JOURNAL OF SYNCHROTRON RADIATION

ISSN 1600-5775

Received 10 December 2014
Accepted 13 March 2015

Edited by M. Yabashi, RIKEN SPring-8 Center, Japan

Keywords: free-electron laser; serial femtosecond crystallography; data processing; protein crystallography.

Supporting information: this article has supporting information at journals.iucr.org/s

Effects of self-seeding and crystal post-selection on the quality of Monte Carlo-integrated SFX data

Thomas Barends,[a]* Thomas A. White,[b] Anton Barty,[b] Lutz Foucar,[a] Marc Messerschmidt,[c] Roberto Alonso-Mori,[c] Sabine Botha,[a] Henry Chapman,[b,d] R. Bruce Doak,[a] Lorenzo Galli,[b,d] Cornelius Gati,[b] Matthias Gutmann,[e] Jason Koglin,[c] Anders Markvardsen,[e] Karol Nass,[a] Dominik Oberthur,[b] Robert L. Shoeman,[a] Ilme Schlichting[a] and Sébastien Boutet[c]

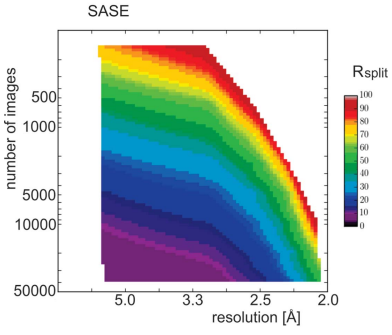
[a]Department of Biomolecular Mechanisms, Max Planck Institute for Medical Research, Jahnstrasse 29, D-69120 Heidelberg, Germany, [b]Center for Free-Electron Laser Science, Deutsches Elektronen-Synchrotron DESY, Notkestrasse 85, 22607 Hamburg, Germany, [c]SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, CA 94025, USA, [d]Physics Department, University of Hamburg, Jungiusstrasse 6, 20355 Hamburg, Germany, and [e]Rutherford Appleton Laboratory, Harwell, Oxford OX11 0QX, England. *Correspondence e-mail: thomas.barends@mpimf-heidelberg.mpg.de

Serial femtosecond crystallography (SFX) is an emerging method for data collection at free-electron lasers (FELs) in which single diffraction snapshots are taken from a large number of crystals. The partial intensities collected in this way are then combined in a scheme called Monte Carlo integration, which provides the full diffraction intensities. However, apart from having to perform this merging, the Monte Carlo integration must also average out all variations in crystal quality, crystal size, X-ray beam properties and other factors, necessitating data collection from thousands of crystals. Because the pulses provided by FELs running in the typical self-amplified spontaneous emission (SASE) mode of operation have very irregular, spiky spectra that vary strongly from pulse to pulse, it has been suggested that this is an important source of variation contributing to inaccuracies in the intensities, and that, by using monochromatic pulses produced through a process called self-seeding, fewer images might be needed for Monte Carlo integration to converge, resulting in more accurate data. This paper reports the results of two experiments performed at the Linac Coherent Light Source in which data collected in both SASE and self-seeded mode were compared. Importantly, no improvement attributable to the use of self-seeding was detected. In addition, other possible sources of variation that affect SFX data quality were investigated, such as crystal-to-crystal variations reflected in the unit-cell parameters; however, these factors were found to have no influence on data quality either. Possibly, there is another source of variation as yet undetected that affects SFX data quality much more than any of the factors investigated here.

1. Introduction

1.1. Serial femtosecond crystallography (SFX)

The extreme peak brightness and ultrashort pulses provided by X-ray free-electron lasers (FELs) allow data collection from micrometre-sized protein crystals (Chapman *et al.*, 2011; Boutet *et al.*, 2012; Redecke *et al.*, 2013) while outrunning radiation damage (Lomb *et al.*, 2011; Barty *et al.*, 2012). Using such highly intense pulses results in the near-immediate destruction of the sample, necessitating the use of a new crystal for each exposure in a scheme called serial femtosecond crystallography (SFX). In this approach, each diffraction snapshot effectively constitutes a separate diffraction experiment, and many parameters vary widely from shot to shot. The variable parameters include properties of the X-ray pulses, pulse energy (intensity), spectral distribution of



the FEL pulse (wavelength and even the detailed shape of the spectrum), as well as of the crystals (diffraction strength, resolution limit, orientation, unit-cell parameters, mosaicity and so on). Moreover, since each snapshot corresponds to a still image, due to the femtosecond duration of the pulses, all observations of reflections are 'partials', with a partiality that varies from observation to observation. In addition, these shot-to-shot variations could possibly even include effects of the experimental detectors currently in use at LCLS (Linac Coherent Light Source).

Ultimately, these variations should be compensated for at the data analysis stage, but progress in achieving this reliably and with general applicability has been slow so far. However, by making very large numbers of measurements of the reflection intensities, all variables affecting the intensities can be averaged out and the partially measured reflections combined into fully integrated intensities in a scheme known as Monte Carlo integration (Kirian *et al.*, 2010, 2011; White *et al.*, 2012).

In Monte Carlo integration, each individual diffraction snapshot is indexed and each Bragg spot integrated. Large numbers of partial observations of a particular reflection h, k, l are then averaged to obtain the final, fully integrated intensity $I(hkl)$ (Kirian *et al.*, 2010, 2011; White *et al.*, 2012). Including more and more measurements leads to higher quality of the final, integrated intensities. Indeed, when plotted *versus* the number of images included in the integration process, quality measures such as R factors are typically seen to converge to a minimum value (Kirian *et al.*, 2011; Boutet *et al.*, 2012) with increasing numbers of images.

When Monte Carlo integration is performed with several thousands of diffraction images, the quality of the structure factors obtained is usually high enough for structure solution by molecular replacement or the detection of ligands in difference density maps. However, applications that require higher quality, such as the detection of anomalous signals or experimental phasing, have to date required the collection of tens of thousands of images (Barends *et al.*, 2013, 2014; Kern *et al.*, 2014). As this requires large amounts of sample and FEL beam time, there is great interest in reducing the number of images that is needed for Monte Carlo integration to converge to obtain high-quality intensities.

Given the nature of the SFX experiment and the mechanics of Monte Carlo integration as described above, one could postulate that reducing the fluctuation in any one experimental variable should speed up Monte Carlo convergence because there are fewer variable parameters over which Monte Carlo integration must be performed.

In particular, it has been suggested that reducing the shot-to-shot spectral variations of the FEL beam could be beneficial to Monte Carlo convergence. Indeed, it may be argued that the spectral properties of FEL pulses constitute a striking difference between 'conventional' crystallography at synchrotron or home sources and crystallography at FEL sources. 'Conventional' sources afford highly stable beams with well defined spectra that can be precisely tailored to the experiment. For instance, synchrotron sources can provide highly monochromatic beams of exactly known wavelength for, for example, MAD (multi-wavelength anomalous diffraction) experiments, as well as polychromatic beams for Laue crystallography. Moreover, for any conventional crystallographic experiment, the precise wavelength and the spectrum (monochromatic or not) of the radiation is always accurately known and this information is used by the data processing software.

For most of the recent crystallographic experiments at FEL sources, the FEL was operated in self-amplified spontaneous emission (SASE) mode. This operation mode results in pulses with broad, spiky spectra that vary in shape, width and intensity from shot to shot (Fig. 1*a*). As the spectral properties of the beam are such important factors in a crystallographic

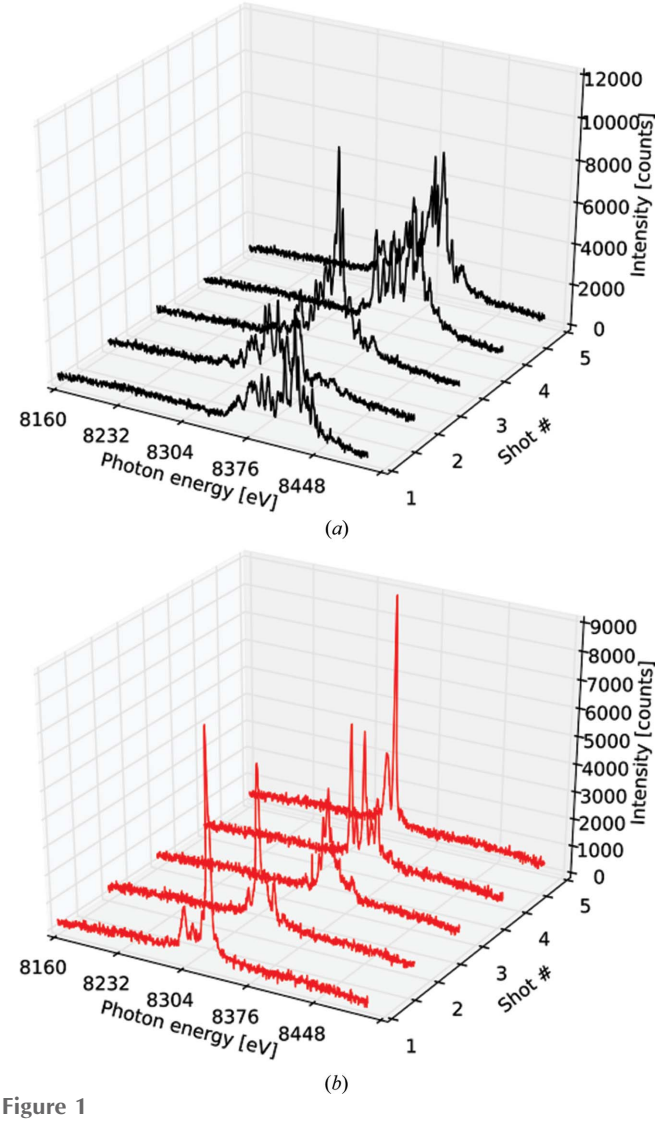


Figure 1
Typical spectra of (*a*) five SASE pulses and (*b*) five self-seeded pulses. The position and shape of the highly spiky SASE spectra vary widely on a shot-by-shot basis. The self-seeded pulse spectra are more monochromatic, depending on the success of the self-seeding process. Pulses 1, 2 and 5 were successfully self-seeded, whereas for pulses 3 and 4 the self-seeding process failed to produce a spectrum with a single peak. However, even these two pulses are more monochromatic than any of the SASE pulses in panel (*a*).

experiment, it has been suggested that these shot-to-shot variations contribute to the need to integrate over a large number of measurements in Monte Carlo integration of SFX data.

However, by using another FEL operation mode, called self-seeding, far more monochromatic pulses can be produced (Feldhaus *et al.*, 1997; Saldin *et al.*, 2001) (Fig. 1*b*). As this would effectively remove one of the varying parameters from the Monte Carlo integration process, it may result in faster Monte Carlo convergence, *i.e.* in higher SFX data quality for the same number of measurements, or, put in another way, it may be that by using the self-seeding mode the same data quality can be achieved using fewer measurements. This study describes two experiments aimed at investigating whether the use of the monochromatic pulses produced by this self-seeding mode indeed results in better data for the same number of images.

1.2. FEL pulse spectra: SASE *versus* self-seeding

SASE is the process by which X-ray pulses are created in the long undulator of a free-electron laser. In the undulator, the electric field component of the radiation interacts with the charge of the electrons. Small stochastic fluctuations in the electron bunch cause some electrons to be slightly faster and some slower. By interacting with the radiation pulse, the faster electrons lose some kinetic energy, whereas the slower ones gain some. Over many undulator periods this results in a periodic modulation of the spatial distribution of electrons in the bunch called microbunching. The microbunching has the same periodicity as the radiation wavelength, and favourable conditions cause the electrons' kinetic energy to be increasingly converted into a pulse of coherent electromagnetic radiation (Huang & Kim, 2007).

The microbunching process in SASE starts from random fluctuations in the electron pulse, and these fluctuations become 'imprinted' on the X-ray pulses. This results in pulses with very spiky spectra of relatively large bandwidth that vary greatly from pulse to pulse in both shape and wavelength distribution (see Fig. 1*a*). Most SFX experiments carried out at FELs to date have been performed with the FEL running in SASE mode.

While placing a monochromator downstream of the FEL undulator results in monochromatic pulses, this comes at the expense of large intensity variations, as a monochromator just takes a small slice out of the varying spectra. An alternative method to reduce the variability of the pulse spectra is known as hard X-ray self-seeding (Feldhaus *et al.*, 1997; Saldin *et al.*, 2001; Amann *et al.*, 2012; Yabashi & Tanaka, 2012). In self-seeded FEL pulse generation, an FEL pulse is generated in the first section of the undulator by the SASE process. A monochromatic part of the initial X-ray pulse is then separated in time from the rest of the pulse using a crystal monochromator downstream of the first undulator section, while the electron bunch is diverted around the monochromator. The electron bunch and the monochromatic part of the X-ray pulse are then recombined and enter the rest of the undulator, thus 'seeding' the SASE process with a spectrally pure X-ray pulse. While this is still a stochastic process, a considerable fraction of FEL pulses produced in this way will display a spectrum which mainly contains a single, narrow spike with a wavelength determined by the monochromator (Fig. 1*b*), although the power of the self-seeded pulses can vary greatly.

As mentioned above, it may be expected that, due to their more consistent wavelength and spectrum, the use of self-seeded pulses might result in SFX data of superior quality to that collected using SASE pulses for the same number of images, or, put differently, that fewer measurements would be required to attain the same data quality. To test this hypothesis, we collected large amounts of SFX data using both SASE and self-seeded pulses from microcrystals of the model protein lysozyme, both native and in complex with gadolinium. We then assessed data quality by several measures.

Using the native lysozyme data, we investigated the effects of seeding on data precision as measured by $R_{\rm split}$ (White *et al.*, 2012), which is a multiplicity-corrected R factor between random half data sets that can be used to track the convergence of the Monte Carlo integration process. We also analysed the effect of seeding on the signal-to-noise ratio of the native data.

Apart from looking at data precision, we used the data from the gadolinium-derivatized lysozyme crystals to investigate the effect of seeding on the strength of the anomalous signal. While there is no reason to expect that anomalous scattering *per se* would be affected by using seeded pulses, we used the strength of the observed anomalous signal to probe possible effects on data quality that cannot be measured by precision indicators such as, for example, $R_{\rm split}$. Moreover, the large volume of data allowed for a systematic investigation of various potential sources of error in SFX data collection.

2. Materials and methods

2.1. Crystallization and data collection

Protein crystals were grown, derivatized and injected as described previously (Boutet *et al.*, 2012; Barends *et al.*, 2014). Briefly, microcrystals (size $\leq 1 \times \leq 1 \times \leq 2$ µm) of hen egg-white lysozyme (Sigma) were grown using batch crystallization (Boutet *et al.*, 2012) and left to settle. To produce gadolinium-derivatized crystals, the supernatant was then exchanged for 8% NaCl, 0.1 *M* sodium acetate buffer pH 4.0 containing 100 m*M* gadoteridol [Gd^{3+}:10-(2-hydroxypropyl)-1,4,7,10-tetraazacyclododecane-1,4,7-triacetic acid (Girard *et al.*, 2002)], after which the crystals were left to incubate at room temperature for at least 30 min before use. SFX diffraction data were collected using X-ray pulses of ∼40 fs duration (electron bunch length) and 8.4 keV photon energy of 0.1 mJ average power for the native crystals, and 8.3 keV photon energy of 0.1 mJ average power for the gadolinium-derivatized crystals, essentially as described previously (Boutet *et al.*, 2012; Barends *et al.* 2014).

A suspension containing ∼30%(*v*/*v*) of lysozyme crystals in their soaking solution was injected into the 100 nm focus

chamber of the CXI instrument (Boutet & Williams, 2010) at LCLS using a 3–4 μm-diameter liquid jet from a gas-dynamic virtual nozzle (Weierstall *et al.*, 2012) running at ∼30 μl min^{-1}. Single-shot diffraction patterns were collected using a CSPAD detector (Hart *et al.*, 2012) at 120 Hz. Two data sets were collected of both the native crystals (LCLS experiment L660, June 2013) and the gadolinium complex (LCLS experiment LA06, November 2013): one using self-seeding and the other using standard SASE-mode operation with the beam attenuated to deliver the same number of photons to the sample per X-ray pulse as in self-seeding mode.

For the determination of FEL pulse spectra, a pair of single-shot X-ray spectrometers was used, both based on a bent, thin, silicon crystal concept developed at LCLS (Zhu *et al.*, 2012). The first spectrometer was installed in the LCLS front-end enclosure (FEE) over 300 m upstream of CXI. This spectrometer was, at the time, a fixed-energy device leading to the choice to operate at 8.3 keV. The FEE spectrometer could not be recorded with the data on a shot-by-shot basis; however, it proved very useful for machine tuning to deliver a good, self-seeded beam and for characterizing the performance of a second spectrometer installed roughly 10 m downstream of the sample at CXI. The position of the second spectrometer was chosen for reasons of space constraint. Because of the high divergence of the nanofocus beam, it was necessary to insert beryllium lenses downstream of the CSPAD detector to project the beam onto the spectrometer. This spectrometer was recorded for every LCLS pulse allowing a post-sorting of the data based on the properties of the spectrum of each shot.

2.2. Data analysis – native crystals

All diffraction images were stored in XTC format together with their respective X-ray spectrum. Synchronization between the diffraction images and the spectrometer was checked by correlating the variations in the signal strengths of the spectrometer and the pulse intensity monitors installed in the LCLS FEE.

For the data from native crystals, individual data frames were screened for crystal diffraction using *Cheetah* (Barty *et al.*, 2014), which also extracted the photon spectrum for each pulse. Not all FEL pulses were successfully self-seeded; for example, for some pulses the spectral distribution of the initial SASE pulse was sufficiently far from the monochromator wavelength that self-seeding failed. In order to exclude these failed attempts at self-seeding in the subsequent crystallographic analysis (see below), we classified frames where the spectrum deviated by not more than 20 pixels on the spectrometer, the Gaussian bandwidth fell between 5 and 30 pixels, and the pulse height was over 350 detector units as successful self-seeding events. No pulse selection (other than hit finding) was performed for SASE mode.

2.3. Data analysis – gadolinium-derivative crystals

In the case of the gadolinium-derivatized crystals, hit finding and identification of successfully self-seeded pulses were performed using *CASS* (Foucar *et al.*, 2012). A first, crude differentiation between self-seeded and non-seeded pulses was obtained by analysing each recorded spectrum for two features: the overall height of the self-seeded spike and its width at a certain fraction of the height. If the peak height was above 9000 counts and the width of the peak was less than 70 pixels (25 eV) at 1/5 the peak height, the pulse was regarded as successfully self-seeded. Individual diffraction patterns were labelled as either successfully seeded or not successfully seeded and written out as individual files in the Hierarchical Data Format version 5 (HDF5) format containing both the diffraction image and the spectrum. Individual diffraction events that were labelled as successfully seeded using the crude initial analysis described above were then further classified according to the spectral purity of the self-seeded FEL pulse using a two-step scheme illustrated in Fig. 2. In step 1, the X-ray pulse spectrum of each indexed image was retrieved, and the lowest value in the entire spectrum was subtracted to reduce the background. In step 2, a window with a width of 50 pixels (corresponding to an energy bandwidth of 18 eV) centred at 8.33 keV was defined, and the area under the spectral curve (Fig. 2, red area) inside this window divided by the area under the entire spectrum (Fig. 2, red and grey areas). Four classes of indexed images were defined using this ratio: *A* with a ratio > 0.3, *B* with a ratio of 0.2–0.3, *C* with a ratio of 0.1–0.2 and *D* with a ratio of 0–0.1.

2.4. Monte Carlo integration

All SFX data sets were processed using *CrystFEL* (White *et al.*, 2012), which was also used to calculate quality metrics. The merged data sets were converted to OLDHKL format and imported into *XSCALE* (Kabsch, 2010), which was used to calculate correlation coefficients between data from individual runs. The correlation coefficients were then used to construct a similarity matrix, which was passed on to the Dendro-UPGMA server [http://genomes.urv.cat/UPGMA/ (Garcia-Vallvé *et al.*, 1999)] for clustering and representation as a 'phylogenetic' tree. *XPREP* (Bruker AXS GmbH) was used to prepare data for substructure searches. Data to 1.9 Å were used and scaling was performed with 100 reflections in the local scaling sphere, after which the F_A values were renormalized using a *B* factor of 20 $Å^2$. The substructure search itself was performed with *SHELXD* (Schneider & Sheldrick, 2002) using 500 trials for each data set.

3. Results

3.1. Native data – self-seeded and unseeded spectra

We first looked at the data collected using native lysozyme crystals, initially comparing the spectra of self-seeded and SASE pulses. In both self-seeded and SASE mode, the photon spectrum for each shot was analysed by fitting a Gaussian peak to the spectrum and measuring the peak height and width. This enabled us to obtain a quick visualization of the difference in shot-to-shot parameters in self-seeded and SASE modes. The results are summarized in Fig. 3, which shows, as

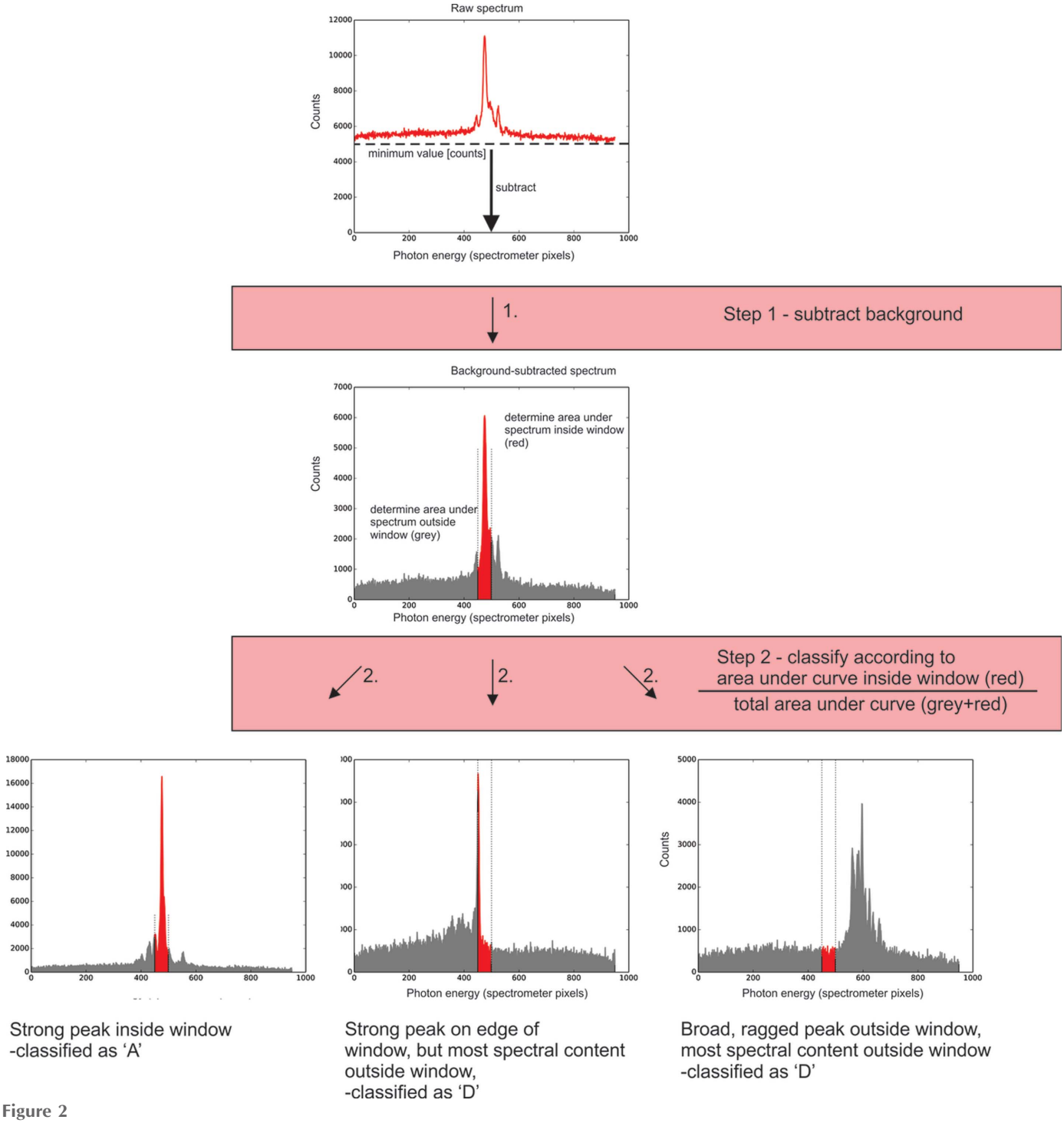


Figure 2
Definition of spectral purity for the classification of images using two steps. In step 1, the lowest value present in the entire, raw spectrum (top) was subtracted to reduce the background. In step 2, the area under the curve inside an 18 eV-wide window (red area) is divided by the area under the entire spectrum (red and grey areas). The resulting ratio was used to classify according to spectral purity using the definitions described in the main text.

expected, a much more stable central wavelength for self-seeding mode, with a much narrower and stronger peak in photon spectral density.

3.2. Native data – Monte Carlo convergence

In order to test the effect of the narrower spectral distribution of the X-ray pulses provided by self-seeding on data convergence, we calculated the $R_{\rm split}$ metric (White *et al.*, 2012) as a function of the number of indexed crystals, for both self-seeded and SASE operation modes. As can be seen from Fig. 4, contrary to expectation, there is essentially no difference between SASE and self-seeded modes in terms of the convergence rate of $R_{\rm split}$ as a function of number of patterns.

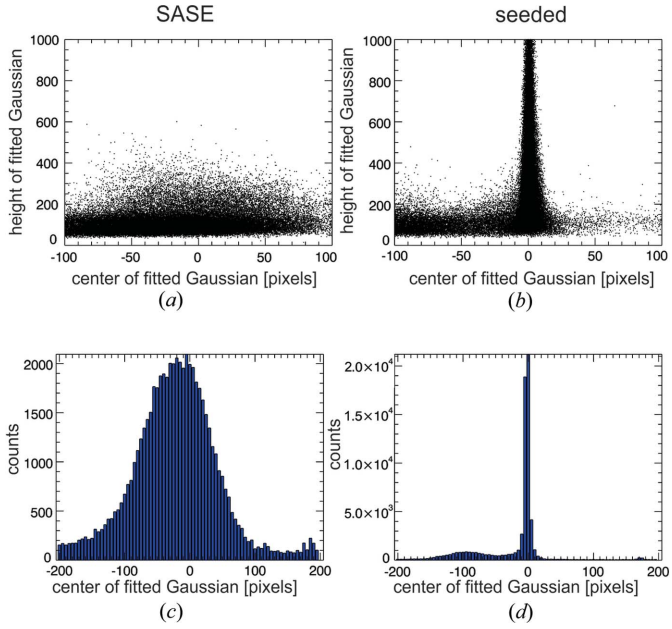


Figure 3
Scatter plot of the height and mean centre of a Gaussian fit to the spectrum in both SASE and self-seeded mode. The bottom row is a histogram of the scatter plot, showing the much sharper spectral distribution afforded by self-seeding.

3.3. Gadolinium-derivative data – data quality and anomalous signal strength

Given this unexpected lack of improvement in Monte Carlo convergence for the native lysozyme data, we proceeded to analyse the strength of the anomalous signal afforded by the gadolinium atoms in the derivative data sets, to investigate whether self-seeding has advantages for the detection of such signals. To this end, we first compared data sets prepared from SASE and successfully self-seeded images (as determined by the criterion described above) which each contained ∼133000 indexed images. Fig. 5 shows the overall quality of the data as measured by $R_{\rm split}$, the redundancy (defined here as the multiplicity of observation of a unique partial reflection) and the signal-to-noise ratio. Here, too, as is clear from the almost perfectly overlapping graphs, there are no significant differences in these metrics between the SASE and self-seeded data sets.

Fig. 6 shows the strength of the anomalous signal afforded by the gadolinium atoms in terms of $R_{\rm ano}/R_{\rm split}$, which is a measure of the anomalous signal-to-noise ratio (Barends *et al.*, 2014; Weiss, 2001) and the correlation of the anomalous signal between half data sets $CC_{\rm ano}$. Here, too, there are no differences in data quality between the SASE and self-seeded data.

We also compared self-seeded and SASE data by evaluating their usefulness in a substructure search. To this end, we prepared data sets from both self-seeded and SASE data containing 10000, 20000, 30000 and 40000 indexed images and performed a dual-space substructure search for two gadolinium atoms using *SHELXD*. No difference between self-seeded and SASE data was apparent for any of these data sets either in terms of the success rate or in terms of how easy

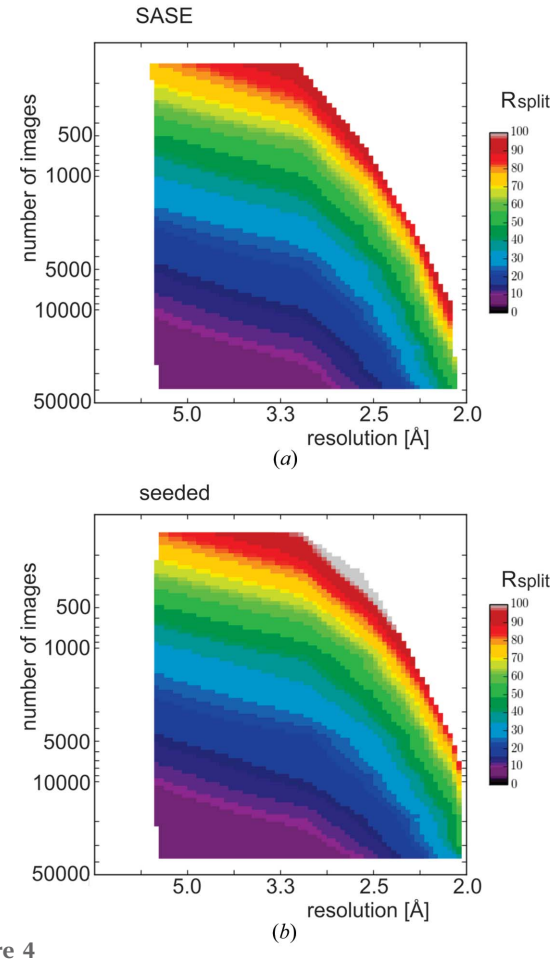


Figure 4
Convergence of the Monte Carlo integration of the native data using SASE (*a*) and self-seeded pulses (*b*). The $R_{\rm split}$ of the data (White *et al.*, 2012), defined as $R_{\rm split} = (1/\sqrt{2})(\sum|I_{\rm even} - I_{\rm odd}|)/\{(1/2)[\sum(I_{\rm even} + I_{\rm odd})]\}$ is shown as a coloured surface, with purple indicating low values and red indicating high values, as a function of resolution and the number of images included in the integration. As expected, the inclusion of more images results in a reduction in $R_{\rm split}$.

it was to distinguish between correct and incorrect solutions (see Fig. S1 in the supporting information).

However, as can be seen in Fig. 1(*b*) and Fig. 3, there is still considerable variation in the spectra of self-seeded pulses. We therefore extended our analysis by grouping the self-seeded data into classes of decreasing spectral purity (*A*, *B*, *C* and *D*, where *A* has the highest spectral purity and *D* has the lowest, see §2). Only ∼10000 images fell inside the high-quality *A* class, so that for each class a final data set for comparison was prepared containing 10000 images. In each class, the multiplicity of the data was the same, as was also seen for the comparison between SASE and self-seeded data.

Here, too, no trend in data quality could be observed. The final $R_{\rm split}$ values to 1.9 Å resolution were 19.9%, 20.7%, 19.5% and 19.9% for the *A*, *B*, *C* and *D* classes, respectively, and the $R_{\rm ano}/R_{\rm split}$ ratios (Barends *et al.*, 2014; Weiss, 2001) were 1.24, 1.24, 1.30 and 1.28. As expected, these values are worse than those obtained for the far larger, complete data set of 133000 images described above due to the necessarily lower multiplicity of the subsets.

We then compared the performance of the *A* and *D* classes in a substructure search as above for self-seeded and SASE data. Here, using 10000 images for both classes, there was a clear advantage to using the *A* data set over the *D* data set, as a much clearer distinction between correct and incorrect substructure solutions was obtained, as can be seen in Fig. S2 (green and blue data points). At first glance, this effect appeared not to be due to a lower signal-to-noise ratio of the data as can be seen from the $R_{\rm ano}/R_{\rm split}$ values noted above as well as from the overall signal-to-noise ratios, which were 3.40 and 3.46 for the *A* and *D* classes using data to 1.9 Å, respectively. However, in *CrystFEL*, the standard deviation of a reflection's intensity is calculated from the standard deviation of all the observations of that reflection, meaning that the signal-to-noise ratios reported for the final Monte Carlo intensities are not necessarily related to the signal-to-noise ratios of the actual diffraction peaks as in conventional crystallography. Indeed, when looking at the pulse intensities of the four classes as indicated by the gas detectors installed at LCLS, the average values for the *A* and *D* classes were 0.67 and 0.44, respectively, indicating that the *A*-class images were collected with pulses that were on average 1.5× more intense than those used for the *D* class. Thus, despite the similar signal-to-noise ratios, the higher performance of the *A*-class data set in substructure searches could also be attributed to this higher intensity. Moreover, when the images from the *A* and *D* classes were combined, the correct solutions became even clearer (Fig. S2, red data points), showing that, in terms of data collection, there is no advantage to splitting a data set into classes according to spectral quality in the way used here after it has been collected. We therefore started to investigate other possible sources of variation that affect SFX data quality, using the data from the gadolinium derivative for these analyses, to enable a comparison on the basis of not only $R_{\rm split}$ but also anomalous signal.

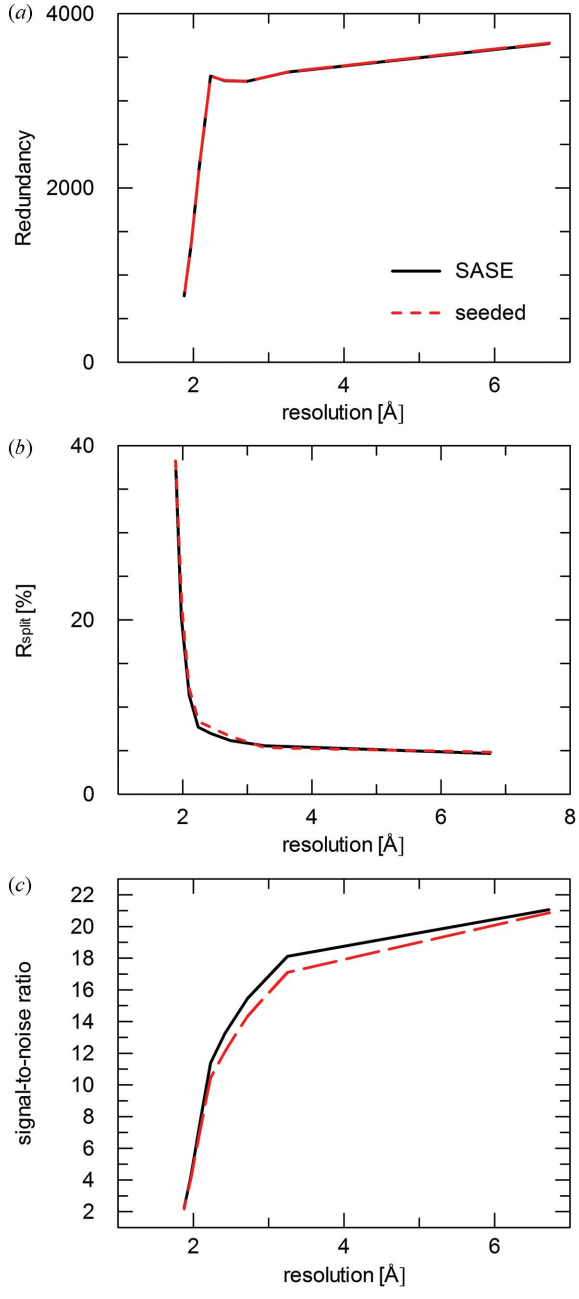


Figure 5
Quality metrics of the SASE (black line) and self-seeded (red line) data. (*a*) Redundancy of the data as a function of resolution. (*b*) $R_{\rm split}$ as a function of resolution and (*c*) signal-to-noise ratio as a function of resolution.

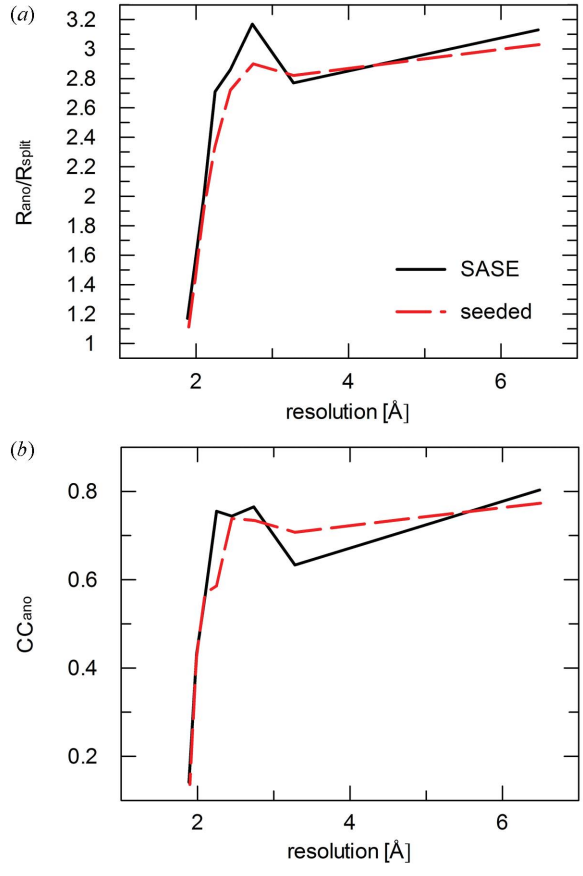


Figure 6
Anomalous signal strength in the SASE (black line) and self-seeded (red line) data sets. (*a*) $R_{\rm ano}/R_{\rm split}$ ratio. (*b*) Anomalous correlation $CC_{\rm ano}$.

3.4. Unit-cell variations

Given that no influence of spectral purity on data quality could be found, either in terms of R_{split} or anomalous signal strength, we suspected that non-isomorphism might introduce large variations in the data. To address this, we attempted to classify the indexed images according to the lengths of the unit-cell diagonals, as done in, for example, *BLEND* (Foadi *et al.*, 2013).

A scatter plot of the *ab*, *bc* and *ac* diagonal lengths did indeed at first appear to show two distinct clusters, but upon closer inspection each class exclusively contained images indexed by one of the two indexing programs used by *CrystFEL* (*DirAX* and *MOSFLM*), and thus probably reflect minor differences in their indexing algorithms.

Nonetheless, we compared data sets prepared from images from self-seeded pulses indexed exclusively by either *DirAX* or *MOSFLM*, containing 100000 self-seeded images each. Again, there was no discernible difference in data quality or anomalous signal strength between the two, the values of R_{split}, R_{ano}/R_{split} and CC_{ano} to 1.9 Å resolution being 6.1%, 2.53 and 0.71, respectively, for the *DirAX*-indexed data and 6.2%, 2.55 and 0.76 for the *MOSFLM*-indexed data. Moreover, for a data set of the same size containing both *DirAX*- and *MOSFLM*-indexed images, R_{split}, R_{ano}/R_{split} and CC_{ano} were again very similar: 6.3%, 2.55 and 0.75, respectively. Thus, while there appeared to be clusters of different unit-cell diagonals, this is due to the indexing program and has no effect on the final data quality.

3.5. Errors introduced by slow variations

In a typical SFX experiment, sets of diffraction patterns are acquired in several batches, or 'runs', each lasting several minutes. In the current experiment, runs typically lasted 10 min, after which the data collection was stopped and the next run started. This allows the comparison of data from different runs to see whether the data are affected by factors which vary on a timescale of minutes to hours.

To investigate this, we employed a clustering scheme as also used in *BLEND* (Foadi *et al.*, 2013) to identify data sets from isomorphous crystals for merging. We processed 133 runs of self-seeded data individually and calculated the pair-wise correlations between the intensities of each of them using *XSCALE* with the goal of identifying clusters of similar partial data sets. This resulted in a 133 × 133-element cross-correlation matrix, with values ranging from −0.12 to 1 (for the diagonal elements). Using the 'unweighted pair group method with arithmetic mean' algorithm as implemented in the DendroUPGMA server [http://genomes.urv.cat/UPGMA/ (Garcia-Vallvé *et al.*, 1999)] a 'phylogenetic' tree was prepared, which showed two major clusters (see Fig. 7). Strikingly, the runs in these clusters were not collected closely in time to each other.

We then prepared a data set from one of these clusters and a data set of comparable size (∼160000 indexed images) from consecutive runs, and compared their quality metrics as above. Again, there was no discernible difference in data quality by any of the metrics used: R_{split}, R_{ano}/R_{split} and CC_{ano} to 1.9 Å resolution were 5.15%, 2.9 and 0.81, respectively, for the cluster, and 5.0%, 3.0 and 0.80 for the data set prepared from consecutive runs. Thus, it appears unlikely that there are factors that change slowly over time that strongly affected SFX data quality in this experiment.

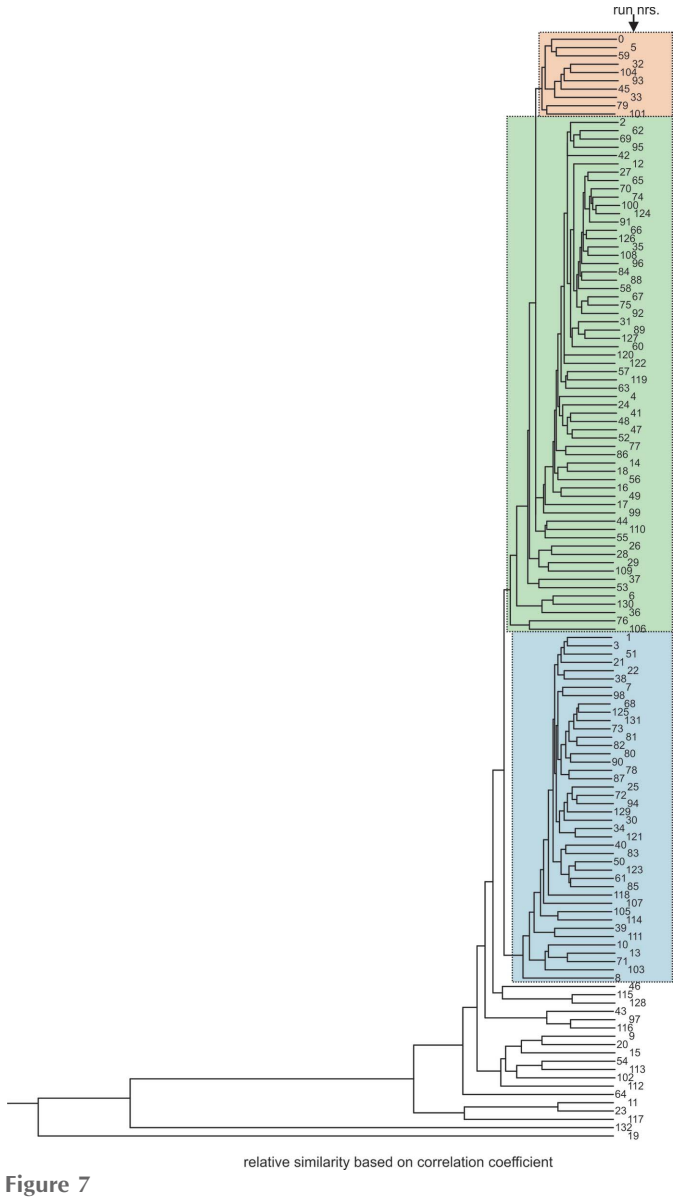


Figure 7
Result of clustering partial data sets from 133 data collection runs by pair-wise correlations on integrated intensities. Several clusters can be identified, such as those indicated by the coloured boxes. The cluster in the green box was investigated as described in the results. The run numbers are indicated on the right.

3.6. Removing 'outlier' images

Finally, we evaluated the effect of removing 'outlier' images, *i.e.* images that have a very low correlation with the final, averaged Monte Carlo data set. To this end, we calculated the Pearson correlation coefficient of the intensities derived from each individual indexed image in the total self-seeded data set, and removed those images with a correlation <0.2.

In this way, about 20000 possibly 'bad' images were removed from the total of 322000 images. However, again, this did not improve data quality. $R_{\rm split}$, $R_{\rm ano}/R_{\rm split}$ and $CC_{\rm ano}$ to 1.9 Å resolution were 3.6%, 4.2 and 0.89 before removal of poorly correlating images, and 3.6%, 4.0 and 0.89 after removal.

Also, when we performed substructure searches using data from 10000 indexed images with and without removal of poorly correlating images, there was no significant difference in either success rate or the clarity of the solution (Fig. S3).

4. Conclusions

We have compared SFX data of lysozyme microcrystals collected using self-seeded and SASE FEL pulses as well as data sets constructed from data selected according to other criteria. Importantly, and contrary to expectation, virtually no influence of the X-ray pulse spectrum on SFX data quality could be found using the Monte Carlo method for data processing. Only when comparing the usefulness in substructure searches did the most spectrally pure data appear to have a clear advantage, but this could also be caused by the higher intensity of the spectrally purer pulses.

Thus, overall the data presented here show that, when using Monte Carlo integration, self-seeding does not afford the large improvements in data quality that had been expected. We have also sorted images according to the underlying indexing program and according to the correlations between the runs in which they were collected, and we have removed poorly correlating images, but in none of these cases did we observe an improvement of data quality. One conclusion from this study is that there is likely another, larger source of error than that introduced by the different photon pulse properties and the others investigated here, one that could not be identified or isolated in the current experiment.

This study has focused on data integrated using 'pure' Monte Carlo integration without scaling, partiality estimation, post-refinement or profile fitting. However, implementations of such techniques have been described within *nXDS* (Kabsch, 2014), *CrystFEL* (White, 2014) and *cctbx.xfel* (Sauter *et al.*, 2014). We anticipate that using such improved data processing methods for SFX data may reveal differences between the seeded and unseeded cases in future. These techniques may even rely on the spectral purity offered by self-seeded pulses to make accurate partiality estimates. Furthermore, it may become possible to use the information offered by the spectra of individual FEL pulses (SASE or self-seeded) in data processing, and the broader SASE spectra may also have advantages for certain experiments because of the increased sampling of reciprocal space that they provide.

Acknowledgements

Portions of this research were carried out at the Linac Coherent Light Source, a National User Facility operated by Stanford University on behalf of the US Department of Energy, Office of Basic Energy Sciences. The CXI instrument was funded by the LCLS Ultrafast Science Instruments (LUSI) project funded by the US Department of Energy, Office of Basic Energy Sciences. We are indebted to S. Pesch and R. van Gessel (Bracco Imaging Konstanz and Singen, Germany) for the kind gift of the sample of gadoteridol. We thank Frank Koeck for excellent computing support, and the staff at the SLAC Main Control Center for the excellent self-seeded FEL beam they provided.

References

Amann, J. *et al.* (2012). *Nat. Photon.* **6**, 693–698.
Barends, T., Foucar, L., Botha, S., Doak, R. B., Shoeman, R. L., Nass, K., Koglin, J. E., Williams, G., Boutet, S., Messerschmidt, M. & Schlichting, I. (2014). *Nature (London)*, **505**, 244–247.
Barends, T. R. M. *et al.* (2013). *Acta Cryst.* D**69**, 838–842.
Barty, A. *et al.* (2012). *Nat. Photon.* **6**, 35–40.
Barty, A., Kirian, R. A., Maia, F. R. N. C., Hantke, M., Yoon, C. H., White, T. A. & Chapman, H. (2014). *J. Appl. Cryst.* **47**, 1118–1131.
Boutet, S. *et al.* (2012). *Science*, **337**, 362–364.
Boutet, S. & Williams, G. J. (2010). *New J. Phys.* **12**, 035024.
Chapman, H. N. *et al.* (2011). *Nature (London)*, **470**, 73–77.
Feldhaus, J., Saldin, E. L., Schneider, J. R., Schneidmiller, E. A. & Yurkov, M. V. (1997). *Opt. Commun.* **140**, 341–352.
Foadi, J., Aller, P., Alguel, Y., Cameron, A., Axford, D., Owen, R. L., Armour, W., Waterman, D. G., Iwata, S. & Evans, G. (2013). *Acta Cryst.* D**69**, 1617–1632.
Foucar, L., Barty, A., Coppola, N., Hartmann, R., Holl, P., Hoppe, U., Kassemeyer, S., Kimmel, N., Küpper, J., Scholz, M., Techert, S., White, T. A., Strüder, L. & Ullrich, J. (2012). *Comput. Phys. Commun.* **183**, 2207–2213.
Garcia-Vallvé, S., Palau, J. & Romeu, A. (1999). *Mol. Biol. Evol.* **16**, 1125–1134.
Girard, É., Chantalat, L., Vicat, J. & Kahn, R. (2002). *Acta Cryst.* D**58**, 1–9.
Hart, P. *et al.* (2012). *Proc. SPIE*, **8504**, 85040C.
Huang, Z. & Kim, K.-J. (2007). *Phys. Rev. ST Accel. Beams*, **10**, 034801.
Kabsch, W. (2010). *Acta Cryst.* D**66**, 125–132.
Kabsch, W. (2014). *Acta Cryst.* D**70**, 2204–2216.
Kern, J. *et al.* (2014). *Nat. Commun.* **5**, 4371.
Kirian, R. A., Wang, X., Weierstall, U., Schmidt, K. E., Spence, J. C. H., Hunter, M., Fromme, P., White, T., Chapman, H. N. & Holton, J. (2010). *Opt. Express*, **18**, 5713–5723.
Kirian, R. A., White, T. A., Holton, J. M., Chapman, H. N., Fromme, P., Barty, A., Lomb, L., Aquila, A., Maia, F. R. N. C., Martin, A. V., Fromme, R., Wang, X., Hunter, M. S., Schmidt, K. E. & Spence, J. C. H. (2011). *Acta Cryst.* A**67**, 131–140.
Lomb, L. *et al.* (2011). *Phys. Rev. B*, **84**, 214111.
Redecke, L. *et al.* (2013). *Science*, **339**, 227–230.
Saldin, E. L., Schneidmiller, E. A., Shvyd'ko, Y. V. & Yurkov, M. V. (2001). *Nucl. Instrum. Methods Phys. Res. Sect. A*, **475**, 357–362.
Sauter, N. K., Hattne, J., Brewster, A. S., Echols, N., Zwart, P. H. & Adams, P. D. (2014). *Acta Cryst.* D**70**, 3299–3309.
Schneider, T. R. & Sheldrick, G. M. (2002). *Acta Cryst.* D**58**, 1772–1779.
Weierstall, U., Spence, J. C. H. & Doak, R. B. (2012). *Rev. Sci. Instrum.* **83**, 035108.
Weiss, M. S. (2001). *J. Appl. Cryst.* **34**, 130–135.
White, T. A. (2014). *Philos. Trans. R. Soc. B*, **369**, 20130330.
White, T. A., Kirian, R. A., Martin, A. V., Aquila, A., Nass, K., Barty, A. & Chapman, H. N. (2012). *J. Appl. Cryst.* **45**, 335–341.
Yabashi, M. & Tanaka, T. (2012). *Nat. Photon.* **6**, 648–649.
Zhu, D., Cammarata, M., Feldkamp, J. M., Fritz, D. M., Hastings, J. B., Lee, S., Lemke, H. T., Robert, A., Turner, J. L. & Feng, Y. (2012). *Appl. Phys. Lett.* **101**, 034103.