jet streaks from the injector that all contribute to the background noise making peak finding difficult. The term background noise used here is defined as any photons on the detector that are non-Bragg reflections. In severe cases, background subtraction algorithms are used to suppress slowly varying noise. There are largely two approaches; radial background subtraction and median background subtraction. Given that the dominant background comes from solvent scattering, the isotropic component of the background noise can be removed by subtracting the radial average of all pixels that are equidistant from the detector center. This algorithm requires *a priori* knowledge of the beam center and detector pixel positions which may not be available. Median background subtraction takes a different approach where a 2D sliding window of pixels is used to calculate the median intensity inside this window and subtracted from the pixel value at the center of the window. This algorithm suppresses both the isotropic and non-isotropic components of the background noise, and also does not require any geometric information.

7.2.3 Finding peaks and identifying hits

There are several software packages that can perform peak finding at FEL facilities, most notably (in alphabetical order) CASS [Foucar(2016)], cctbx.xfel [Hattne et al(2014)],

Cheetah [Barty et al(2014), Nakane et al(2016)], and psocake [Thayer et al(2017)]. All these packages implement their own unique algorithms that share common image processing ideas:

1. Locate all local maxima in a detector tile as potential Bragg spots. Brightest connected pixels of the local maxima are counted as signal and surrounding pixels as the estimation of local background noise.
2. Calculate peak properties, such as the signal-to-noise ratio (SNR), area of the spot, and sum of pixel values. The SNR is often evaluated to determine whether the peak found is significant enough to be a Bragg spot or just spurious noise

$$\text{SNR} = I/\sigma(I) = \mu_s/\sigma_n,$$
 where μ_s is the mean of the signal S and σ_n is the standard deviation of noise estimate N . This SNR should not be confused with $I/\sigma(I)$ used in crystallography tables for data collection and refinement statistics. Poisson statistics of photon arrival dictates the noise associated with the measurement of a Bragg spot intensity (I),

$$I = \text{var}(I) = \sigma^2(I).$$
3. Return centroid of the peaks as peak positions if the peak properties meet the user-defined criteria.

A diffraction pattern is considered to be a hit if it has more than some significant number of peaks, typically around 15. A much lower number of reflections, as low as 2, may in principle be sufficient to determine the orientation of the crystal. However 15 is a heuristically determined practical minimum number. Accurate peak finding is important; erroneous peak detection with many false negatives (i.e. unable to find Bragg peaks that are present) obviously leads to less hits being found ultimately leading to a poorer protein structure. On the other hand, peak detection with many false positives (i.e. finding peaks that are not Bragg peaks) artificially increases the number of hits found, giving an illusion of having collected more crystal data.

In Fig. 7.6, the peaks found are indicated by cyan squares. Note that some Bragg spots may have intensities close to zero, because the Bragg spot intensities are modulated by the molecular transform of the protein, which have zero intensities, and also systematic absences due to a screw axis of the unit cell can make certain Bragg spots disappear. Predicted lattice points are also shown in magenta rings which will be discussed later (section 7.3.1). Identification of hits is an active area of research required for rapid structure solution during beamtime and may one day be performed more accurately and efficiently with machine learning techniques.

The peak finding software packages mentioned above interface with various FEL facilities to access the diffraction patterns and save the crystal hits in a predefined data format that is easily accessible by the user regardless of where the data was collected.

Fig. 7.6 Representative peak finding (cyan) and indexing results (magenta) from a selenobiotinyl streptavidin crystal [Yoon et al(2017)] at the CXI instrument, LCLS. Figure ©The Author, licensed under CC-BY-4.0.

7.2.4 Facility frameworks and data formats

The large data volumes and high rates of acquisition calls for dedicated systems for recording, storing and accessing the data. Instead of simply providing the data on disk in a familiar file format, most x-ray FEL facilities provide a software framework for accessing the data. The access may be "online" — during the data acquisition process itself on shared memory with minimal delay — or "offline", sometime afterwards when the data arrives on disk. LCLS provides a graphical online monitoring tool called the Analysis Monitoring Interface (AMI) [Thayer et al(2017)], and a software framework "psana" (Photon Science ANALysis) [Damiani et al(2016)] for online and offline analysis. Psana offers programming interfaces in both C++ and Python, and allows access to all data generated by the data acquisition system. Software packages such as Cheetah [Barty et al(2014)], OnDA [Mariani et al(2016)], IOTA [Lyubimov et al(2016)] and cctbx.xfel [Hattne et al(2014)] use these interfaces to access the data and perform tasks such as calculating "live" hit rates and Bragg spot saturation levels, online, or writing the hits out to separate files, offline. European XFEL is the other facility that has developed a dedicated software framework, Karabo [Heisen et al(2013)] for accessing and analysing data. SACLA offers

its users a crystallography data processing pipeline through the SACLA data acquisition application programming interface (API) [Nakane et al(2016)].

For storing hits, the HDF5 (Hierarchical Data Format, version 5) file format is commonly used across all of FEL science. This is a "container format" allowing a large amount of flexibility in the layout and representation of the data structures. For scattering data, a predefined schema by CXIDB [Maia(2012)] or NEXUS [Konnecke et al(2015)] is often used.

7.3 Indexing, integration and merging

Having brought the data volume under control by isolating the patterns containing apparently useful diffraction signal, the next steps of data processing are concerned with turning those diffraction patterns into structure factor estimates which can be used by established crystallography software. Several software packages are now available to perform the processing, notably CrystFEL [White et al(2013), White et al(2016)], cctbx.xfel [Hattne et al(2014)], nXDS [Kabsch(2014)] and cpxfel [Ginn et al (2016a)]. These packages have differences in their implementation, for instance in how they interface with the facility's data processing framework, but are alike in more ways than they are different. They all sequentially perform the essential processing steps of indexing, integration, merging and finally evaluating the data quality. Each of these steps is described below.

7.3.1 Indexing

Indexing a diffraction pattern means assigning Miller indices to the Bragg peaks in the diffraction pattern. Implicitly, this involves determining the lattice parameters of the crystal and its orientation relative to a reference ("laboratory") frame. It also acts as a powerful filter of data, because it is very unlikely that a frame containing only spurious peaks will pass through this process successfully.

Several algorithms and pieces of software exist for indexing rotation series data in macromolecular crystallography [Powell(1999), Duisenberg(1992), Kabsch(1988)]. It has been found that the same algorithms and software are able to index snapshot diffraction patterns, without the advantage of recording a three-dimensional wedge of reciprocal space [Kirian et al(2010)].

In a serial crystallography experiment, we can usually assume that each crystal has very similar lattice parameters. However, the indexing algorithms usually determine the parameters *ab initio* for each diffraction pattern. Each indexing result must therefore be checked for consistency with a reference set of lattice parameters provided by the user. If the parameters are unknown, they can be determined by plotting histograms of each of the parameters (a , b , c , α , β and γ) and finding the most common values for a representative part of the dataset. SFX data processing programs

provide graphical tools to assist this process. The indexing process can then be repeated using the parameters so determined. Some indexing algorithms can make use of prior information about the lattice parameters to increase their success rate [Gildea et al(2014)], or even require this information to work [Ginn et al(2016)].

Once the lattice parameters and orientation of the crystal have been determined, they can be used to calculate the positions where Bragg peaks are expected to appear on the detector. Successful indexing relies on having an accurate description of the detector geometry, which means that the positions of the detector panels (see section 7.1.2) must be known accurately and precisely. Provided that the initial geometry is accurate enough to index at least a few patterns, it can be refined by comparing the observed and calculated positions of peaks on the detector. Since the indexing solution uses information about spot positions from the entire detector, the calculated peak positions can be taken as a reference, and a mispositioned panel will show a systematic offset between the observed and calculated peak locations [Hattne et al(2014)]. After correcting the panel locations, the indexing process can be repeated until the detector geometry is known with high precision [Yefanov et al(2015)]. This process is made easier by the use of a strongly diffracting and readily available calibration sample such as lysozyme or thermolysin.

If the distance between the sample and detector (the ‘camera length’) is set incorrectly in the description of the detector geometry, this will manifest as a systematic offset in the spot positions [Yefanov et al(2015)]. Small offsets of the camera length can also be seen in the histograms of lattice parameters. The peaks will be sharpest for the correct values, and become wider or even bimodal as the camera length deviates further from the true value.

If the concentration of crystals in the delivery medium is high, there may be a significant number of frames that contain multiple diffraction patterns (see section 7.1.2). Most indexing algorithms assume that all the peaks in the diffraction pattern come from one lattice, and so can fail when presented with two or more overlapping lattices. As a result of this, algorithms for multiple lattices have recently been developed. Some of these are based on the delete and retry method. Here, the pattern is indexed assuming that it is a single lattice, and the peaks which could be accounted for by the resulting lattice removed from the peak list prior to making another indexing attempt using the remaining peaks. This algorithm has been available in `cctbx.xfel` since the earliest released versions [Hattne et al(2014)], and has been extended to larger numbers of lattices [Ginn et al(2016), Gildea et al(2014)].

This delete and retry method has advantages of simplicity, not requiring much extra code to be added to software, however it relies on the first indexing attempt succeeding in finding one of the lattices despite the extra peaks from other lattices. Another algorithm, called “FELIX”, has recently been developed which operates on completely different principles [Schmidt et al(2017)]. Instead of treating the image as if it contained only one lattice at the outset and finding subsequent lattices in sequence afterwards, the FELIX algorithm assumes that there are multiple lattices present, and searches for them simultaneously. It can index large numbers of overlapping lattices from a single snapshot: ten or more depending on the resolution of the patterns and the unit cell parameters.

The process for determination of the lattice parameters, described above, is quite a coarse one. Populations of crystals with different parameters will be identified only if at least one of the parameters is different from the rest of the population by more than the width of the distribution of that parameter. A *Data Exploration Toolkit* has been written [Zeldin et al(2015)], which, amongst other operations, can perform hierarchical clustering according to the Andrews-Bernstein distance metric to compare lattice parameters. Clustering the crystals in this way can reveal populations of crystals with subtly different lattice parameter, and appeared to reduce the number of outlying intensity values.

Some crystal symmetry classes produce an extra complication with serial data acquisition. These are the classes which allow the crystal structure to be rotated, usually by 90 or 180 degrees, in such a way as to overlap the reciprocal lattice points with their original locations while the structure itself does not look the same. Tables of which symmetry classes are affected can be found elsewhere [White et al(2013)]. As well as exact overlaps, there can also be approximate overlaps for certain unit cell parameters [Barends et al(2015)]. In these cases, the indexing solution is ambiguous between two (or, rarely, more) possibilities. While indexing is usually a purely geometrical procedure, which uses only the positions of the reflections, indexing ambiguities can only be resolved by additionally comparing the intensities of the reflections. Because of the large amounts of noise in the individual measurements, this was a significant problem in the first SFX experiments. However, the Brehm-Diederichs algorithm was later developed, which applies a clustering scheme to the intensities and effectively resolves the ambiguities [Brehm and Diederichs (2014)]. This algorithm, or a simplified variant of it [White et al(2016), Ginn et al(2015b)], has now been implemented in all SFX data processing software packages.

7.3.2 *Integration and Merging*

The main aim of data processing in this chapter is to measure the structure factor moduli, which manifest themselves in the intensities of the Bragg peaks. There are many confounding factors affecting the relationship between the two, some of which were mentioned in section 7.1.1. Nevertheless, the process begins by measuring the intensity above the background level at a location determined using the reciprocal lattice basis vectors (lattice parameters and orientation) determined by the indexing algorithm. There are several methods for doing this, the simplest of which, and the most commonly used one in SFX, being to estimate the background level from an annular region around the peak location and the intensity from a circular region inside it, as discussed in Section 7.2.3 and shown in Figure 7.7. The average background level is subtracted from each pixel value in the peak region, and their sum calculated. Several other techniques have been used. These include two dimensional profile fitting, where the shape of each peak is fitted using either an average of the shapes of the strong peaks [Rossmann(1979)] or a shape calculated from the properties of the crystals [Kroon-Batenburg et al(2015)], or even

using the crystal parameters to calculate exactly which pixels should contain signal [Hattne et al(2014)].

Fig. 7.7 Detailed view of a reflection integration ‘shoebox’. Detector pixels are shown as squares in a grid, and pixel intensity values by the darkness of their shading. Note that the integration fiducial, which is the calculated location of the reflection, is not aligned with the pixel grid. Note further that it does not necessarily coincide with the centroid of the actual peak, because the detector geometry, cell parameters, crystal orientation and other geometrical parameters may not be perfectly determined. Figure ©The Author, licensed under CC-BY-4.0.

Since the structure factor moduli can have very small values as well as large ones, this integration procedure must be performed even when no obvious peak is present. This introduces some additional considerations. The diffraction model, which includes the indexing solution as well as estimates of crystal parameters such as mosaicity and crystal size, must be as accurate as possible to avoid missing the true reflection locations. It also needs to be accurate to avoid making a large number of false measurements, integrating reflections which are not truly excited in the diffraction pattern (regardless of their structure factors). Most current software therefore performs a refinement stage, where these parameters are refined before integrating the reflections [White et al(2016), Ginn et al(2015b), Sauter et al (2014)].

Compared with single-particle x-ray imaging, as described in **Chapter 14**, crystallography has the great advantage that the indices for each intensity measurement can be calculated geometrically, as described in the previous section. The intensities themselves are not important for indexing. The intensity measurements can be averaged together in ‘buckets’ according to their indices (except for the indexing ambiguities previously discussed), and this averaging process is very powerful: apart from certain types of systematic effect, any confounding factor reducing the precision of the individual measurements can be overcome by using more measurements. For the initial experiments, hardly any attempt was made whatsoever to overcome any of these factors, and this approach was referred to as ‘Monte Carlo Integration [Kirian et al(2010)]’ because of its similarity to the numerical Monte Carlo integration procedures for calculating integrals. Since then, many techniques have been developed to compensate for the factors affecting the intensity of each peak, which are described in the next section.

The progress of the Monte Carlo procedure can be tracked by plotting a self-consistency figure of merit such as R_{split} (described in section 7.3.5) against the number of crystals, n . R_{split} is proportional to $1/\sqrt{n}$, with the constant of proportionality depending on the dataset.

7.3.3 *Scaling the intensity measurements*

Getting the most out of the data means modeling and accounting for as many aspects of crystal and x-ray pulse variation, mentioned in section 7.1.1, as possible. Perhaps the most obvious ways to begin doing this is to scale the intensities to account for variations in the pulse intensity and crystal size. The weaker intensities from smaller crystals can be scaled up, and the stronger intensities from larger crystals scaled down, to bring everything to a common scale. This can be done using similar algorithms to rotation crystallography with a synchrotron or home source. Some extra considerations are needed because of the large crystal-to-crystal variations between frames, such as using logarithms of scaling factors instead of the scaling factors themselves, to make the calculation more numerically stable [Kabsch(2014)]. These methods can be extended to determine an effective Debye-Waller parameter for each crystal. This accounts for the falloff of intensity as resolution increases, which is greater for less well-ordered crystals. The high-resolution reflections from less ordered crystals can then be scaled up relative to others [White et al(2016), Uervirojnangkoorn et al(2015), Sauter(2015)].

As was mentioned above, accurate values are needed for the parameters affecting which spots are to be integrated, to avoid ‘over-prediction’ or ‘under-prediction’. Automatically determining or refining these parameters for each crystal is another way to improve the modeling of the diffraction process, and has been found to improve the data quality [Uervirojnangkoorn et al(2015), Sauter(2015)]. If the orientation of the crystal is not accurate, it is common to find that the prediction parameters, such as the spectral bandwidth of the x-ray pulse, is over-estimated by the software

in an attempt to fit the visible peaks despite the inaccurate orientation. These parameters have therefore been successfully used as a figure of merit in a grid search technique to determine the optimum processing parameters [Lyubimov et al(2016)]. Combining geometrical refinement and parameter auto-determination techniques with scaling produces a significant combined improvement [White et al(2016)].

7.3.4 *Partiality and post-refinement*

The idea of compensating for reflection *partiality* has seen a lot of discussion since the first x-ray FEL SFX experiments [Chapman et al(2011)]. The ‘partiality’ of a reflection in a given diffraction pattern is a quantity which describes the fact that not all regions of the crystal, nor all the x-rays in the beam, contribute simultaneously to the reflection. A partiality of 1 would mean that the entire crystal and all of the x-rays contribute to the reflection. The ideal situation, leading to the most straightforward data processing, would be for all reflections to have the same partiality. However, there are several reasons why they are not all equal.

Alongside the factors mentioned in section 7.3.3, partiality is suspected to be a strong determiner of data quality in SFX. However, several schemes for modeling the diffraction conditions have been proposed, and do not yet appear to agree on the most appropriate model. The simplest model is shown in Figure 7.8, and has been used in the nXDS software [Kabsch(2014)] and cctbx.prime [Uervirojnangkoon et al(2015)]. In this model, the x-ray pulses are assumed to be monochromatic, so the Ewald sphere is an infinitely thin shell. The partiality of a reflection is simply related to the distance between the reciprocal lattice point and the Ewald sphere. To compensate for the partiality, reflections which are further away from the Ewald sphere must be scaled up by a larger factor than ones closer to the exact Bragg condition. The model can be extended by including a description of the disorder of the crystal. Protein crystals are usually modeled as being made up of a large number of mosaic blocks, each perfectly ordered within itself, but with an orientational disorder between them. In this case, not all of the mosaic blocks need be in the Bragg condition at the same time. Small rotations of the crystal, as performed during rotation crystallography, would allow all the mosaic blocks to be sampled and therefore can record the intensity from the entire crystal. However, with x-ray pulses on the femtosecond timescale, there is no possibility to do this. Therefore, the partiality model includes an angular “smearing” of the reciprocal lattice points, which become larger with increasing distance from the origin of reciprocal space.

The x-ray beam used for SFX experiments is typically not monochromatic, but has a small bandwidth of around 0.1%. In this case, not all the wavelengths may satisfy the Bragg condition. The same applies if the incident x-ray beam is not completely collimated and therefore has a convergence or divergence angle, which is also usually the case due to the focusing optics. Furthermore, these effects may be convoluted with those mentioned above in a single snapshot. For example, a subset of mosaic blocks may be at the exact Bragg condition for one wavelength, while

Fig. 7.8 Simple model of reflection partialities using a monochromatic x-ray beam. Point "O" is the origin of reciprocal space, and the small discs represent reciprocal lattice points. Point A is closer to the Ewald sphere than point B, therefore point A has a higher partiality than B. If the two reflections were to have the same structure factor, point A would produce a brighter reflection in the diffraction pattern. Figure © The Author, licensed under CC-BY-4.0. Reproduced from T. A. White, "Processing of XFEL data", in "Protein Crystallography", Wlodawer, Dauter and Jaskolski (eds), Methods in Molecular Biology 1607 (2017).

another subset of mosaic blocks may satisfy it for a different wavelength. A model based on a finite bandwidth x-ray beam has also been used successfully with x-ray FEL data [Ginn et al(2015b)].

Partialities are different for each reflection and each diffraction pattern. However, we can assume that the underlying structure factors should be the same among all patterns for each symmetrically unique reflection. The partialities are strongly affected by geometrical parameters such as the orientation of the crystal. This allows a refinement procedure to be performed where the geometrical parameters are optimized for each diffraction pattern in turn, aiming to maximize the fit between the underlying structure factor estimates derived from each pattern by modeling the partialities. This process is called post-refinement. Post-refinement has been implemented for rotation crystallography for many years [Rossmann et al(1979)]. For snapshot data, the first step was to determine whether the process could be performed stably with simulated data, which was found to be the case [White(2014)]. The first implementation of post-refinement aimed at experimental SFX data was by [Sauter(2015)] in `cctbx.xfel`. This was followed by the software `cctbx.prime`, which interfaces with `cctbx.xfel` and is developed specifically for post-refinement and merging, including scaling steps as well. Partiality modeling and post-refinement has also been implemented in `cppxfel` [Ginn et al(2015a)]. Whereas `cctbx.xfel` and `cctbx.prime` assume, at least in the versions described by recent literature, that the x-ray pulses are monochromatic, `cppxfel` uses a different geometrical model of the

diffraction process where the spectral bandwidth of the x-ray pulse is the dominant resolution-dependent effect.

Finally, comparing intensity values is not the only way to refine the factors affecting partiality. The EVAL15 software [Schreurs et al(2010)] instead compares simulated and experimental peak shapes and positions for each reflection, using a ray tracing method to model the diffraction geometry. This approach has been successfully tested on snapshot data from a laboratory source [Kroon-Batenburg et al(2015)], and is now being investigated for SFX data.

7.3.5 Evaluating the data quality

Many of the conventional figures of merit for a crystallographic dataset also apply to SFX data. These include the intensities compared to their estimated errors ($I/\sigma(I)$) and the number of measurements per symmetrically unique reflection, for example. Due to the large errors in an individual intensity measurement, arising from all the causes mentioned above including SASE fluctuations, difficulties accurately indexing snapshots and un-modeled partialities, the minimum number of measurements per reflection considered acceptable in SFX is much higher than in rotation crystallography, where a single measurement can suffice.

The most widely-used figures of merit for SFX data are those based on the self-consistency of the dataset. In the ideal case, all the factors introducing noise into the individual intensity measurements would be corrected for, thereby obtaining ‘perfect’ data from only one measurement per symmetrically unique reflection. In this case, repeating the experiment and data processing under identical conditions would produce an identical set of intensity measurements. In practice, some amount of variation between separate measurements of the same reflection must be accepted. By quantifying this variation, we can estimate how much random error is contained in the data, which places a limitation on how well the model can fit the data. The variation can be quantified by splitting the experimental dataset into two halves, alternating patterns to avoid systematic variations between the start and end of the experiment, merging each one separately, then comparing the two half-datasets using a correlation coefficient, R-factor or other metric. $CC_{\frac{1}{2}}$, the name given to the correlation coefficient between two half-datasets, was introduced and is closely related to CC^* for conventional crystallography [Karplus and Diederichs(2012)]. R_{split} expresses the same correlation as an R-factor, given by

$$R_{\text{split}} = \frac{1}{\sqrt{2}} \frac{\sum |I_{\text{even}} - I_{\text{odd}}|}{\frac{1}{2} \sum (I_{\text{even}} + I_{\text{odd}})}. \quad (7.2)$$

This formula divides the total absolute difference between the intensity from the ‘odd’ and ‘even’ half-datasets by the total of their mean values. Dividing by $\sqrt{2}$ aims to adjust the figure of merit to account for the half-datasets containing only half the amount of data as the full dataset, to give an estimate of the error in the

complete dataset. As has been described earlier, R_{split} usually decreases in inverse proportion to the square root of the number of diffraction patterns.

Several different methods have been proposed for estimating the standard error in each individual merged intensity. In CrystFEL, this is done by measuring the standard deviation of the individual measurements for one symmetrically unique reflection, then dividing this by $\sqrt{(N)}$, where N is the number of measurements for that reflection. This gives an estimate of the standard deviation of the mean value, which would hypothetically be found if the experiment were to be repeated many times under the same conditions and many mean values calculated. A problem with this method is that the standard deviation can only be calculated accurately when the number of measurements for the reflection is large. Unfortunately, this is precisely the case in which the merged intensity itself would be the most precise, and therefore when we *least* need a good estimate of the error!

The ease with which these metrics can be calculated means that they are very commonly used. However, it is much better to use figures of merit which measure the accuracy as well as the precision of the data. For these, the fit between the final model and the data (R_{work} and R_{free}) and the anomalous measurability [Dauter(2006)] are useful.

7.3.6 Solving the structure

Once the final set of reflection intensities has been produced by merging the individual measurements from all the diffraction patterns, solving the structure can follow the same procedure as for conventional rotation crystallography. As for conventional macromolecular crystallography, molecular replacement is the most commonly used structure solution method. This method relies on a ‘search model’ which is similar in structure to the macromolecule under investigation. Because the search model provides a large amount of information itself — needing only small modifications to arrive at the final structure — this method has relatively low requirements on the data quality.

Experimental phasing techniques such as Single Anomalous Diffraction (SAD) have now been applied successfully to SFX data. These methods have much more stringent requirements on the data quality, because they are based on differences between intensities that are small fractions of the absolute intensities, and therefore can provide a useful independent validation of the data quality in SFX. In SAD, these are the differences between intensities of reflections with inverse indices (hkl and $\bar{h}\bar{k}\bar{l}$). These differences can be made larger by incorporating atoms of a strongly scattering species into the structure, and recording data using an x-ray wavelength close to the resonance condition for that species. Improvements in SFX data acquisition and processing can be traced by following the progression of SAD phasing results, starting with the observation of an anomalous signal from sulphur atoms [Barends et al(2013)] progressing to phasing using a very strong anomalous signal from gadolinium atoms [Barends et al(2014)], phasing of

a well-ordered protein using the weak anomalous signal from sulphur and chlorine atoms natively present in the protein [Nakane et al(2015)] and then arriving at the native phasing of a membrane protein [Batyuk et al(2016)]. Along the way, it has been found that Single Isomorphous Replacement with Anomalous Scattering (SIRAS), in which two sets of data are compared, with and without the additional heavy atom, gives a large reduction in the number of patterns required for successful phasing [Yamashita et al(2015)]. It has also been found that the number of patterns required can be significantly reduced by careful calibration of the detector geometry [Nass et al(2016)].

The most commonly used strongly scattering species for SAD phasing in conventional crystallography is selenium. It would be very useful to establish conditions for Se-SAD phasing at x-ray FEL facilities, because techniques for incorporating selenium atoms into macromolecular structures are widely understood. The resonance condition for selenium is at a wavelength of just less than 1 Å, which is currently difficult to achieve at LCLS, but routinely available at SACLA. However, Se-SAD phasing was recently demonstrated using LCLS [Hunter et al(2016)] and SACLA [Yamashita et al (2017)], and is expected to be possible at future XFEL facilities including LCLS-II, PAL-XFEL and the European XFEL.

The radiation damage mechanisms are quite different for FEL data as for synchrotron data (see **Chapter 6**). In FEL data, electronic damage processes such as ‘bleaching’ of atoms dominate over disintegration of the crystal due to processes such as free radical diffusion, which occur on a timescale much longer than the FEL pulse length. The degree of ionization of atoms is expected to depend on the x-ray pulse fluence, with more bleaching expected at high intensities. This principle may lead to a new experimental phasing method similar to Radiation Induced Phasing (RIP). The heaviest atoms should be more ionized and hence scatter less strongly at high intensity than at low intensity. A pair of diffraction datasets at low and high intensities could therefore be used similarly to the datasets in Single Isomorphous Replacement (SIR). These differences have been observed experimentally for heavy atoms [Galli et al(2015a)] and natively occurring atoms [Galli et al(2015b)].

Other aspects of FEL data, such as Fourier truncation fringes arising from the use of very small crystals and coherent x-rays, may give additional information as discussed in **Chapter 8**.

7.4 Conclusion

Data processing for SFX, as well as other related techniques using x-ray FELs and synchrotrons, has become an active research field in its own right. There are several themes to this research, including addressing the technical challenges of storing and archiving data, understanding the fundamental physics underlying the diffraction and radiation damage processes, and the software engineering challenges of creating processing software that meet these challenges while at the same time being accessible to a rapidly growing number of non-expert users.

The amount of x-ray FEL beamtime available worldwide is oversubscribed several times over. The amount of data required per experiment has been reduced over the years, using the techniques described in this chapter. However, this is unlikely to translate to a reduction in the size of the data mountain. Instead, more experiments will be performed, using a smaller amount of measurement time each. This has already been seen at LCLS, which since 2013 has offered "protein crystal screening" shifts of only six hours to allow crystals to be tested in the LCLS beam before applying for time for a more ambitious experiment. Although intended only for a quick check that the crystals produce sufficient diffraction signal for the more ambitious experiment, several of these shifts have led to protein structures being determined, for example [Conrad et al (2015), Hutchison et al(2017)].

Another factor affecting the future size of the data mountain is the type of experiment. In the past years, determination of static protein structure has dominated. However, dynamic experiments with many time points are on the rise, involving as many as 14 individual data sets [Barends et al(2015)]. When looking for small intensity differences between the datasets, consistency of acquisition conditions is paramount, and so all the time points should be recorded during one block of experiment time. This can result in hundreds of thousands of diffraction patterns (corresponding to many million detector frames) being processed for a single experiment [Pande et al(2016)].

Finally, there are high hopes that the next generation of high repetition rate x-ray FELs will allow us to record an entire data set in a matter of seconds. At this speed, the experimenters themselves would become the limiting factor for the speed of the experiment, and automated systems for injecting a lineup of samples would be useful. This could produce a whole new data mountain, consisting not just of carefully acquired individual data sets, but instead of systematic parameter space investigation, for example of different crystallization or ligand binding conditions. The data mountain climbers will not be able to return to base camp for a long time yet!

Acknowledgements TAW acknowledges the Helmholtz Association via Programme-Oriented Funds. Portions of this research were carried out at the Linac Coherent Light Source (LCLS) at the SLAC National Accelerator Laboratory. LCLS is an Office of Science User Facility operated for the US Department of Energy Office of Science by Stanford University. Use of the Linac Coherent Light Source (LCLS), SLAC National Accelerator Laboratory, is supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences under Contract No. DE-AC02-76SF00515.

References

- [Hutchison et al(2017)] Hutchison CDM, Cordon-Preciado V, Morgan RML, Nakane T, Ferreira J, Dorlhiac G, Sanchez-Gonzalez A, Johnson AS, Fitzpatrick A, Fare C, Marangos JP, Yoon CH, Hunter MS, DePonte DP, Boutet S, Owada S, Tanaka R, Tono K, Iwata S, van Thor J (2017) X-ray Free Electron Laser Determination of Crystal Structures of Dark and Light

- States of a Reversibly Photoswitching Fluorescent Protein at Room Temperature. *Int. J. Mol. Sci.* 18, 1918, DOI <https://doi.org/10.3390/ijms18091918>
- [Boutet et al(2015)] Boutet S, Foucar L, Barends TRM, Botha S, Doak RB, Koglin JE, Messerschmidt M, Nass K, Schlichting I, Seibert MM, Shoeman RL, Williams GJ (2015) Characterization and use of the spent beam for serial operation of LCLS. *J Synchrotron Rad* 22, 634–643, DOI <https://doi.org/10.1107/S1600577515004002>
- [Mancuso et al(2013)] Mancuso AP, Aquila A, Borchers G, Giewekemeyer K, Reimers N (2013) Technical Design Report: Scientific Instrument Single Particles, Clusters, and Biomolecules (SPB), DOI [dx.doi.org/10.3204/XFEL.EU/TR-2013-004](https://doi.org/10.3204/XFEL.EU/TR-2013-004)
- [Zhu et al(2014)] Zhu D, Feng Y, Stoupin S, Terentyev SA, Lemke HT, Fritz DM, Chollet M, Glowonia JM, Alonso-Mori R, Sikorski M, Song S, van Driel TB, Williams GJ, Messerschmidt M, Boutet S, Blank VD, Shvy'dko YV, Robert A (2014) Performance of a beam-multiplexing diamond crystal monochromator at the Linac Coherent Light Source. *Review of Scientific Instruments* 85(6): 063106.
- [Gruner] Pixel array detectors. <http://bigbro.biophys.cornell.edu/research/pad>. Accessed 20 Nov 2017
- [Blaj et al(2015)] Blaj G, Caragiulo P, Carini G, Carron S, Dragone A, Freytag D, Haller G, Hart P, Hasi J, Herbst R, Herrmann S, Kenney C, Markovic B, Nishimura K, Osier S, Pines J, Reese B, Segal J, Tomada A, Weaver M (2015) X-ray detectors at the Linac Coherent Light Source. *J Synchrotron Rad* 22(3), 577–583, DOI [http://dx.doi.org/10.1107/S1600577515005317](https://doi.org/10.1107/S1600577515005317)
- [Yoon et al(2017)] Yoon CH, DeMirici H, Sierra RG, Dao H, Ahmadi R, Aksit F, Aquila AL, Batyuk A, Ciftci H, Guillet S, Hayes M, Hayes B, Lane TJ, Liang M, Lundström U, Koglin JE, Mgbam P, Rao Y, Rendahl T, Rodríguez E, Zhang L, Wakatsuki S, Boutet S, Holton JM, Hunter MS (2017) Se-SAD serial femtosecond crystallography datasets from selenobiotinyl-streptavidin. *Scientific Data* 4, 170055, DOI [http://dx.doi.org/10.1038/sdata.2017.55](https://doi.org/10.1038/sdata.2017.55)
- [Konnecke et al(2015)] Könnecke M, Akeroyd FA, Bernstein HJ, Brewster AS, Campbell SI, Clausen B, Cottrell S, Hoffmann JU, Jemian PR, Männicke D, Osborn R, Peterson PF, Richter T, Suzuki J, Watts B, Wintersberger E, Wuttke J (2015) The NeXus data format. *J Appl Crystallogr* 48, 301–305, DOI [http://dx.doi.org/10.1107/S1600576714027575](https://doi.org/10.1107/S1600576714027575)
- [Maia(2012)] Maia FRNC (2012) The Coherent X-ray Imaging Data Bank. *Nat Methods* 9(9), 854–855, DOI [http://dx.doi.org/10.1038/nmeth.2110](https://doi.org/10.1038/nmeth.2110)
- [Nakane et al(2016)] Nakane T, Joti Y, Tono K, Yabashi M, Nango E, Iwata S, Ishitani R, Nureki O (2016) Data processing pipeline for serial femtosecond crystallography at SACLA. *J Appl Crystallogr* 49, 1035–1041.
- [Foucar(2016)] Foucar L (2016) CFEL-ASG Software Suite (CASS): usage for free-electron laser experiments with biological focus. *J Appl Crystallogr* 49(4), 1336–1346.
- [Thayer et al(2017)] Thayer J, Damiani D, Ford C, Dubrovin M, Gaponenko I, O'Grady CP, Kroeger W, Pines J, Lane TJ, Salnikov A, Schneider D, Tookey T, Weaver M, Yoon CH, Perazzo A (2017) Data systems for the Linac coherent light source. *Adv Struct Chem Imaging* 3(1):3, DOI [http://dx.doi.org/10.1186/s40679-016-0037-7](https://doi.org/10.1186/s40679-016-0037-7)
- [Lyubimov et al(2016)] Lyubimov AY, Uervirojnangkoorn M, Zeldin OB, Brewster AS, Murray TD, Sauter NK, Berger JM, Weis WI, Brunger AT (2016) IOTA: integration optimization, triage and analysis tool for the processing of XFEL diffraction images. *J Appl Crystallogr* 49, 1057–1064.
- [Heisen et al(2013)] Heisen BC, Boukhelef D, Esenov S, Hauf S, Kozlova I, Maia L, Parenti A, Szuba J, Weger K, Wrona K, Youngman C (2013) Karabo: An integrated software framework combining control, data management, and scientific computing tasks. *Proceedings of ICALEPCS*, San Francisco, U.S.A.
- [Mariani et al(2016)] Mariani V, Morgan A, Yoon CH, Lane TJ, White TA, O'Grady CP, Kuhn M, Aplin S, Koglin J, Barty A, Chapman HN (2016) OnDA: online data analysis and feedback for serial X-ray imaging. *J Appl Crystallogr* 49(3), 1073–1080.
- [Damiani et al(2016)] Damiani D, Dubrovin M, Gaponenko I, Kroeger W, Lane TJ, Mitra A, O'Grady CP, Salnikov A, Sanchez-Gonzalez A, Schneider D, Yoon C H (2016) Linac Coherent Light Source data analysis using it psana. *J Appl Crystallogr* 49, 672–679.

- [Barty et al(2014)] Barty A, Kirian RA, Maia FRNC, Hantke M, Yoon CH, White TA, Chapman HN (2014) Cheetah: software for high-throughput reduction and analysis of serial femtosecond X-ray diffraction data. *J Appl Crystallogr* 47(3), 1118–1131.
- [Mozzanica et al(2014)] Mozzanica A, Bergamaschi A, Cartier S, Dinapoli R, Greiffenberg D, Johnson I, Jungmann J, Maliakal D, Mezza D, Ruder C, Schaedler L, Schmitt B, Shi X, Tinti G (2014) Prototype characterization of the JUNGFRUA pixel detector for SwissFEL. *Journal of Instrumentation* 9, C05010.
- [Allahgholi et al(2015)] Allahgholi A, Becker J, Bianco L, Delfs A, Dinapoli R, Goettlicher P, Graafsma H, Greiffenberg D, Hirsemann H, Jack S, Klanner R, Klyuev A, Krueger H, Lange S, Marras A, Mezza D, Mozzanica A, Rah S, Xia Q, Schmitt B, Schwandt J, Sheviakov I, Shi X, Smoljanin S, Trunk U, Zhang J, Zimmer M (2015) AGIPD, a high dynamic range fast detector for the European XFEL. *Journal of Instrumentation* 10, C01023.
- [Barty et al(2008)] Barty A, Boutet S, Bogan MJ, Hau-Riege S, Marchesini S, Sokolowski-Tinten K, Stojanovic N, Tobey R, Ehrke H, Cavalleri A, Dusterer S, Frank M, Bajt S, Woods BW, Seibert MM, Hajdu J, Treusch R, Chapman HN (2008) Ultrafast single-shot diffraction imaging of nanoscale dynamics. *Nat Phot* 2, 415–419, DOI <http://dx.doi.org/10.1038/nphoton.2008.128>
- [Bajt et al(2008)] Bajt S, Chapman HN, Spiller EA, Alameda JB, Woods BW, Frank M, Bogan MJ, Barty A, Boutet S, Marchesini S, Hau-Riege SP, Hajdu J, Shapiro D (2008) Camera for coherent diffractive imaging and holography with a soft-x-ray free-electron laser. *Appl Opt* 47:1673–83
- [Chapman et al(2006)] Chapman HN, Barty A, Bogan MJ, Boutet S, Frank M, Hau-Riege SP, Marchesini S, Woods BW, Bajt S, Benner WH, London RA, Plonjes E, Kuhlmann M, Treusch R, Dusterer S, Tschentscher T, Schneider JR, Spiller E, Moller T, Bostedt C, Hoener M, Shapiro DA, Hodgson KO, van der Spoel D, Burmeister F, Bergh M, Coleman C, Hult G, Seibert MM, Maia FRNC, Lee RW, Szoke A, Timneanu N, Hajdu J (2006) Femtosecond diffractive imaging with a soft-X-ray free-electron laser. *Nat Phys* 2, 839, DOI <http://dx.doi.org/10.1038/nphys461>
- [Hunter et al(2014)] Hunter MS, Segelke B, Messerschmidt M, Williams GJ, Zatsepin NA, Barty A, Benner WH, Carlson DB, Coleman M, Graf A, Hau-Riege SP, Pardini T, Seibert MM, Evans J, Boutet S, Frank M (2014) Fixed-target protein serial microcrystallography with an x-ray free electron laser. *Scientific Reports* 4:6026, DOI <http://dx.doi.org/10.1038/srep06026>
- [Carini et al(2014)] Carini GA, Boutet S, Chollet M, Dragone A, Haller G, Hart PA, Herrmann SC, Kenney CJ, Koglin J, Messerschmidt M, Nelson S, Pines J, Robert A, Song S, Thayer JB, Williams GJ, Zhu D (2014) Experience with the CSPAD during dedicated detector runs at LCLS. *J. Phys. Conf. Series* 493, 012011.
- [Casanas et al(2016)] Casanas A, Warshamange R, Finke AD, Panepucci E, Olieric V, Nöll A, Tampé R, Brandstetter S, Förster A, Mueller M, Schulze-Bries C, Bunk O, Wang, M (2016) EIGER detector: application in macromolecular crystallography. *Acta Crystallographica. Section D, Structural Biology*, 72(9):1036-1048, DOI <http://doi.org/10.1107/S2059798316012304>
- [Kraft et al(2009)] Kraft P, Bergamaschi A, Broennimann C, Dinapoli R, Eikenberry EF, Heinrich B, Johnson I, Mozzanica A, Schlepütz CM, Willmott PR, Schmitt B (2009) Performance of single-photon-counting PILATUS detector modules. *Journal of Synchrotron Radiation*. 16(3):368-375, DOI <http://doi.org/10.1107/S0909049509009911>
- [Barends et al(2015)] Barends TRM, Foucar L, Ardevol A, Nass K, Aquila A, Botha S, Doak RB, Falahati K, EHartmann, Hilpert M, Heinz M, Hoffmann MC, Köfinger J, Koglin JE, Kovascova G, Liang M, Milathianaki D, Lemke H, Reinstei n J, Roome CM, Shoeman RL, Williams GJ, Burghardt I, Hummer G, Boutet S, Schlichting I (2015) Direct observation of ultrafast collective motions in CO myoglobin upon ligand dissociation. *Science* 350:445–450
- [Chapman et al(2011)] Chapman HN, Fromme P, Barty A, White TA, Kirian RA, Aquila A, Hunter MS, Schulz J, DePonte DP, Weierstall U, Doak RB, Maia FRNC, A V Martin IS, Lomb L, Coppola N, Shoeman RL, Epp SW, Hartmann R, Rolles D, Rudenko A, Foucar L, Kimmel N, Weidenspointner G, Holl P, Liang M, Barthelmess M, Coleman C, Boutet S, Bogan MJ, Krzywinski J, Bostedt C, Bajt S, Gumprecht L, Rudek B, Erk B, Schmidt C, Homke

- A, Reich C, Pietschner D, Strüder L, Hauser G, Gorke H, Ullrich J, Herrmann S, Schaller G, Schopper F, Soltau H, Kühnel K, Messerschmidt M, Bozek JD, Hau-Riege SP, Frank M, Hampton CY, Sierra RG, Starodub D, Williams GJ, Hajdu J, Timneanu N, Seibert MM, Andreasson J, Rocker A, Jonsson O, Svenda M, Stern S, Nass K, Andritschke R, Schröter C, Krasniqi F, Bott M, Schmidt KE, Wang X, Grotjohann I, Holton JM, Barends TRM, Neutze R, Marchesini S, Fromme R, Schorb S, Rupp D, Adolph M, Gorkhover T, Andersson I, Hirse-mann H, Potdevin G, Graafsma H, Nilsson B, Spence JCH (2011) Femtosecond x-ray protein nanocrystallography. *Nature* 470:73–77
- [Dauter(2006)] Dauter Z (2006) Estimation of anomalous signal in diffraction data. *Acta Cryst D* 62:867–876
- [Duisenberg(1992)] Duisenberg AJM (1992) Indexing in Single-Crystal Diffractometry with an Obstinate List of Reflections. *J Appl Cryst* 25:92–96
- [Gildea et al(2014)] Gildea RJ, Waterman DG, Parkhurst JM, Axford D, Sutton G, Stuart DI, Sauter NK, Evans G, Winter G (2014) New methods for indexing multi-lattice diffraction data. *Acta Cryst D* 70:2652–2666
- [Ginn et al(2015a)] Ginn HM, Brewster AS, Hattne J, Evans G, Wagner A, Grimes JM, Sauter NK, Sutton G, Stuart DI (2015a) A revised partiality model and post-refinement algorithm for X-ray free-electron laser data. *Acta Cryst D* 71:1400–1410
- [Ginn et al(2015b)] Ginn HM, Messerschmidt M, Ji X, Zhang H, Axford D, Gildea RJ, Winter G, Brewster AS, Hattne J, Wagner A, Grimes JM, Evans G, Sauter NK, Sutton G, Stuart DI (2015b) Structure of CPV17 polyhedrin determined by the improved analysis of serial femtosecond crystallographic data. *Nature Communications* 6:6435
- [Ginn et al(2016)] Ginn HM, Roedig P, Kuo A, Evans G, Sauter NK, Ernst O, Meents A, Mueller-Werkmeister H, Miller RJD, Stuart DI (2016) TakeTwo: an indexing algorithm suited to still images with known crystal parameters. *Acta Cryst D* 72:956–965
- [Hattne et al(2014)] Hattne J, Echols N, Tran R, Kern J, Gildea RJ, Brewster AS, Alonso-Mori R, Glöckner C, Hellmich J, Laksmono H, Sierra RG, Lassalle-Kaiser B, Lampe A, Han G, Gul S, DiFiore D, Milathianaki D, Fry AR, Miahnahri A, White WE, Schafer DW, Seibert MM, Koglin JE, Sokaras D, Weng TC, Sellberg J, Latimer MJ, Gatzel P, Zwart PH, Grosse-Kunstleve RW, Bogan MJ, Messerschmidt M, Williams GJ, Boutet S, Messinger J, Zouni A, Yano J, Bergmann U, Yachandra VK, Adams PD, Sauter NK (2014) Accurate macromolecular structures using minimal measurements from X-ray free-electron lasers. *Nature Methods* 11:545–548
- [Kabsch(1988)] Kabsch W (1988) Evaluation of single-crystal x-ray diffraction data from a position-sensitive detector. *J Appl Cryst* 21:916–924
- [Kabsch(2014)] Kabsch W (2014) Processing of X-ray snapshots from crystals in random orientations. *Acta Cryst D* 70:2204–2216
- [Karplus and Diederichs(2012)] Karplus PA, Diederichs K (2012) Linking Crystallographic Model and Data Quality. *Science* 336:1030–1033
- [Kirian et al(2010)] Kirian RA, Wang X, Weierstall U, Schmidt KE, Spence JCH, Hunter M, Fromme P, White T, Chapman HN, Holton J (2010) Femtosecond x-ray protein nanocrystallography — data analysis methods. *Optics Express* 18:5713–5723
- [Kroon-Batenburg et al(2015)] Kroon-Batenburg LMJ, Schreurs AMM, Ravelli RBG, Gros P (2015) Accounting for partiality in serial crystallography using ray-tracing principles. *Acta Cryst D* 71:1799–1811
- [Lyubimov et al(2016)] Lyubimov AY, Uervirojnangkoorn M, Zeldin OB, Brewster AS, Murray TD, Sauter NK, Berger JM, Weis WI, Brunger AT (2016) IOTA: integration optimization, triage and analysis tool for the processing of XFEL diffraction images. *J Appl Cryst* 49:1057–1064
- [Powell(1999)] Powell HR (1999) The Rossmann Fourier autoindexing algorithm in *MOSFLM*. *Acta Crystallographica Section D* 55(10):1690–1695, DOI 10.1107/S0907444999009506
- [Rossmann(1979)] Rossmann MG (1979) Processing Oscillation Diffraction Data for Very Large Unit Cells with an Automatic Convolution Technique and Profile Fitting. *J Appl Cryst* 12:225–238

- [Rossmann et al(1979)] Rossmann MG, Leslie AGW, Abdel-Meguid SS, Tsukihara T (1979) Processing and post-refinement of oscillation camera data. *J Appl Cryst* 12:570–581
- [Sauter(2015)] Sauter NK (2015) XFEL diffraction: developing processing methods to optimize data quality. *J Synchrotron Rad* 22:239–248
- [Schreurs et al(2010)] Schreurs AMM, Xian X, Kroon-Batenburg LMJ (2010) EVAL15: a diffraction data integration method based on ab initio predicted profiles. *J Appl Cryst* 43:70–82
- [Uervirojnangkoorn et al(2015)] Uervirojnangkoorn M, Zeldin OB, Lyubimov AY, Hattne J, Brewster AS, Sauter NK, Brunger AT, Weis WI (2015) Enabling X-ray free electron laser crystallography for challenging biological systems from a limited number of crystals. *eLife* 4:e05421
- [White(2014)] White TA (2014) Post-refinement method for snapshot serial crystallography. *Phil Trans R Soc B* 369:20130,330
- [White et al(2013)] White TA, Barty A, Stellato F, Holton JM, Kirian RA, Zatsepin NA, Chapman HN (2013) Crystallographic data processing for free-electron laser sources. *Acta Crystallographica D* 69:1231–1240
- [White et al(2016)] White TA, Mariani V, Brehm W, Yefanov O, Barty A, Beyerlein KR, Chervinskii F, Galli L, Gati C, Nakane T, Tolstikova A, Yamashita K, Yoon CH, Diederichs K, Chapman HN (2016) Recent developments in CrystFEL. *J Appl Cryst* 49:680–689
- [Yefanov et al(2015)] Yefanov O, Mariani V, Gati C, White TA, Chapman HN, Barty A (2015) Accurate determination of segmented X-ray detector geometry. *Optics Express* 23:28,459
- [Zeldin et al(2015)] Zeldin OB, Brewster AS, Hattne J, Uervirojnangkoorn M, Lyubimov AY, Zhou Q, Zhao M, Weis WI, Sauter NK, Brunger AT (2015) Data Exploration Toolkit for serial diffraction experiments. *Acta Cryst D* 71:352–356
- [Nakane et al(2015)] Nakane T, Song C, Suzuki M, Nango E, Kobayashi J, Masuda T, Inoue S, Mizohata E, Nakatsu T, Tanaka T, Tanaka R, Shimamura T, Tono K, Joti Y, Kameshima T, Hatsui T, Yabashi M, Nureki O, Iwata S, Sugahara M (2015) Native sulfur/chlorine SAD phasing for serial femtosecond crystallography. *Acta Cryst D* 71:2519–2525
- [Nass et al(2016)] Nass K, Meinhardt A, Barends TRM, Fourcar L, Gorel A, Aquila A, Botha S, Doak RB, Koglin J, Liang M, Shoeman RL, Williams G, Boutet S, Schlichting I (2016) Protein structure determination by single-wavelength anomalous diffraction phasing of X-ray free-electron laser data. *IUCrJ* 3:180–191
- [Hunter et al(2016)] Hunter MS, Yoon CH, DeMirci H, Sierra RG, Dao EH, Ahmadi R, Aksit F, Aquila AL, Ciftci H, Guillet S, Hayes MJ, Lane TJ, Liang M, Lundström U, Koglin JE, Mgbam P, Rao Y, Zhang L, Wakatsuki S, Holton JM, Boutet S (2016) Selenium single-wavelength anomalous diffraction de novo phasing using an X-ray-free electron laser. *Nature Communications* 7:13,388
- [Barends et al(2013)] Barends TRM, Foucar L, Shoeman RL, Bari S, Epp SW, Hartmann R, Hauser G, Huth M, Kieser C, Lomb L, Motomura K, Nagaya K, Schmidt C, Strecker R, Anielski D, Boll R, Erk B, Fukuzawa H, Hartmann E, Hatsui T, Holl P, Inubushi Y, Ishikawa T, Kassemeyer S, Kaiser C, Koeck F, Kunishima N, Kurka M, Rolles D, Rudek B, Rudenko A, Sato T, Schroeter CD, Soltan H, Strueder L, Tanaka T, Togashi T, Tono K, Ullrich J, Yase S, Wada Si, Yao M, Yabashi M, Ueda K, Schlichting I (2013) Anomalous signal from S atoms in protein crystallographic data from an X-ray free-electron laser. *Acta Crystallographica D* 69:838–842
- [Barends et al(2014)] Barends TRM, Foucar L, Botha S, Doak RB, Shoeman RL, Nass K, Koglin JE, Williams GJ, Boutet S, Messerschmidt M, Schlichting I (2014) De novo protein crystal structure determination from X-ray free-electron laser data. *Nature* 505:244–247
- [Batyuk et al(2016)] Batyuk A, Galli L, Ishchenko A, Han GW, Gati C, Popov PA, Lee MY, Stauch B, White TA, Barty A, Aquila A, Hunter MS, Liang M, Boutet S, Pu M, Liu ZJ, Nelson G, James D, Li C, Zhao Y, Spence JCH, Liu W, Fromme P, Katritch V, Weierstall U, Stevens RC, Cherezov V (2016) Native phasing of x-ray free-electron laser data for a G protein-coupled receptor. *Science Advances* 2:e1600,292
- [Yamashita et al(2015)] Yamashita K, Pan D, Okuda T, Sugahara M, Kodan A, Yamaguchi T, Murai T, Gomi K, Kajiyama N, Mizohata E, Suzuki M, Nango E, Tono K, Joti Y, Kameshima T,

- Park J, Song C, Hatsui T, Yabashi M, Iwata S, Kato H, Ago H, Yamamoto M, Nakatsu T (2015) An isomorphous replacement method for efficient de novo phasing for serial femtosecond crystallography. *Scientific Reports* 5:14,017
- [Galli et al(2015a)] Galli L, Son SK, Barends TRM, White TA, Barty A, Botha S, Boutet S, Caleman C, Doak RB, Nanao MH, Nass K, Shoeman RL, Timneanu N, Santra R, Schlichting I, Chapman HN (2015a) Towards phasing using high X-ray intensity. *IUCrJ* 2:627–634
- [Galli et al(2015b)] Galli L, Son SK, Klinge M, Bajt S, Barty A, Bean R, Betzel C, Beyerlein KR, Caleman C, Doak RB, Duszenko M, Fleckenstein H, Gati C, Hunt B, Kirian RA, Liang M, Nanao MH, Nass K, Oberthür D, Redecke L, Shoeman R, Stellato F, Yoon CH, White TA, Yefanov O, Spence J, Chapman HN (2015b) Electronic damage in S atoms in a native protein crystal induced by an intense X-ray free-electron laser pulse. *Structural Dynamics* 2:041,703
- [Pande et al(2016)] Pande K, Hutchison CDM, Groenhof G, Aquila A, Robinson JS, Tenboer J, Basu S, Boutet S, DePonte DP, Liang M, White TA, Zatsepin NA, Yefanov O, Morozov D, Oberthuer D, Gati C, Subramanian G, James D, Zhao Y, Koralek J, Brayshaw J, Kupitz C, Conrad C, Roy-Chowdhury S, Coe JD, Metz M, Xavier PL, Grant TD, Koglin JE, Ketawala G, Fromme R, Srajer V, Henning R, Spence JCH, Ourmazd A, Schwander P, Weierstall U, Frank M, Fromme P, Barty A, Chapman HN, Moffat K, van Thor JJ, Schmidt M (2016) Femtosecond structural dynamics drives the trans/cis isomerization in photoactive yellow protein. *Science* 352:725–729
- [Schmidt et al(2017)] Beyerlein K, White TA, Yefanov O, Gati C, Kazantsev IG, Fog-Gade N, Larsen PM, Chapman HN, Schmidt S (2017) FELIX: an algorithm for indexing multiple crystallites in X-ray free-electron laser snapshot diffraction images. *J. Applied Crystallography* 50:1075-1083
- [Brehm and Diederichs (2014)] Brehm W and Diederichs K (2014) Breaking the indexing ambiguity in serial crystallography. *Acta Cryst. D*70:101-109
- [Sauter et al (2014)] Sauter NK, Hattne J, Brewster AS, Echols N, Zwart PH, Adams PD (2014) Improved crystal orientation and physical properties from single-shot XFEL stills. *Acta Cryst. D*70:3299-3309
- [Ginn et al (2016a)] Ginn HM, Evans G, Sauter NK, Stuart DI (2016) On the release of cpxfel for processing X-ray free-electron laser images. *J. Applied Cryst.* 49:1065-1072
- [Yamashita et al (2017)] Yamashita K, Kuwabara N, Nakane T, Murai T, Mizohata E, Sugahara M, Pan D, Masuda T, Suzuki M, Sato T, Kodan A, Yamaguchi T, Nango E, Tanaka T, Tono K, Joti Y, Kameshima T, Hatsui T, Yabashi M, Manya H, Endo T, Kato R, Senda T, Kato H, Iwata S, Ago H, Yamamoto M, Yumoto F, Nakatsu T (2017) Experimental phase determination with selenome-thionine or mercury-derivatization in serial femtosecond crystallography. *IUCrJ* 4:639-647
- [Conrad et al (2015)] Conrad CE, Basu S, James D, Wang D, Schaffer A, Roy-Chowdhury S, Zatsepin NA, Aquila A, Coe J, Gati C, Hunter MS, Koglin JE, Kupitz C, Nelson G, Subramanian G, White TA, Zhao Y, Zook J, Boutet S, Cherezov V, Spence JCH, Fromme R, Weierstall U, Fromme P (2015) A novel inert crystal delivery medium for serial femtosecond crystallography. *IUCrJ* 2:421-430